MISSING FEATURE IMPUTATION OF LOG-SPECTRAL DATA FOR NOISE ROBUST ASR

Bengt J. Borgstrom and Abeer Alwan

Department of Electrical Engineering,
University of California, Los Angeles
jonas, alwan@ee.ucla.edu

ABSTRACT

In this paper, we present a missing feature (MF) imputation algorithm for log-spectral data with applications to noise robust ASR. Drawing from previous work [1], we adapt the previously proposed spectrographic reconstruction solution to the liftered log-spectral domain by introducing log-spectral flooring (LS-FLR). LS-FLR is shown to be an efficient and effective noise robust feature extraction technique. When LS-FLR is integrated in deriving the novel log-spectral data imputation framework, the overall system is shown to provide significant improvements in noise robust speech recognition.

Index Terms— Missing Features, Feature Extraction, Compressibility, Noise Robust Automatic Speech Recognition.

1. INTRODUCTION

In-vehicle speech recognition remains a challenging task due to varying levels of background noise [2],[3]. In this paper, we present a front-end framework designed to improve the robustness of speech recognition to vehicular environments.

The missing feature (MF) approach to noise robust ASR has been shown to be successful in difficult acoustic environments [5]. Specifically, data imputation may improve ASR in noisy scenarios by reconstructing unreliable spectrographic components prior to recognition. Prior studies have utilized correlation-based and cluster-based methods to obtain estimates of clean spectral data [4]. In [1], we propose an imputation solution utilizing theory from compressive sensing [6], which exploits sparsity in the underlying clean data to reconstruct the signal of interest given an incomplete set of reliable components.

Prior missing feature data imputation studies have typically focussed on linear or Mel-filtered spectrographic data [4]. In this paper, we present an efficient imputation technique for liftered log-spectral data, which draws upon the spectral reconstruction solution from [1]. The log-spectral domain offers an attractive alternative for data imputation due to increased spectral correlation.

This work was supported in part by NSF.

In order to adapt the solution from [1] to the log-spectral domain, we present log-spectral flooring as a novel noise robust front-end processing technique. This technique minimizes variability due to corruptive noise, as well as reducing the dynamic range of otherwise unbounded log-spectral data. Log-spectral flooring outperforms traditional flooring methods in the spectral domain, and is shown to serve on its own as an efficient and effective noise robust algorithm.

When adapted to the log-spectral domain, the imputation solution of [1] provides significant improvements to ASR noise robustness.

This paper is organized as follows: in Section 2 we present log-spectral flooring. In Section 3, we review the reconstruction solution from [1] and adapt it to the liftered log-spectral domain. Experimental results are provided in Section 4. Finally, we provide conclusions in 5.

2. LOG-SPECTRAL FLOORING

The aim of designing noise robust front-end features for ASR is to maximize discriminative spectral information while minimizing variability due to noise [7]. It has been widely reported that discriminative speech information tends to lie in high amplitude spectral peaks, whereas noise tends to lie in spectral valleys. Such observations have been motivated by human perception [8] and shown quantitatively [7]. Resulting efforts to overcome ASR performance degradation due to noise include spectral flooring [7] and peak isolation [8].

In this section we propose a simple algorithm for front-end feature extraction which aims to minimize feature variability due to noise by flooring observed data in the log-spectral domain. Let $\mathbf{x}$ represent the Mel-filtered spectrogram of an observed speech signal. The liftered log-Mel-spectrogram is obtained via:

$$\mathbf{x}_L = \mathbf{C}^{-1}\mathbf{L}_{cep}. \times \mathbf{C}\log\left(\mathbf{x}\right), \tag{1}$$

where $\mathbf{C}$ denotes the discrete Cosine transform (DCT) and $\mathbf{L}_{cep}$ denotes the cepstral lifter. Also $.\times$ denotes element-by-element multiplication.

The logarithmic function is applied during feature extraction (Eq. 1) as it simulates the human auditory system, and

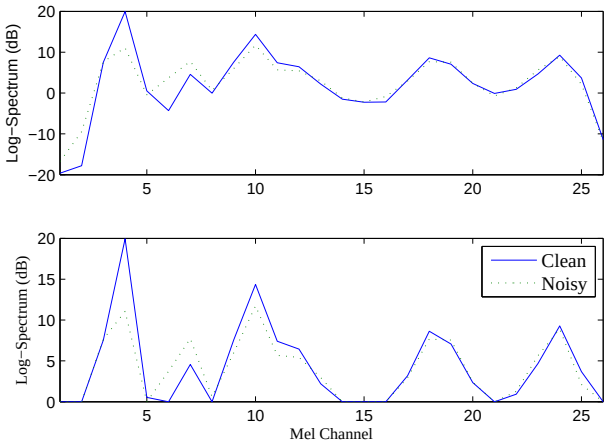

Fig. 1. Log-Spectral Flooring: The top panel shows the clean and noisy versions of liftered log-spectra corresponding to an active frame of speech (/ε/ from "seven"). The bottom panels shows the same spectra after log-spectral flooring. The noisy speech was degraded by vehicular noise at 5 dB SNR.

