

SPARSE IMPUTATION FOR NOISE ROBUST SPEECH RECOGNITION USING SOFT MASKS

J. F. Gemmeke, B. Cranen

Dept. of Linguistics, Radboud University
P.O. Box 9103, NL-6500 HD
Nijmegen, The Netherlands

{J.Gemmeke, B.Cranen}@let.ru.nl

ABSTRACT

In previous work we introduced a new missing data imputation method for ASR, dubbed *sparse imputation*. We showed that the method is capable of maintaining good recognition accuracies even at very low SNRs provided the number of mask estimation errors is sufficiently low. Especially at low SNRs, however, mask estimation is difficult and errors are unavoidable. In this paper, we try to reduce the impact of mask estimation errors by making soft decisions, i.e., estimating the probability that a feature is reliable. Using an isolated digit recognition task (using the AURORA-2 database), we demonstrate that using soft masks in our sparse imputation approach yields a substantial increase in recognition accuracy, most notably at low SNRs.

Index Terms— Speech recognition, Robustness, Redundancy

1. INTRODUCTION

Automatic speech recognition (ASR) performance degrades substantially when speech is corrupted by background noise that was not seen during training. Missing Data Techniques (MDTs) provide a powerful way to mitigate the impact of both stationary and non-stationary noise for a wide range of Signal-to-Noise (SNR) ratios [1; 2]. The general idea behind MDT is that it is possible to estimate—prior to decoding— which spectro-temporal elements of the acoustic representations are reliable (i.e., dominated by speech) and which are unreliable (i.e., dominated by background noise). These reliability estimates, referred to as a *spectrographic mask*, can then be used to treat reliable and unreliable features differently, for instance for replacing the unreliable features by clean speech estimates (i.e. *missing data imputation*).

Most missing data imputation methods work on a frame-by-frame basis (i.e. strictly local in time). At low SNRs (≤ 0 dB) a substantial number of frames may contain few, if any, reliable features. Providing clean speech estimates for these frames is difficult and hence ASR performance at low SNRs is severely reduced. In [3; 4], we introduced a new missing data imputation method, *sparse imputation*, to address this issue. Sparse imputation uses a time window which is (much) wider than a single frame. The method works by finding a sparse representation of the reliable features of an unknown word in an overcomplete basis of noise-free example words. The projection of these sparse representations in the basis is then used to provide clean speech estimates to replace the unreliable features.

In [3; 4] we showed that using the sparse imputation method significantly improved recognition accuracies even at low SNRs pro-

vided the number of mask estimation errors is sufficiently low. However, in practical settings, especially at low SNRs, missing data mask estimation errors are unavoidable. Previous studies [5; 6; 7] have shown that the influence of mask estimation errors can be reduced when the binary reliability score is replaced by the probability that a spectral component is reliable: *Soft* (or *fuzzy*) masks. In this paper we will present an extension to the sparse imputation method which enables it to use soft masks. The goal of this paper is to explore to what extent recognition accuracy improves when soft masks are used in the sparse imputation framework.

The rest of the paper is organized as follows. In Section 2 we introduce binary mask MDT and the sparse imputation framework. In Section 3 we extend this framework for use with soft masks. In Section 4 we compare recognition accuracies between binary masks and soft masks using isolated digits extracted from AURORA-2. We discuss the results in Section 5 and we present our conclusions in Section 6.

2. SPARSE IMPUTATION USING BINARY MASKS

2.1. Sparse representation of speech

In ASR, speech is represented as a spectro-temporal distribution of acoustic power, a *spectrogram*. We express the $K \times T$ spectrogram matrix (with K being the number of frequency bands and T the number of time frames) of clean speech S as a single vector s of dimension $D = K \cdot T$ by concatenating T subsequent time frames. We assume T to be fixed. This can be achieved, for example, by time-normalizing all utterances we want to process.

As in [3; 4], we consider s as a non-negative linear combination of exemplar spectrograms \mathbf{a}_n , where n denotes a specific exemplar ($1 \leq n \leq N_A$) in the set of N_A available exemplars. We write:

$$\mathbf{s} = \sum_{n=1}^{N_A} x_n \mathbf{a}_n = \mathbf{A} \mathbf{x} \quad (1)$$

with weights $x_n \geq 0 \in \mathbb{R}$, \mathbf{x} an N_A -dimensional weight vector, and $\mathbf{A} = (\mathbf{a}_1 \ \mathbf{a}_2 \ \dots \ \mathbf{a}_{N_A})$ a matrix with size $D \times N_A$.

