Noise reduction through Compressed Sensing

J. F. Gemmeke, B. Cranen

Dept. of Linguistics, Radboud University, Nijmegen, The Netherlands

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

Abstract

We present an exemplar-based method for noise reduction using missing data imputation: A noise-corrupted word is sparsely represented in an over-complete basis of exemplar (clean) speech signals using only the uncorrupted time-frequency elements of the word. Prior to recognition the parts of the spectrogram dominated by noise are replaced by clean speech estimates obtained by projecting the sparse representation in the basis. Since at low SNRs individual frames may contain few, if any, uncorrupted coefficients, the method tries to exploit all reliable information that is available in a word-length time window. We study the effectiveness of this approach on the Interspeech 2008 Consonant Challenge (VCV) data as well as on AURORA-2 data. Using oracle masks, we obtain obtain accuracies of 36-44% on the VCV data. On AURORA-2 we obtain an accuracy of 91% at SNR -5 dB, compared to 61% using a conventional frame-based approach, clearly illustrating the great potential of the method.

Index Terms: Automatic Speech Recognition, Missing Data Techniques, Compressed Sensing

1. Introduction

Automatic speech recognition (ASR) performance degrades substantially when speech is corrupted by background noises. Missing Data Techniques (MDT) [1, 2] provide a powerful way to mitigate the impact of both stationary and non-stationary noise for moderate Signal-to-Noise (SNR) ratios. 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 energy) and which are unreliable (i.e., dominated by background noise). These reliability estimates are used to treat reliable and unreliable features differently and are referred to as a *spectrographic mask*. This information can for instance be used to replace the unreliable features by clean speech estimates, a process called *imputation* ([3]).

Although recognition accuracy can be improved substantially with MDT, at SNRs ($\leq$ 0 dB) the gain in recognition performance appears generally too small to be of practical use. A possible explanation is that at SNRs $\leq$ 0 dB a substantial number of frames may contain few, if any, reliable features. As the number of reliable coefficients in a frame decreases, it becomes more difficult to safely impute the missing coefficients because in most MDT approaches imputation is performed on a frame by frame (i.e. strictly local) basis. This cannot but degrade recognition performance.

Due to continuity constraints imposed by the speech production system, speech energy is distributed over the time-frequency plane in patches which cannot be of arbitrary small size. Using a (much) wider time window than a single frame might provide a way to exploit this continuity over time for imputing missing data thus avoiding local information scarcity. In this paper we propose a novel, exemplar based, data imputation front-end that tries to take advantage of the dependencies between neighboring frames by using a larger spectro-temporal context. The technique is dubbed *sparse imputation* and is based on work in the emergent field of *Compressed Sensing* [4, 5].

Following an approach similar to [6], we treat entire words as units and represent them by fixed length vectors. We represent unknown words as a linear combination of as few as possible exemplar words in a training database. Work in *Compressed Sensing* has shown that if such a linear combination is *sparse*, the weight vector can be determined using only a small part of the elements of the feature vector representing the unknown item. We exploit this property by using only the features that were considered reliable in the noisy input according to the spectrographic mask. Next, the linear combination of clean exemplar words is used for reconstructing the unreliable coefficients of the noisy words. Finally, the imputed feature vectors are processed by a conventional HMM-based ASR.

In this paper we present a feasibility study to show that the *Compressed Sensing* approach is also potentially beneficial for speech recognition. We explore the effectiveness of our method by applying it to the Interspeech 2008 Consonant Challenge (VCV) data [7]) using two different mask types: the 'oracle' mask [1] and a mask which estimates reliability based on a harmonic decomposition, dubbed harmonicity mask ([8]). Additionally, we compare the recognition performance that we obtained by means of our new, whole word based sparse imputation method with the results from a classical, frame based imputation approach using the AURORA-2 digit recognition task, which also allows us to investigate relative improvement as a function of SNR. We explain why the improvement is larger for oracle than for harmonicity masks.

2. Method

2.1. Speech materials and classification task

Two word recognition tasks were performed. First, we performed intervocalic English consonant recognition using the VCV data (consisting of 1 clean and 6 noisy subsets) which are described in [7]. Second, we carried out a single-digit recognition/classification task using test set A from the AURORA-2 corpus which comprises 1 clean and 24 noisy subsets, with four noise types (subway, car, babble, exhibition hall) at six SNR levels, SNR$= 20, 15, 10, 5, 0, -5$ dB

Due to the constraint that the current implementation of our exemplar-based imputation technique operates on whole-

[1] Oracle masks are masks in which reliability decisions are based on exact knowledge about the extent to which each time-frequency cell is dominated by either noise or speech

word units, we constructed training and test sets in which the words were surrounded by only a minimal amount of "silence". The noisy single-digit data were created by extracting individual digits from the utterances in the AURORA-2 corpus using the segmentation obtained from a forced alignment of the clean speech utterances with their reference transcription. From the VCV utterances we removed leading and trailing silences using the offset data provided.