effectively compands the energy scale. This serves to emphasize discriminative speech patterns present in the spectral envelope, which are considered important for recognition [8]. However, the logarithm results in an unbounded lower limit, leading to an increased dynamic range for spectral valleys. This may cause confusion during recognition since spectral valleys typically contain little discriminative speech energy. To reduce the confusability introduced by the logarithmic function on small spectral values, we define log-spectral flooring (LS-FLR) as:

$$\mathbf{x}_L(m,n) = \begin{cases} \mathbf{x}_L(m,n), \text{ if } \mathbf{x}_L(m,n) \geq \alpha_{fl} \\ \alpha_{fl}, \text{ otherwise} \end{cases} . \quad (2)$$

In Eq. 2, m denotes Mel channel index, and n denotes frame index. Furthermore, α_{fl} determines the level of flooring. In this study, the flooring parameter was empirically optimized to $\alpha_{fl} = 0.0$ dB.

Fig. 1 provides an illustrative example of log-spectral flooring. The top panel shows the clean and noisy versions of liftered log-spectra corresponding to an active frame of speech (/ε/ from "seven"). The bottom panels shows the same spectra after log-spectral flooring. The noisy speech was degraded by vehicular noise at 5 dB SNR. LS-FLR reduces variability between clean and noisy spectral features, which is visible in lower Mel channels. Additionally, flooring reduces the dynamic range of log-spectra by instituting a lower bound below which little discriminative speech energy typically lies.

The proposed LS-FLR algorithm was applied to the Aurora-2 database to assess its ability to improve noise robust ASR.

Table 1. *Word-Accuracies for Log-Spectral Flooring (LS-FLR): Spectral flooring (S-FLR) is included for comparison. Results are averaged across all noise types in Set A of the Aurora-2 database. All systems are MFCC-based.*

SNR (dB)	20	15	10	5	0	-5
Vehicular Noise						
none	96.1	89.0	67.3	25.2	1.0	0.0
S-FLR	96.9	92.5	79.7	48.4	7.0	0.1
LS-FLR	98.4	97.1	91.1	71.4	27.7	9.1
Average Across All Noise Types						
none	95.8	89.6	70.4	34.6	5.2	0.2
S-FLR	96.8	93.3	82.3	55.5	14.2	0.8
LS-FLR	97.6	95.2	88.2	69.6	34.2	11.9

Table 1 provides word-accuracy results for vehicular noise in particular, as well as results averaged across all noise types in Set A. For comparison, spectral flooring [7] is included. It can be observed in Table 1 that flooring in the log-spectral domain provides significant performance improvements relative to traditional spectral flooring, and thus serves as an efficient and effective front end noise robust tool.

3. MISSING FEATURE DATA IMPUTATION

3.1. Reconstruction in the Spectral Domain

In the missing feature (MF) approach to noise robust ASR, unreliable spectro-temporal components of observed signals are detected and compensated for [5]. MF algorithms can be grouped into two main approaches. The first, marginalization, deemphasizes the effect of unreliable components on back-end posterior probabilities. The second, data imputation, reconstructs unreliable components prior to recognition.

In [1], we propose a MF data imputation algorithm drawing on compressive sensing theory [6]. We provide qualitative analysis on the compressibility of spectrographic speech data. Furthermore, we explore the effect of induced sparsity on ASR performance.