Typically, the number of exemplar spectrograms will be much larger than the dimensionality of the acoustic representation ($N_A \gg D$). Therefore, the system of linear equations (1) has no unique solution. Research in the field of *compressed sensing* [8; 9] has shown, however, that if \mathbf{x} is *sparse*, \mathbf{x} can be determined *exactly* by solving:

$$\min_{\mathbf{x}} \{ \|\mathbf{x}\|_0 \} \text{ subject to } \mathbf{s} = \mathbf{A} \mathbf{x} \quad (2)$$

with $\|\cdot\|_0$ the l^0 zero norm (i.e., the number of nonzero elements). The combinatorial problem in Eq. 2 is NP-hard and therefore unfeasible for practical applications. However, it has been proven that, with mild conditions on the sparsity of \mathbf{x} and the structure of \mathbf{A} , \mathbf{x} can be determined [10] by solving:

$$\min_{\mathbf{x}} \{ \|\mathbf{A}\mathbf{x} - \mathbf{s}\|_2 + \lambda \|\mathbf{x}\|_1 \} \quad (3)$$

with a regularization parameter λ and a non-negativity constraint on \mathbf{x} . The resulting vector \mathbf{x} is a sparse representation of the clean speech vector \mathbf{s} .

2.2. Binary missing data masks

Assuming noise is additive the spectrogram of noisy speech, denoted by \mathbf{Y} , can be described as the sum of the individual spectrograms of clean speech \mathbf{S} and noise \mathbf{N} , i.e., $\mathbf{Y} = \mathbf{S} + \mathbf{N}$. Elements of \mathbf{Y} that predominantly contain speech or noise energy are distinguished by introducing a spectrographic mask \mathbf{M} . With all spectrograms represented as $K \times T$ dimensional matrices, a mask is also a $K \times T$ matrix. The elements of a *binary* mask \mathbf{M} are either 1, meaning the corresponding cell of \mathbf{Y} is dominated by speech ('reliable') or 0, meaning it is dominated by noise ('unreliable' c.q. 'missing'). Thus, we write:

$$M(k, t) = \begin{cases} 1 \stackrel{def}{=} \text{reliable} & \text{if } S(k, t)/N(k, t) > \theta \\ 0 \stackrel{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \quad (4)$$

with frequency band k ($1 \leq k \leq K$), time frame t ($1 \leq t \leq T$) and constant threshold θ . If log-spectral energy features are used, reliable noisy speech coefficients can be used directly as estimates of the clean speech features since $\log[|S(k, t) + N(k, t)|] = \log[|S(k, t)(1 + N(k, t)/S(k, t))|] \approx \log[|S(k, t)|]$.

In experiments with artificially added noise, the so-called *oracle masks* can be computed directly using Eq. 4. In realistic situations, however, the masks must be estimated. In Section 4 we will use an oracle mask and one estimated mask (i.c. harmonicity mask [11]). We refer to [12] and the references therein for a more complete overview of mask estimation techniques.

2.3. Imputation

By concatenating subsequent time frames of \mathbf{M} , similarly as we did for the spectrogram \mathbf{Y} , we construct a mask vector \mathbf{m} . We consider an observation vector \mathbf{y} derived from the noisy speech spectrogram \mathbf{Y} . We denote \mathbf{y}_r as the reliable coefficients of \mathbf{y} for which the corresponding elements of mask vector \mathbf{m} are equal to one. Rather than solving Eq. 3, we use the reliable \mathbf{y}_r as an approximation for \mathbf{s} and solve:

$$\min_{\mathbf{x}} \{ \|\mathbf{A}_r \mathbf{x} - \mathbf{y}_r\|_2 + \lambda \|\mathbf{x}\|_1 \} \quad (5)$$

with \mathbf{A}_r pertaining to the rows of \mathbf{A} for which $\mathbf{m} = 1$. As suggested in [13] it is possible to use the sparse representation \mathbf{x} obtained from Eq. 5 to estimate the missing values of \mathbf{y} by reconstruction:

$$\hat{\mathbf{y}} = \begin{cases} \hat{\mathbf{y}}_r = \mathbf{y}_r \\ \hat{\mathbf{y}}_u = \mathbf{A}_u \mathbf{x} \end{cases} \quad (6)$$

yielding the estimated clean speech vector $\hat{\mathbf{y}}$. \mathbf{A}_u and $\hat{\mathbf{y}}_u$ pertain to the rows of \mathbf{A} and $\hat{\mathbf{y}}$ for which $\mathbf{m} = 0$. A version of $\hat{\mathbf{y}}$ that is reshaped into a $K \times T$ matrix can be considered a denoised spectrogram of the underlying speech signal.

