Utilizing Compressibility in Reconstructing Spectrographic Data, with Applications to Noise Robust ASR

Bengt J. Borgström, *Student Member, IEEE,* and Abeer Alwan, *IEEE Fellow*

***Abstract*— In this letter we propose a novel algorithm for reconstructing unreliable spectrographic data, a method applicable to missing feature-based automatic speech recognition (ASR). We provide quantitative analysis illustrating the high compressibility of spectrographic speech data. The existence of sparse representations for spectrographic data motivates the spectral reconstruction solution to be posed as an optimization problem minimizing the ℓ_1-norm. When applied to the Aurora-2 database, the proposed missing feature estimation algorithm is shown to provide significant improvements in recognition accuracy, relative to the baseline MFCC system. Even without an oracle mask, performance approaches that of the ETSI advanced front end (AFE) [1], with less complexity.**

***Index Terms*— Spectral Reconstruction, Missing Features, Compressibility, Noise Robust Automatic Speech Recognition.**

I. Introduction

The missing feature (MF) approach to robust automatic speech recognition (ASR) is effective in unfavorable acoustic environments [5]. MF algorithms can be grouped into two main categories; marginalization [2] and data imputation [3]. This letter focuses on data imputation techniques, which aim to reconstruct unreliable spectrographic components prior to recognition [3].

In this letter, we propose a novel missing feature data imputation algorithm for noise robust ASR based on the notion of compressibility. We provide quantitative analysis on the compressibility of spectrographic speech data, which motivates spectral reconstruction to be posed as an optimization problem minimizing the ℓ_1-norm. The proposed missing feature algorithm is shown to provide significant improvements in word-accuracy rates, relative to the baseline system, when applied to the Aurora-2 database [12]. Even without an oracle mask, performance approaches that of the ETSI advanced front end (AFE) [1].

II. The Compressibility of Speech

A. Signal Recovery from Incomplete Observations

Compressive sampling (CS) theory states that perfect reconstruction of signals can be achieved with far fewer observations than required by the traditional Nyquist sampling rate [6]-[7]. As discussed in [7], recovery of signals from an incomplete set of observations is made possible by the *sparsity* of the signal of interest, and by the *incoherence* of utilized sensing functions. In this section, we present a brief introduction to CS theory, specifically the notion of sparsity, and provide motivation for the use of linear programming in the current problem of missing feature estimation.

Let $\mathbf{f} \in \mathbb{R}^N$ represent a signal of interest, and let the set $\phi_k \in \mathbb{R}^N$, for $k = 1, \ldots, M$, represent sensing functions used to obtain M observations in $\mathbf{y}$ according to $\mathbf{y} = \Phi\mathbf{f}$, where Φ is comprised of row vectors ϕ_k. The design of sensing functions is an important aspect of many CS applications, such as imaging [8], and involves the concept of minimizing the coherence of bases (see [7] for details).

An underlying reason for the success of CS theory is that many signals can be described efficiently when expressed in terms of a proper basis. Let $\Psi \in \mathbb{R}^{N\times N}$ represent a suitable orthonormal basis for $\mathbf{f}$, such that $\mathbf{f} = \Psi^*\mathbf{v}$, where the * operator represents the conjugate transpose. Here, $\mathbf{v}$ is the sparse representation of $\mathbf{f}$, expanded in the basis Ψ, also referred to as the representation basis. The compressible or sparse nature of $\mathbf{v}$ states that it is comprised of only a few large magnitude terms, and implies that discarding the small terms will result in little or no distortion.

Define a signal as *S-sparse* if it contains at most S nonzero terms. Furthermore, let $\mathbf{v}_S$ be the vector comprised of the S largest magnitude terms of $\mathbf{v}$, with the remaining terms set to zero. The recovered version of the original signal can be expressed as $\mathbf{f}_S$=$\Psi^*\mathbf{v}_S$. If the original signal is truly *S-sparse*, the reconstructed signal will be perfect. However, the original signal may be *compressible*, so that the magnitude of terms in $\mathbf{v}$ decreases quickly, which will result in a reconstructed signal with little distortion.