2.2. Speech decoders

For recognition of the VCV words, we use the baseline HTK decoder described in [7]. Imputation was carried out on mel frequency log power spectra (FBANK_E), after which the reconstructed spectra were converted to standard mel cepstral coefficients (MFCC_Z_D_A_E) prior to recognition. With separate HMMs for initial and final vowels we used 30 3-state monophone models (24 consonants plus $2 \times 3 = 6$ vowels) consisting of 24 Gaussian mixtures.

For recognition of the AURORA-2 digits and comparison with a frame-based imputation method, we used a MATLAB implementation of the missing data recognition system described in [3]. Acoustic feature vectors consisted of mel frequency log power spectra, which are then converted to PROSPECT features [3]. We trained 11 whole-word models with 16 states per word using clean speech. In the baseline recognizer, unreliable features are replaced by estimated values using maximum likelihood per Gaussian-based imputation [3]. In the sparse imputation system, the spectrographic data are first cleaned with the method described below after which they are recognized using the baseline decoder with a spectrographic mask that considers every time-frequency cell as reliable.

2.3. Fixed length vector representation of words

Since the method described in the following sections works on observation vectors of fixed size, we converted the acoustic feature representations to a time normalized version (a fixed number of acoustic feature frames). The re-sampling was done by applying spline interpolation to the spectrographic representation and then re-sampling the 23 mel frequency log-energy coefficients individually such that a fixed number of acoustic vectors per word resulted. In our experiment we used 60 time frames per word for the VCV-data and 35 time frames per word for AURORA-2 digit data (i.e., the mean number of time frames per word in the training sets). For the sparse imputation technique the time-frames were then concatenated to form a single, fixed length observation vector. The baseline recognizer used the same, time-normalized spectra. A pilot study revealed that the recognition accuracies did not decrease after applying the resampling procedure.

2.4. Sparse representation

Following [6] we consider a test word $\boldsymbol{y}$ to be a linear combination of exemplar words $\boldsymbol{w}_n$, where the index n denotes a specific exemplar word ($1 \leq n \leq N$) and N the total number of exemplar words in the training corpus. We write:

$$\boldsymbol{y} = \sum_{n=1}^{N} \alpha_n \boldsymbol{w}_n$$

with weights $\alpha_n \in \mathbb{R}$.

Denoting the k^{th} vector element of $\boldsymbol{w}_n$ by w_n^k, and recalling that each word in the example set is represented by a vector with dimensionality K, we write our set of N exemplar words as a matrix A with dimensionality $K \times N$:

$$A = \begin{pmatrix} w_1^1 & w_2^1 & \dots & w_{N-1}^1 & w_N^1 \\ w_1^2 & w_2^2 & \dots & w_{N-1}^2 & w_N^2 \\ \vdots & & & \vdots & \\ w_1^K & w_2^K & \dots & w_{N-1}^K & w_N^K \end{pmatrix}$$

We can now express any word $\boldsymbol{y}$ as

$$\boldsymbol{y} = A\boldsymbol{x} \tag{1}$$

with $\boldsymbol{x} = [\alpha_1 \alpha_2 \dots \alpha_{N-1} \alpha_N]^T$ an N-dimensional vector that will be sparsely represented in A (i.e., most coefficients α are zero).

For the VCV-data the set of exemplar words comprised the entire training set: $N = 6,664$. The number of exemplar words that could be taken from the clean train set of AURORA-2 is $27,748$ words. In order to make classification times practical, we did not use all of them and reduced the number of columns N in A by randomly selecting a subset of the training set. We used $N = 4,000$ (yielding an average of about 360 tokens for each of the 11 digit words), which in a pilot study was found to give nearly the same recognition accuracy as using the full set.