3. SPARSE IMPUTATION USING SOFT MASKS

3.1. Soft missing data masks

We define a *soft* mask which represents the probability that the clean speech dominates the background noise as follows:

$$M(k, t) = P(S(k, t)/N(k, t) > \theta) \quad (7)$$

with $M(k, t)$ now taking continuous values between 0 and 1. If the value is close to 1, the spectral component has a high probability of being dominated by speech. A soft mask can be generated directly using the probabilistic output of machine learning techniques [6], or by the approach followed in [5; 7], e.g. by the substitution of Eq. 4 in a sigmoid function:

$$M(k, t) = \frac{1}{1 + \exp^{-(S(k, t)/N(k, t) - \theta)}} \quad (8)$$

with sigmoid center θ .

3.2. Imputation

In order to use the probabilistic information provided by the soft mask we need to modify the optimization problem described in Eq. 5. We propose to do this by carrying out a *weighted* norm minimization instead:

$$\min_{\mathbf{x}} \{ \|\mathbf{W}\mathbf{A}\mathbf{x} - \mathbf{W}\mathbf{y}\|_2 + \lambda \|\mathbf{x}\|_1 \} \quad (9)$$

with \mathbf{W} a diagonal matrix of which the elements are determined directly by the soft missing data mask \mathbf{M} . The weights on the diagonal are given by using the mask vector representation \mathbf{m} : $\text{diag}(\mathbf{W}) = \mathbf{m}$. After recovering the sparse representation \mathbf{x} , the clean speech is estimated as:

$$\hat{\mathbf{y}} = \mathbf{A}\mathbf{x} \quad (10)$$

Using a binary mask is equivalent to using \mathbf{W} as a row selector picking only those rows of \mathbf{A} and \mathbf{y} that are assumed to contain reliable data. In the case of a soft mask the weights on the diagonal influence the reconstruction error allowed for each spectrographic element.

4. EXPERIMENTS

In order to explore to what extent using soft masks improves recognition accuracy in the sparse imputation framework, we compare digit recognition accuracies obtained with binary masks (generated using Eq. 4) and soft masks (generated using the sigmoid function described in Eq. 8). Two different mask generation techniques are studied, viz. the oracle mask and the harmonicity mask [11]. The oracle mask gives us an upper bound on the recognition accuracy that can be obtained with the sparse imputation method. The harmonicity mask serves as an example of a mask that is obtained when no a priori information about the clean speech signal is available.

4.1. Experimental setup

In this paper, we study a single-digit recognition/classification task using speech data from the AURORA-2 corpus. The single-digit speech data was created by extracting individual digits from the utterances in the AURORA-2 corpus [14] using the segmentation information obtained from a forced alignment of the clean speech utterances with the reference transcription. We used the segments from test set A, which comprises 1 clean and 24 noisy subsets, with four

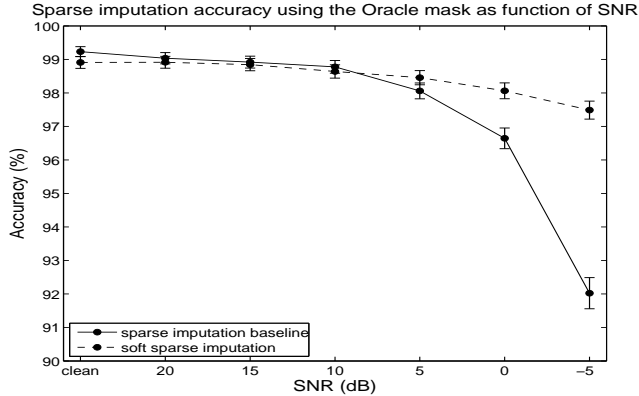


Fig. 1. The figure shows AURORA-2 recognition accuracies for binary and soft sparse imputation using the oracle mask. The range of the vertical axis is [90, 100]. The vertical bars around the data points indicate 95% confidence intervals

noise types (subway, car, babble, exhibition hall) at six SNR values, SNR= 20, 15, 10, 5, 0, -5 dB to evaluate recognition accuracy of the sparse imputation method as a function of SNR and mask type. Recognition accuracies were averaged over the four noise types, resulting in 13176 test digits per SNR condition.