The signal reconstruction $\mathbf{f}_S$ can not generally be reproduced since it requires oracle information regarding the locations of large magnitude terms in $\mathbf{v}$. However, the discussion of sparsity motivates the use of the ℓ_1-norm during signal recovery, since optimization problems which minimize the ℓ_1-norm tend to solutions comprised of few nonzero terms [9]. In [6]-[7], the CS solution to the signal recovery problem is given by $\tilde{\mathbf{f}} = \Psi^*\tilde{\mathbf{v}}$, where $\tilde{\mathbf{v}}$ is determined by:

$$\min_{\tilde{\mathbf{v}}\in\mathbb{R}^N} \|\tilde{\mathbf{v}}\|_{\ell_1} \quad \text{subject to:} \quad \mathbf{y} = \Phi\Psi^*\tilde{\mathbf{v}}. \tag{1}$$

Thus, the reconstructed signal $\tilde{\mathbf{f}}$, given an incomplete set of observations $\mathbf{y}$, is the function which minimizes the ℓ_1-norm of the sparse representation $\tilde{\mathbf{v}}$. Although the cost function in Equation 1 is nonlinear, the problem statement can be rearranged as a linear program, and thus be solved quite efficiently [9].

B. The Compressibility of Spectrographic Speech Data

Compressive sensing has been successfully applied to both image compression and image denoising [7], [8]. At the heart of these applications lies the fact that images typically have sparse representations when expanded on certain transform bases. In this section, we explore the compressibility of spectrographic speech data. The discussion will motivate the use of linear programming in the proposed missing feature estimation algorithm.

Let $\mathbf{X}(k,n)$ be the spectrographic representation of an input speech signal, where n denotes frame number, and k denotes frequency channel index. In this section we provide compressibility analysis for Mel-filtered spectral data as an illustrative example.

Let $\mathbf{x}$ represent the vector representation of $\mathbf{X}(k,n)$ formed by lexicographic ordering. We assume the existence of an orthonormal basis Ψ which reveals a concise representation of $\mathbf{x}$, namely $\mathbf{v}$. Additionally, assuming oracle information regarding the location of large magnitude terms within $\mathbf{v}$, we can extract the *S-sparse* vector $\mathbf{v}_S$, and recover the approximation of the original speech signal, $\mathbf{x}_S$. Let β be the portion of terms within $\mathbf{v}$ retained, and be defined as $\beta = \frac{N_s}{N}$, where N_s is the number of nonzero terms retained, and N is the total number of elements. The quality of the recovered signal, as a function of β, can be analyzed to assess the compressibility of the original data.

In image processing, the mean-square error (MSE) distortion provides a reliable metric for measuring the degradation of a compressed image [10]. However, in speech processing, such a distance metric applied to spectrographic speech data does not directly reflect the quality of the underlying speech data, and instead the performance of the overall speech processing system must be analyzed. In automatic speech recognition (ASR), one can study the effect of induced sparsity on the resulting recognition accuracy rates.

Figure 1 shows an example of the sparsity of Mel-filtered spectrographic speech data. Analysis was performed on the word "three" extracted from the Aurora-2 database. The top panel shows the sparse representation of the input Mel-filtered spectrographic data in vector form, utilizing the discrete Haar transform (DHT) [10]. The DHT was chosen due to its common usage in compressive sensing. Other transforms such as the Discrete Cosine Transform (DCT) and the Karhunen-Loeve Transform (KLT) were tested, but with less success. The bottom panel shows the absolute value of the sparse representation, sorted by magnitude. As can be concluded from the rapidly decreasing values in the bottom panel, the input spectrographic data is highly compressible.