2.5. l^1minimization

In order to determine the sparse vector $\boldsymbol{x}$ representing a word $\boldsymbol{y}$, we need to solve the system of linear equations of Eq. 1. Typically, the number of exemplar words will be much larger than the dimensionality of the feature representation of the vowels ($K \ll N$). Thus, the system of linear equations in Eq. 1 is *under-determined* and has, generally speaking, no unique solution. Research in the field of *Compressed Sensing* [4, 5], however, has shown that if $\boldsymbol{x}$ is *sparse*, $\boldsymbol{x}$ *can* be determined by solving:

$$min||\boldsymbol{x}||_1 \text{ subject to } \boldsymbol{y} = A\boldsymbol{x} \tag{2}$$

with $||.||_1$ the l^1norm (i.e. minimization of the sum of absolute values of elements) which serves as an approximation of the l^0 norm (i.e., the number of nonzero elements). The approximation is necessary since minimizing the l^0 norm is an NP-hard combinatorial problem [9], while l^1minimization can be done efficiently in polynomial time. Since in practice it may be impossible to express a word exactly as a superposition of exemplar words, we use a noise robust version of Eq. 2 (cf. [10]):

$$min||\boldsymbol{x}||_1 \text{ subject to } ||\boldsymbol{y} - A\boldsymbol{x}||_2 \leq \epsilon \tag{3}$$

with a small constant ϵ such that the error $\boldsymbol{e}$ satisfies $||e||_2 < \epsilon$.

2.6. Spectrographic mask

A spectrographic mask M is a matrix with the same dimensions as the spectrographic representation of a word. We used two different masks to describe the reliability of time-frequency cells in the spectrographic representation of a word: 1) an oracle mask and 2) an estimated mask: The harmonicity mask [8]. For the computation of the harmonicity mask the noisy speech signal is first decomposed into a harmonic and a random part. Next, a time-frequency cell is defined as unreliable if the energy of the random part exceeds that of the harmonic part.

For use in the sparse imputation framework, we reshape the mask M to form a vector $\boldsymbol{m}$ by concatenating subsequent time frames as described in 2.3.

Table 1: VCV *consonant recognition accuracy.*

	test set						
method	1	2	3	4	5	6	7
baseline	86.7	7.6	5.0	5.5	3.9	8.9	5.5
oracle	-	44.8	43.0	36.0	39.6	41.7	40.9
harmonicity	-	7.3	9.9	7.8	9.6	7.0	6.0

2.7. Sparse imputation

Given an observation vector $\boldsymbol{y}$ (representing an entire word), we denote $\boldsymbol{y}_r$ consisting of the reliable coefficients of $\boldsymbol{y}$. These are the elements for which the corresponding elements of mask vector $\boldsymbol{m}$ are equal to one. Similarly, we denote the unreliable coefficients of $\boldsymbol{y}$ (for which the corresponding elements of mask vector $\boldsymbol{m}$ are equal to zero) by $\boldsymbol{y}_u$. Without loss of generality we reorder $\boldsymbol{y}$ and A as in [11] so that we can write:

$$\begin{bmatrix} \boldsymbol{y}_r \\ \boldsymbol{y}_u \end{bmatrix} = \begin{pmatrix} A_r \\ A_u \end{pmatrix} \boldsymbol{x} \tag{4}$$

with A_r and A_u pertaining to the rows of A indicated by the reliable and unreliable coefficients in $\boldsymbol{y}$. Since we consider the values of the $\boldsymbol{y}_r$ to be valid representatives of clean speech, we solve Eq. 3 using only $\boldsymbol{y}_r$ instead of $\boldsymbol{y}$. After obtaining the sparse representation $\boldsymbol{x}$ we use this vector to impute clean estimates $\boldsymbol{y}_i$ for the unreliable coefficients $\boldsymbol{y}_u$ using the support of $\boldsymbol{x}$ in the basis A_u:

$$\hat{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{y}_r \\ \boldsymbol{y}_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_r \\ A_u \boldsymbol{x} \end{bmatrix} \tag{5}$$

yielding a new observation vector $\hat{\boldsymbol{y}}$. In order to perform recognition we restore the original ordering and reshape $\hat{\boldsymbol{y}}$ of Eq. 5 to a time framed spectrographic representation.

Obviously, no restoration of the unreliable coefficients in $\boldsymbol{y}$ is possible if there are no reliable coefficients to base the estimation on. If we denote the number of reliable coefficients in $\boldsymbol{y}$ by $K_r = \dim(\boldsymbol{y}_r)$, in practice, effective restoration of the unreliable coefficients will be difficult below some threshold $K_r < \delta$. For several reasons it is impossible to determine δ theoretically. First, it is impossible to predict the sparsity of $\boldsymbol{x}$ obtained in Eq. 4. Second, δ will depend on the structure of the spectrographic mask and therefore on the underlying speech signal and environmental noise: features cannot be restored if the remaining reliable features do not carry sufficient information to predict the value of the unreliable ones. Accepting that faulty restorations of $\boldsymbol{y}$ are unavoidable when the number of reliable features drops below the unknown threshold K_r, we decided to always perform sparse imputation except when $K_r = 0$.