Following [3; 4] we converted the acoustic feature representations to a time normalized representation (a fixed number of acoustic feature frames) using spline interpolation. In our experiment we used 35 time frames per word i.e., the mean number of time frames per word in the training set. Comparison with previously reported recognition accuracies of AURORA-2 clean speech (cf. [7] in which the same decoder was used as in the current study), shows that the recognition accuracies of the clean speech test set have not decreased because of the normalization procedure.

For the computation of the harmonicity mask the noisy speech signal was first decomposed into a harmonic and a random part using the procedure in [11]. Next, the harmonic energy was used as an estimator of the clean speech energy and the random residual as an estimator for the noise energy of the speech signal, for use in Eqs. 4 and 8. Following [11; 3] we have chosen $20 \log_{10}(\theta) = -3$ dB for the oracle mask and -9 dB for the harmonicity mask in Eqs. 4 and 8.

The sparse imputation method was implemented in MATLAB. The l^1 minimization was carried out using the LARS algorithm [15] and implemented as part of the SparseLab toolbox (www.sparselab.stanford.edu).

For recognition we used a MATLAB implementation of the automatic speech recognition system described in [16]. Acoustic feature vectors of the isolated digits consisted of mel frequency log power spectra (23 bands with center frequencies starting at 100 Hz). After imputation of the missing (static) acoustic features, delta and delta-delta coefficients were calculated on these individual digits. During decoding the acoustic representations are converted to PROSPECT features, an alternative feature representation that allows the combination of missing data processing in the spectral domain with the attractive properties of cepstral coefficients [16]. As in [16] we trained 11 whole-word models with 16 states per word, as well as two silence words with 1 and 3 states respectively, using the AURORA-2 clean speech train set.

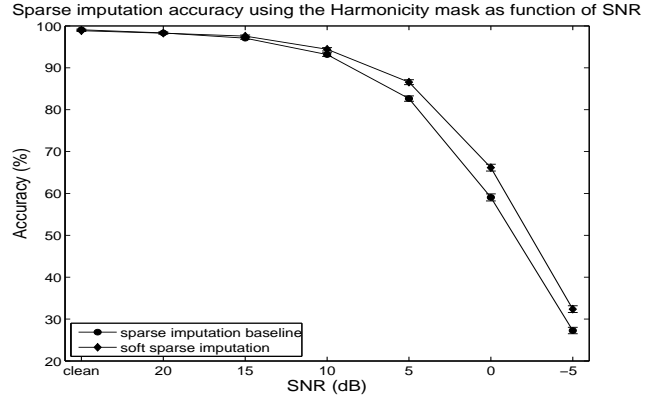


Fig. 2. The figure shows AURORA-2 recognition accuracies for binary and soft sparse imputation using the harmonicity mask. The range of the vertical axis is [20, 100]. The vertical bars around the data points indicate 95% confidence intervals.

4.2. Oracle mask experiment

The oracle mask results shown in Fig. 1 indicate that the sparse imputation method using soft masks achieves higher recognition accuracies at SNRs 0 and -5 dB than when using binary masks. At SNR -5 dB, the soft imputation method has an accuracy of 97.5%, 5.5% higher than the 92% obtained with binary masks. Accuracy of the soft imputation method at SNRs > 0 dB are comparable to those obtained using a binary mask, except for clean speech, which has slightly lower recognition accuracies when using soft masks.

4.3. Harmonicity mask experiment

The harmonicity mask results displayed in Fig. 2 show that our sparse imputation method also achieves higher recognition accuracies with realistic soft masks. The difference in recognition accuracy increases at lower SNRs: At SNR -5 dB, the soft imputation method has an accuracy of 33%, 6% points higher than the 27% obtained by the sparse imputation method using binary masks.

5. DISCUSSION

The results from both experiments indicate that the use of soft masks increases the recognition accuracy obtained with the sparse imputation technique. The recognition results obtained with soft oracle masks (97.5% at SNR -5 dB) indicate that the sparse imputation method can deliver excellent noise robust single digit recognition results, if the mask can be properly estimated.

The improvements in recognition accuracy obtained using soft masks notwithstanding, there remains a substantial gap between the recognition accuracies obtained with oracle (ideal) and harmonicity (estimated) masks at SNRs < 15 dB. As the SNR decreases it becomes increasingly more difficult for the harmonicity mask to distinguish between reliable and unreliable spectral features. The detection of harmonic components, which are always treated as reliable, depends on accurate pitch detection, a process that is volatile at low SNR values. In addition, the harmonicity mask may not always treat noises with a harmonic structure, such as babble noise, correctly. Thus, there is a clear need for developing more powerful mask estimation methods.