Figure 2 provides quantitative analysis of the compressibility of Mel-filtered spectrographic speech data, presenting the time-average MSE, along with word-accuracies, as a function of induced sparsity. The representation basis used was again the discrete Haar kernel. Compressibility analysis was performed on clean speech from the Aurora-2 database [12]. It can be concluded from Figure 2 that approximately 90% (β=0.90) of terms in the sparse domain can be zeroed without significantly affecting recognition results.

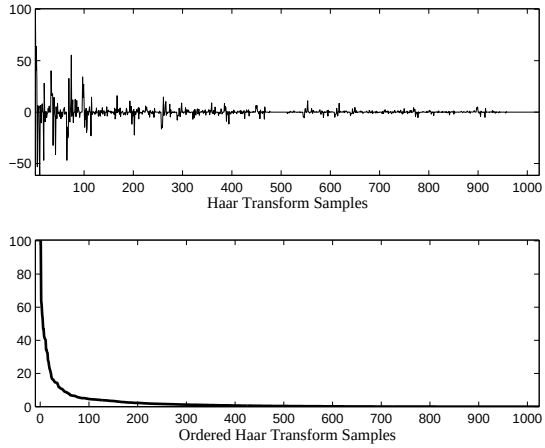

Fig. 1. The Compressibility of Spectrographic Speech Data: Analysis was performed on the clean word "three" extracted from the Aurora-2 database [12]. The top panel shows the sparse representation of the input spectrographic data in vector form, utilizing the discrete Haar transform (DHT). The bottom panel shows the absolute value of the sparse representation, sorted by magnitude.

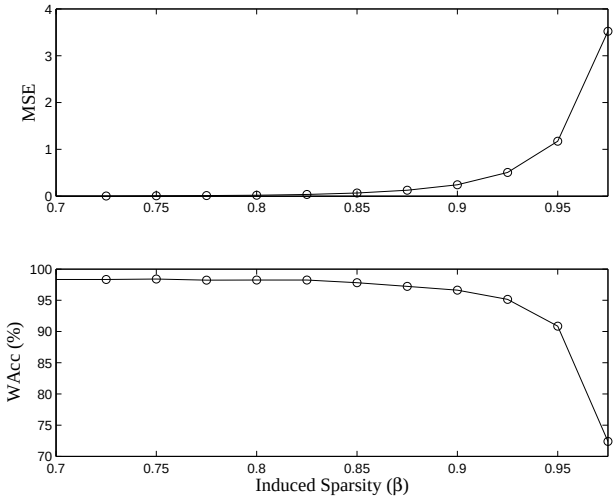

Fig. 2. Quantitative Analysis of the Compressibility of Spectrographic Speech Data: The top panel illustrates the MSE distortion resulting from induced sparsity in $\mathbf{x}_S$, utilizing the discrete Haar transform (DHT). The bottom panel represents the word-accuracies corresponding to the recovered Mel-filtered spectra used in the top panel.

III. Reconstruction of Missing Features for Noise Suppression in Speech

In real world applications, speech signals will generally suffer degradation due to acoustic noise, resulting in decreased performance for recognition. In this section we present a missing feature estimation method for noise suppression of spectrographic speech data, which in turn is applicable to automatic speech recognition.

A. The Proposed Missing Feature Estimation Algorithm

Assuming an additive noise model and assuming independence of speech and noise components, an observed speech

signal can be approximated in the spectral domain as:

$$\mathbf{X}(k,n) = \mathbf{S}(k,n) + \mathbf{D}(k,n), \text{ for } 1 \leq k \leq N, \quad (2)$$

where $\mathbf{D}(k,n)$ is the corrupting noise, and $\mathbf{S}(k,n)$ is the underlying clean speech signal. Let $\Psi \in \mathbb{R}^{N\times N}$ be the representation basis revealing a compressible representation of $\mathbf{x}$, namely $\mathbf{v}$.