Here, we briefly review the reconstruction algorithm presented in [1]. Let $\mathbf{x} \in \mathbb{R}^N$ represent the vector-form Mel-filtered spectrographic representation of an observed noisy speech signal. Let $\mathbf{y} \in \mathbb{R}^K$ be the subset of reliable components obtained via:

$$\mathbf{y} = \mathbf{A}_R \mathbf{x}, \quad (3)$$

where $K \ll N$. Details regarding the selection matrix $\mathbf{A}_R$ can be found in [1]. Additionally, let Ψ be a suitable basis which reveals a sparse representation of $\mathbf{x}$ as:

$$\mathbf{v} = \Psi\mathbf{x}. \tag{4}$$

The spectral reconstruction solution from [1], referred to as $MF-SP$, is given as:

$$\min_{\tilde{\mathbf{v}}\in\mathbb{R}^N} \|\tilde{\mathbf{v}}\|_1 \quad \text{s.t.} \quad (i)\ \mathbf{A}_R\mathbf{x} = \mathbf{A}_R\Psi^*\mathbf{v} \tag{5}$$
$$(ii)\ \Psi^*\mathbf{v} \leq \mathbf{x}$$
$$(iii)\ \Psi^*\mathbf{v} \geq 0$$

In Eq. 5, constraint (i) acts to fix the values of reliable components during imputation. Constraint (ii) follows from the additive model of noisy speech, and constraint (iii) follows from the nonnegative property of spectral data.

3.2. Extension to Imputation of Log-Spectral Data

Success of the previous data imputation algorithm [1] depends on the existence of a compressible representation of the observed signal. Fig. 2 provides qualitative analysis of the compressibility of spectro-temporal data in the log-spectral domain. As in [1], the discrete Haar transform (DHT) is used as the representation basis.

In Fig. 2, the top panel illustrates the log Mel-filtered spectrogram of the clean utterance "three zero eith two," from the Aurora-2 database. The middle panel provides the DHT of the vector-form log-Mel-spectrogram. The bottom panel shows the ordered magnitudes of the DHT, revealing high compressibility of the original log-Mel-spectrographic data.

Motivated by observations from Fig. 2, we adapt the spectral reconstruction solution from [1] to data imputation in the log-spectral domain. With $\mathbf{x}_L$ previously defined, we let:

$$\mathbf{y}_L = \mathbf{A}_R\mathbf{x}_L, \tag{6}$$

and:

$$\mathbf{v}_L = \Psi\mathbf{x}_L. \tag{7}$$

Note that constraint (ii) from Eq. 5 is valid for log-spectral data due to the monotonic nature of the logarithm. Constraint (iii), on the other hand, does not hold since log-spectral values become unbounded with respect to the lower limit. However, by applying log-spectral flooring from Section 2, constraint (iii) can be stated alternatively, and the data imputation solution in the log-spectral domain, referred to as $MF\text{-}LOG$, can be expressed as:

$$\min_{\tilde{\mathbf{v}_L}\in\mathbb{R}^N} \|\tilde{\mathbf{v}_L}\|_1 \quad \text{s.t.} \quad (i)\ \mathbf{A}_R\mathbf{x}_L = \mathbf{A}_R\Psi^*\mathbf{v}_L \tag{8}$$
$$(ii)\ \Psi^*\mathbf{v}_L \leq \mathbf{x}_L$$
$$(iii)\ \Psi^*\mathbf{v}_L \geq \alpha_{fl}$$

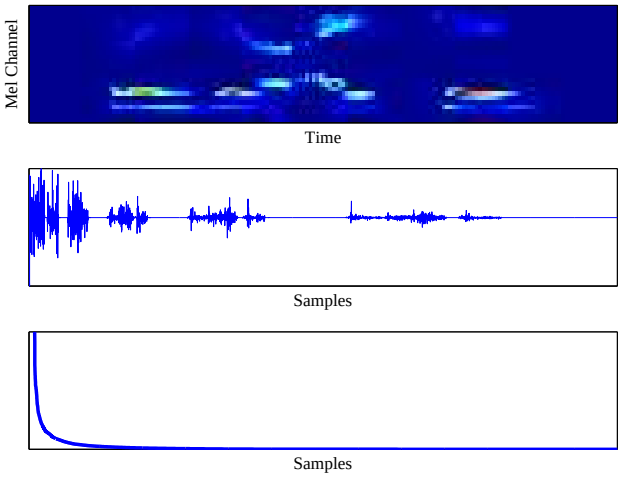

Fig. 2. The Compressibility of Spectro-Temporal Speech Data in the Log-Spectral Domain: The top panel shows the liftered log-Mel-spectrogram of the utterance "three zero eight two," the middle panel provides the DHT, and the bottom panel shows the ordered magnitudes of the DHT.