The slight and non-significant drop in recognition accuracy of the soft masking technique on clean speech observed when using the oracle mask is due to (small) reconstruction errors in reliable features. When using the binary masks reliable features are copied from the input, while in the current implementation of the soft masking framework all features are replaced by clean speech *estimates*. Since Eq. 5 minimizes the reconstruction error rather than enforcing exact reconstruction, the reconstructed spectrogram might not be exactly identical to the observed spectrogram, even if all features have a high probability of being reliable. This might occasionally lead to recognition errors. The drop in accuracies for clean speech could therefore be avoided by not using clean speech estimates for spectro-temporal elements which have very high probability of being reliable.

The improvement found when using soft masks can be understood from the nature of the minimization carried out in Eqs. 5 and 9. When using a binary mask, cells may be occasionally labeled reliable even though the clean speech energy and noise energy are very close. In this situation the assumption that the reliable cells are good estimators for clean speech may be less than ideal. This situation could be avoided by changing the threshold θ in Eq. 4. However, in [3] we showed that this approach reduces the number of reliable elements in \mathbf{y}_r in Eq. 5 which hurts the imputation as well: If the number of reliable elements in \mathbf{y}_r gets too low, there is not enough information to uniquely determine the sparse representation \mathbf{x} . In the soft sparse imputation framework, however, this situation is handled differently: Features are used regardless whether the speech energy exceeds the noise energy or not. The influence they may exert on the sparse representation that is found, is controlled by weights. For features of which the clean speech energy and the noise energy are very close, the assigned weights will be close to 0.5. As a result, the influence of individual features will be gradually reduced and not abruptly switched on or off as in the binary mask case. Thus the soft mask sparse imputation technique is more robust, especially when noise energy rivals clean speech energy.

The sparse imputation approach presented in this paper can be extended to connected digit and continuous speech recognition. One option that we are investigating is imputation in a sliding time window (cf. [17]).

6. CONCLUSIONS

We have proposed an extension to our previous MDT-technique for binary masks so that it can be applied with soft masks. We have tested the recognition accuracies obtained with the soft masking technique using both ideal (oracle) and estimated masks on single digits extracted from the AURORA-2 corpus. While the results showed there remains a substantial performance gap between oracle and harmonicity mask recognition accuracies, we have demonstrated that the noise robustness of the sparse imputation method is further improved by using soft masks instead of binary masks. Our isolated digit recognition experiments have shown an increase of up to 6% absolute in word recognition accuracy at SNR -5 dB. By using soft masks the influence of mask estimation errors is diminished and the influence of cells more resembling clean speech is increased, leading to increased performance both when oracle masks and when realistic masks are used.

Future work will address more advanced missing data masks based on machine learning techniques. In addition, we are investigating techniques for generalizing the soft mask sparse imputation technique to continuous speech recognition.

Acknowledgments

The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program.

7. REFERENCES

- [1] 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.
- [2] B. Raj, *Reconstruction of incomplete spectrograms for robust speech recognition*, Ph.D. thesis, Carnegie Mellon University, 2000.
- [3] J. Gemmeke and B. Cranen, "Using sparse representations for missing data imputation in noise robust speech recognition," *Proc. of EUSIPCO 2008*, 2008.
- [4] J. Gemmeke and B. Cranen, "Noise reduction through compressed sensing," *Proc. of INTERSPEECH 2008*, 2008.
- [5] J. Barker, L., M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP-2000*, 2000, pp. 373–376.
- [6] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.
- [7] M. Van Segbroeck and H. Van hamme, "Robust speech recognition using missing data techniques in the prospect domain and fuzzy masks," in *Proc. of IEEE ICASSP*, 2008, pp. 4393–4396.
- [8] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.
- [9] E. J. Candes, "Compressive sampling," in *Proc. of the International Congress of Mathematicians*, 2006.
- [10] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.
- [11] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. of IEEE ICASSP*, 2004, vol. 1, pp. 213–216.
- [12] C. Cerisara, S. Demange, and J-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 443–457, 2007.
- [13] Y. Zhang, "When is missing data recoverable?," *CAAM Technical Report TR06-15*, Rice University, Houston, 2006.
- [14] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop, Paris, France*, 2000, pp. 181–188.
- [15] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.
- [16] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. INTERSPEECH-2004*, 2004, pp. 101–104.
- [17] J. Gemmeke and B. Cranen, "Time-continuous sparse imputation," *Technical Report*, <http://arxiv.org/abs/0901.2416>, 2009.