In MF-based noise robust ASR systems, spectral reconstruction algorithms must be preceded by mask estimation. We utilize two types of binary masks, oracle masks and ones based on speech presence probability (SPP), which classify each term in $\mathbf{x}$ as reliable, corresponding to strong speech signal presence relative to noise, or unreliable, corresponding to a high level of corruption due to noise. In this study, oracle masks were determined via a simple SNR comparison between the observed spectrum and a noise spectrum obtained from linear spectral subtraction. A hard threshold of 0 dB was then used to differentiate between reliable and unreliable components. On the other hand, SPP-based masks were determined by calculating speech presence probabilities [11] throughout spectro-temporal locations, and using a hard probability threshold of 0.4 to determine a binary mask. Thresholds were optimized empirically.

A spectral reliability mask can be expressed analytically via the selection matrix $\mathbf{A}_R \in \mathbb{R}^{M\times N}$, corresponding to the M reliable components of $\mathbf{x}$. $\mathbf{A}_R$ is defined as follows:

$$\mathbf{A}_R(i,j) = \begin{cases} 1, \text{ if } \mathbf{x}(j) \text{ is the } i^{th} \text{ reliable term in } \mathbf{x} \\ 0, \quad \text{otherwise} \end{cases}. \quad (3)$$

The goal of signal reconstruction can be restated as estimating the components of the sparse representation $\mathbf{v}$, given the incomplete reliable observations ($\mathbf{A}_R\mathbf{x}$). Motivated by the discussions on CS from Section II-A, and on the compressibility of speech from Section II-B, the missing feature estimation task can be posed as an optimization problem minimizing the ℓ_1-norm:

$$\min_{\tilde{\mathbf{v}}\in\mathbb{R}^N} \|\tilde{\mathbf{v}}\|_{\ell_1} \quad \text{subject to:} \quad \mathbf{A}_R\mathbf{x} = \mathbf{A}_R\Psi^*\tilde{\mathbf{v}} \quad (4)$$

Utilizing Equation 4, the sparse representation of the estimated underlying speech spectrum can be determined. The reconstructed clean speech spectrum can be found as $\tilde{\mathbf{s}} = \Psi^*\tilde{\mathbf{v}}$.

The spectral reconstruction solution expressed in Equation 4 does not take into account any information specific to spectrographic speech data. Basic properties of spectral signals can be integrated as constraints in the optimization problem of Equation 4 to better estimate the underlying clean speech spectrum. First, spectral coefficients are inherently nonnegative. Also, following the additive model from Equation 2, it can be concluded that clean speech components must be less than or equal to observed spectral components. Thus, the optimization from Equation 4 becomes:

$$\min_{\tilde{\mathbf{v}}\in\mathbb{R}^N} \|\tilde{\mathbf{v}}\|_{\ell_1} \quad \text{subject to:} \quad \begin{aligned} &\mathbf{A}_R\mathbf{x} = \mathbf{A}_R\Psi^*\tilde{\mathbf{v}} \\ &\Psi^*\tilde{\mathbf{v}} \geq 0 \\ &\Psi^*\tilde{\mathbf{v}} \leq \mathbf{x} \end{aligned} \quad (5)$$

The additional constraints of Equation 5 provide boundaries for the solution $\tilde{\mathbf{v}}$, specific to Mel-filtered spectral data, which may not exist for other types of spectrographic data. These constraints were found to result in more accurate solutions than cases with no such constraints.

As stated previously, the optimization in Equation 5 can be carried out efficiently by posing the problem as a linear program (LP) [9]. In this study, the primal-dual LP method was used during optimization. Also, the algorithm is able to run in real-time by processing one spectral frame (25 ms with a 10 ms overlap) at a time.

It is interesting to note that the subspace approach to noise suppression [13] also exploits the sparsity of speech for noise-robust processing. Such techniques explicitly construct the sparse design of speech signals given predetermined threshold(s). In the proposed algorithm, however, the sparse design is determined implicitly during the minimization of the ℓ_1-norm of the sparse representation $\tilde{\mathbf{v}}$.