3. Results

Table 1 shows that the accuracy on clean speech (86.7%) of our baseline recognizer is slightly lower than the accuracies reported in [7] (88.5%). This is probably caused by the use of log energy only rather than the combination of log energy and zeroth cepstral coefficient. As can be inferred from Table 1, using a harmonicity mask to control the reconstruction process does not yield consistent improvements. In fact, in some cases the accuracies even deteriorate. The recognition accuracies obtained with the sparse imputation method using an oracle mask range from 36-45%, which amounts to a very substantial improvement over the baseline.

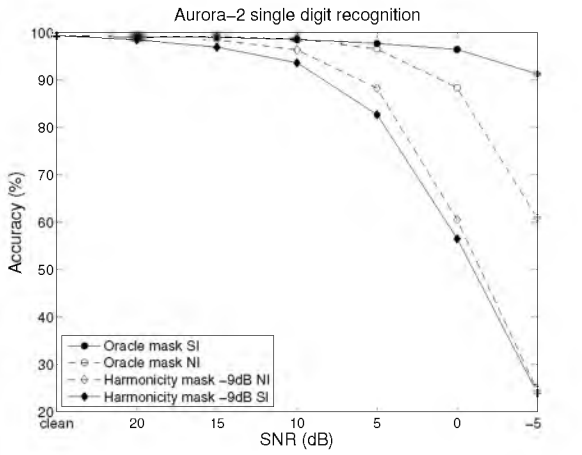

Figure 1: AURORA-2 *single digit recognition accuracy. The figure shows results for both normal Missing Data Imputation (NI) as well as sparse imputation (SI) for the oracle mask and the harmonicity mask respectively.*

Figure 1 shows the recognition accuracies (i.e., averages over the four noise types in test set A) on the single-digit task (AURORA-2). The figure shows recognition accuracies using the baseline missing data recognizer, as well as the sparse imputation method. With the estimated (harmonicity) mask sparse imputation does not improve and even slightly degrades performance. Also, for both imputation methods recognition accuracies at SNRs $\leq$ 0 dB, the SNRs in the VCV data, decrease rapidly. With the oracle masks, however, the whole word sparse imputation technique substantially outperforms the frame based imputation technique at low SNRs: Recognition accuracy at SNR -5 dB is 91% compared to 61%.

4. Discussion

4.1. Reconstruction using estimated masks

The small but consistent differences in recognition accuracy between the imputation methods observed on AURORA-2 are most likely due to the fact that only in the frame-based imputation method *bounded imputation* is employed: The estimated coefficients are bounded by the observation energy assuming additive noise, while the sparse imputation makes no such assumption. Apparently, bounded imputation is better suited to compensate for the little amount of reliable information marked as such by the harmonicity mask. More important, however, is the fact that at low SNRs the use of a wider context does not help to improve recognition performance compared to the frame based approach. It was shown in [12] that the harmonicity mask labels significantly fewer time-frequency cells reliable when compared to the oracle mask: Our results therefore indicate that at low SNRs this specific mask estimation method simply does not generate enough reliable coefficients (according to the criterion $K_r > \delta$ cf. section 2.7) to allow successful reconstruction of the unreliable ones. Additionally, more speech-like noises like present in the VCV test sets inevitably causes mislabeling since the mask estimation method uses harmonic decomposition to label cells reliable. If the corrupting noise is a competing speaker, such as in test set 2, the harmonicity mask is likely to label noisy spectrographic regions reliable and possibly even label target

speech regions unreliable, causing the two lower-than-baseline accuracies.

Obviously more advanced mask estimation techniques are required (cf. [13] for a comprehensive survey) if the full potential of missing data techniques at low SNRs is to be exploited.

4.2. Reconstruction using oracle masks

As might be expected, recognition accuracies using the oracle mask show significant improvement over baseline results. For the VCV-data accuracies obtained with the the sparse imputation method range from 36-45% indicating that a substantial amount, but by far not all corrupted features can be reconstructed with our method. This is not unexpected since even native listeners display a decrease in recognition accuracy of 14-27% [7]. Still, the fact that the drop in accuracy of human listeners is lower than in our method indicates that our reliable features contain not enough information: The sparse imputation technique, and imputation based MDT in general, can only reconstruct unreliable coefficients if the reliable coefficients alone are sufficient to discriminate between the alternatives in the vocabulary. For AURORA-2 this condition holds much better since even at an SNR of -5dB an accuracy of 91% is attainable.