Table 2. *Word-Accuracies for LS-FLR by itself, and when incorporated in MF-LOG Data Imputation: MF-SP refers to the previously proposed spectral reconstruction solution from [1], and SB and ORC refer to the use of signal-based and oracle masks, respectively. ETSI AFE refers to the ETSI advanced front-end [9]. All systems were MFCC-based. Results were obtained on the Aurora-2 database.*

SNR (dB)	20	15	10	5	0
Vehicular Noise					
none	96.1	89.0	67.3	25.2	1.0
LS-FLR	98.4	97.1	91.1	71.4	27.7
MF-LOG (SB)	**98.6**	**98.0**	**95.6**	**86.8**	**62.5**
MF-SP (SB) [1]	97.7	96.4	89.6	73.1	45.2
MF-LOG (ORC)	98.7	98.3	96.7	91.4	75.6
ETSI AFE [9]	98.7	97.8	94.2	81.7	53.8
Average Across All Noise Types					
none	95.8	89.6	70.4	34.6	5.2
LS-FLR	97.6	95.2	88.2	69.6	34.2
MF-LOG (SB)	**98.3**	**97.2**	**94.1**	**84.5**	**60.3**
MF-SP (SB) [1]	97.7	95.8	88.9	72.3	42.4
MF-LOG (ORC)	98.7	98.0	96.4	91.6	78.7
ETSI AFE [9]	98.2	97.0	92.2	79.1	51.1

Thus, the reconstructed log-spectral signal in the sparse domain is that which minimizes the ℓ_1-norm, fixed by observed reliable log-spectral components, and constrained by upper and lower boundaries (ii) and (iii). Although the solution in Eq. 8 can be applied to groups of frames to exploit temporal correlation, in this study we applied it on a single-frame basis to eliminate delay requirements. Additionally, although the cost function in Eq. 8 is nonlinear, it can be rearranged as a linear program (LP) and solved quite efficiently [10]. Specifically, we implemented a primal-dual LP solver with a maximum of 5 iterations.

4. EXPERIMENTAL RESULTS

An integral component of missing feature systems is mask estimation, which determines the reliability of observed spectrographic components. In MF studies, imputation methods are commonly tested both with oracle masks and signal-based masks [5]. Oracle masks require information regarding the clean version of the input speech signal, and therefore do not represent a realistic scenario. However, they provide an upper performance bound for data imputation techniques. Signal-based masks, on the other hand, utilize solely the observed noisy signal. Further details regarding mask estimation can be found in [4] and [1].

The log-spectral data imputation algorithm proposed in Section 3 was applied to the Aurora-2 database to assess its effectiveness. The algorithm was used both with oracle masks (ORC) and signal-based masks (SB) [5]. Results are provided in Table 2. We examine the performance of LS-FLR by itself, as well as when incorporated into the MF-LOG technique. Table 2 includes results for the spectral reconstruction solution of Eq. 5 presented in [1]. Additionally, the ETSI AFE [9] is included for comparison.

It can be observed that the proposed log-spectral data imputation algorithm (MF-LOG) provides significant improvements for noise robust ASR, and outperforms the algorithm presented in [1] for the case of signal-based masks. Even without the use of oracle masks, the proposed technique provides better noise robustness than the ETSI AFE.

5. CONCLUSIONS

In this paper, we presented an algorithm for missing feature data imputation in the log-spectral domain. Drawing from previous work in [1], we adapted the previously proposed solution for spectrographic reconstruction to the log-spectral domain by introducing log-spectral flooring. LS-FLR is shown to serve as an efficient and effective noise robust feature extraction technique on its own. The overall MF-LOG data imputation system is shown to provide significant improvements in noise robust recognition when combined with signal-based masks. When combined with oracle masks, the system provides an impressive upper performance bound.

6. REFERENCES

[1] B. J. Borgstrom and A. Alwan, *Utilizing Compressibility in Reconstructing Spectrographic Data, with Applications to Noise Robust ASR*, Signal Processing Letters, vol. 16, Issue 5, pp. 398-401, 2009.

[2] W. Kim and J. H. L. Hansen, *Feature compensation in the cepstral domain employing model combination*, Speech Communication 51(2): 83-96, 2009.

[3] M. Akbacak and J. H. L. Hansen, *Environmental Sniffing: Noise Knowledge Estimation for Robust Speech Systems*, IEEE Trans. Audio, Speech and Language Processing, vol. 15, no.2, pp. 465-477, Feb. 2007.

[4] 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.

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

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

[7] Q. Zhu and A. Alwan, *Non-linear feature extraction for robust recognition in stationary and non-stationary noise*, Computer, Speech, and Language, 17(4): 381-402, Oct. 2003.

[8] B. Strope A. and Alwan, *A model of dynamic auditory perception and its application to robust word recognition*, IEEE Transactions on Speech and Audio Processing, Vol. 5, No. 5, pp. 451-464, September 1997.

[9] ETSI ES 202 050 Ver. 1.1.5, 2007.

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

[11] 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.