B. Comparisons with Compressive Sensing

As can be interpreted from Equations 1 and 4, great similarity exists between compressive sensing and the proposed missing feature estimation algorithm. Both techniques aim to reconstruct signals from an incomplete set of observations. Additionally, both techniques rely on the notion of compressibility, and specifically each method minimizes the ℓ_1-norm of the given signal in a sparse domain. However, the concept of sensing is quite different in each case.

In CS applications, such as imaging [8], signals are sampled at low rates by utilizing sensing functions. Sensing functions are designed to comprise an orthonormal basis, Φ, which minimizes the coherence measure with the representation basis, Ψ. (See [7] for a detailed discussion.)

In the proposed missing feature estimation algorithm, observations are not sampled in the same sense as in CS applications, but their origins are instead decided by the reliability of terms in the mask estimation domain. The sensing matrix, which for the proposed algorithm can be written as Φ=$\mathbf{A}_R$, is defined entirely by the corruptive effect of noise on the input speech signal. Thus, sensing functions cannot be actively designed, and the notion of coherence does not play a role.

IV. Experimental Results

The proposed spectral reconstruction algorithm was applied to Set A of the Aurora-2 database [12], with training performed on clean data. The system extracted 39-dimensional MFCCs (with first and second derivatives), and the recognizer used 16-states, 3-mixture word models. The proposed algorithm was run one frame at a time, and the Haar kernel was used for analysis/reconstruction. Table I provides word-accuracy rates for the proposed algorithm when combined

TABLE I

Word-Accuracies for the Proposed Missing Feature Estimation Technique, using oracle masks (ORC) and SPP-based masks (SPP). The baseline refers to a standard MFCC front end. Results were obtained on Set A of the Aurora-2 database [12].

SNR (dB)	20	15	10	5	0
Subway Noise					
baseline	96.84	91.71	73.29	36.08	5.74
ORC	99.02	98.62	97.11	91.34	80.96
SPP	97.48	95.09	87.93	73.41	46.91
Babble					
baseline	97.79	94.89	77.99	37.73	3.69
ORC	98.40	98.22	97.52	95.19	86.37
SPP	98.00	96.43	90.93	74.15	40.57
Vehicular Noise					
baseline	97.49	92.04	68.95	22.37	0.89
ORC	98.54	98.48	97.05	93.02	81.39
SPP	97.70	96.39	89.56	73.13	45.15
Exhibition Hall					
baseline	96.85	91.55	73.71	33.51	4.35
ORC	98.55	98.21	96.33	89.36	75.59
SPP	97.75	95.19	87.01	68.40	36.93
Average					
baseline	97.24	92.55	73.49	32.42	3.67
ORC	98.65	98.38	97.00	92.23	81.08
SPP	97.73	95.77	88.86	72.27	42.39

with oracle reliability masks (ORC) [3] and masks based on speech presence probability (SPP). In general, results obtained utilizing oracle masks provide an upper performance bound for missing feature methods. Results obtained through mask estimation represent a more realistic scenario, since they utilize information solely from the noisy input signal with no prior knowledge. The proposed spectral reconstruction algorithm is shown to provide a high upper performance bound when combined with oracle reliability masks. Additionally, when combined with SPP-based masks, the proposed algorithm provides significant improvements in word-accuracy rates, relative to the baseline system which uses a standard MFCC front end.

The proposed spectral reconstruction framework is compatible with various proven noise robust feature extraction and post-processing techniques. Table II provides word-accuracy results for the combination of the proposed spectral reconstruction algorithm with Peak Isolation (PK-ISO) post-processing [4], averaged across all conditions included in Set A of the Aurora-2 database. As can be concluded from Table II, combining the proposed spectral reconstruction algorithm with Peak Isolation provides further improvements in system performance. For comparison, Table II includes results obtained using solely Peak Isolation. Additionally, Table II includes comparisons with the noise robust ETSI AFE front-end [1]. Even without an oracle mask, the algorithm provides performance similar to the ETSI AFE, with less complexity with respect to processing time.