The difference in recognition performance between the VCV and AURORA-2 task using oracle masks can be understood from the nature of the recognition tasks: consonants typically have lower energy than vowels. At SNRs $\leq$ 0 dB, in general only the vowels will have spectrographic regions dominating the noise. With the digits in AURORA-2 the vowels will often contain sufficient information to predict the surrounding consonants, if only because not every consonant occurs in all possible vowel contexts. In this respect, the VCV set is much less predictable: Each consonant can occur all vowel contexts. As a consequence, the vowels themselves will be of little use to discriminate between words, and only the VC and CV transition regions may help to discriminate between consonants.

Unfortunately we did not have a MDT frame-based decoder for the VCV-data at our disposal. From comparing the recognition accuracy of 91% at SNR -5 dB on AURORA-2 to the 61% recognition accuracy of the baseline MDT-decoder, an increase of 30% absolute, it is clear that a frame based approach is only partially successful in exploiting redundant information. The results on AURORA-2 suggest that using a wider time context when doing imputation is beneficial to avoid local information scarcity and that sparse imputation can be a powerfull method to utilize this wider time context.

4.3. Future work

The success of the sparse imputation method using oracle masks makes further research desirable. In order to be used as a general front-end for ASR systems the method needs to be extended to work in a continuous time setting. One way to do this is to use a sliding time-window using several neighboring time frames as generally used in frame-based Support Vector Machine and Neural Net classification tasks. Another approach is used in [14], in which a larger basis is defined using time-shifted copies of the original basis. The practical applicability and computational feasibility of either method is left as future work.

5. Conclusions

We introduced a missing data imputation front-end which works by finding a sparse representation of the noisy speech signal, using only the information of the speech signal labeled reliable by a spectrographic mask. The sparse representation is found by expressing entire words as a linear combination of exemplar words. The sparse representation is then used to estimate the missing (unreliable) coefficients of the speech signal after which classic speech recognition can take place. Results on both the Interspeech Consonant Challenge data and the AURORA-2 digits underline that recognition accuracy depends on the success with which the spectrographic mask can be estimated and to what extent the reliable features carry information about the unreliable ones. Experiments on AURORA-2 using an oracle mask, however, also clearly show the potential of the presented method: A recognition accuracy of 91% at SNR = -5 dB is obtained, an increase of 30% absolute over a state-of-the art missing data speech recognizer using frame by frame imputation. This shows that even at very low SNRs enough information about the speech signal may be preserved to successfully perform imputation solely on the basis of reliable time-frequency cells provided enough time-context is used.

6. Acknowledgments

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

7. References

[1] B. Raj, R. Singh, and R. Stern, "Inference of missing spectrographic features for robust automatic speech recognition," in *Proceedings International Conference on Spoken Language Processing*, 1998, pp. 1491–1494.

[2] M. Cooke, P. Green, L. Josifovksi, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[3] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Interspeech-2004*, 2004, pp. 101–104.

[4] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] E. J. Candes, "Compressive sampling," in *Proceedings of the International Congress of Mathematicians*, 2006.

[6] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Feature selection in face recognition: A sparse representation perspective," *submitted to IEEE Transactions Pattern Analysis and Machine Intelligence*, August 2007.

[7] M. Cooke and O. Scharenborg, "The interspeech 2008 consonant challenge," *submitted to Interspeech 2008*, 2008.

[8] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proceedings of IEEE ICASSP*, vol. 1, 2004, pp. 213–216.

[9] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, no. 2, pp. 227–234, 1995.

[10] D. L. Donoho, "For most large underdetermined systems of equations, the minimal l1-norm near-solution approximates the sparsest near-solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 7, pp. 907–934, 2006.

[11] Y. Zhang, "When is missing data recoverable?" *Technical Report*, 2006.

[12] J. Gemmeke, B. Cranen, and L. ten Bosch, "On the relation between statistical properties of spectrographic masks and recognition accuracy," in *Proceeding (599) Signal Processing, Pattern Recognition, and Applications - 2008*, 2008, pp. 200–206.

[13] 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.

[14] M. Mœrup and M. N. Schmidt, "Shift invariant sparse coding of image and music data," *Submitted to Journal of Machine Learning Research*, 2008.