TABLE II

Word-Accuracies for the Proposed Missing Feature Estimation Technique in Combination with Peak Isolation [4] Post-Processing (PK-ISO). Results for the ETSI AFE [1] are included for comparison. Results were averaged across all conditions included in Set A of the Aurora-2 database.

SNR (dB)	20	15	10	5	0
MFCC (baseline)	97.24	92.55	73.49	32.42	3.67
MFCC (ORC)	98.65	98.38	97.00	92.23	81.08
MFCC + PK-ISO (ORC)	98.57	98.47	97.68	95.60	88.99
MFCC (SPP)	97.73	95.77	88.86	72.27	42.39
MFCC + PK-ISO (SPP)	97.98	96.25	91.44	78.54	51.10
PK-ISO	97.04	94.37	83.67	56.15	19.15
ETSI AFE	98.46	96.96	92.22	79.13	51.11

V. Conclusion

This letter presents a novel algorithm for the reconstruction of unreliable spectral speech components based on the notion of compressibility. We provide quantitative analysis on the compressibility of spectrographic speech data, which motivates the use of minimization of the ℓ_1-norm in the proposed missing feature estimation technique. The proposed spectral reconstruction method is shown to provide significant improvements in word-accuracy rates relative to the MFCC baseline system, when applied to the Aurora-2 database.

Future work will focus on the design of spectral reliability masks. Masks which better differentiate between reliable and unreliable spectral components can be expected to perform closer to the bound obtained by using oracle masks.

VI. Acknowledgments

This work was supported in part by the NSF.

References

[1] ETSI Standard Doc., *Speech Processing, Transmission, and Quality Aspects (STQ); Distributed Speech Recognition; Advanced Front-end Feature Extraction Algorithms; Compression Algorithms*, ETSI ES 202 050 V1.1.5 (2007-01).

[2] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, *Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data*, Speech Communication, Vol. 34, pp. 267-285, 2001.

[3] B. Raj, M. L. Seltzer, and R. M. Stern, *Reconstruction of Missing Features for Robust Speech Recognition*, Speech Communication, vol. 43, pp. 275-296, 2004.

[4] B. Strope and A. Alwan, *A Model of Dynamic Auditory Perception and its Application to Robust Word Recognition*, IEEE Trans. on Speech and Audio Processing, Vol. 5, No. 5, pp. 451-464, 1997.

[5] B. Raj and R. Stern, *Missing Feature Approaches in Speech Recognition*, IEEE Signal Processing Magazine, Vol. 22, Issue 5, pp. 101-116, 2005.

[6] D. L. Donoho, *Compressed Sensing*, IEEE Trans. on Information Theory, Vol. 52, No. 4, pp. 1289-1306, 2006.

[7] E. J. Candes and M. B. Wakin, *An Introduction to Compressive Sampling*, IEEE Signal Processing Magazine, Vol. 25, No. 2, pp. 21-30, 2008.

[8] J. Romberg, *Imaging via Compressive Sensing*, IEEE Signal Processing Magazine, Vol. 25, No. 2, pp. 14-20, 2008.

[9] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[10] A. K. Jain, *Fundamentals of Digital Image Processing*, Prentice Hall, 1989.

[11] P. Vary and R. Martin, *Digital Speech Transmission: Enhancement, Coding and Error Concealment*, New York: Wiley, 2006.

[12] D. Pearce, *Enabling New Speech Driven Services For Mobile Devices: An Overview of the ETSI Standards Activities for Distributed Speech Recognition Front-Ends*, AVIOS 2000: Speech Appl. Conf., Vol. 5, May 2000.

[13] K. Hermus, P. Wambacq, and H. Van hamme, *A Review of Signal Subspace Speech Enhancement and Its Application to Noise Robust Speech Recognition*, EURASIP Journal on Advances in Signal Processing, Vol. 2007, pp. 1-15, 2007.