USING SPARSE REPRESENTATIONS FOR MISSING DATA IMPUTATION IN NOISE ROBUST SPEECH RECOGNITION

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

Noise robustness of automatic speech recognition benefits from using missing data imputation: Prior to recognition the parts of the spectrogram dominated by noise are replaced by clean speech estimates. Especially at low SNRs each frame contains at best only a few uncorrupted coefficients. This makes frame-by-frame restoration of corrupted feature vectors error-prone, and recognition accuracy will mostly be sub-optimal. In this paper we present a novel imputation technique working on entire words. A word is sparsely represented in an overcomplete basis of exemplar (clean) speech signals using only the uncorrupted time-frequency elements of the word. The corrupted elements are replaced by estimates obtained by projecting the sparse representation in the basis. We achieve recognition accuracies of 92% at SNR -5 dB using oracle masks on AURORA-2 as compared to 61% using a conventional frame-based approach. The performance obtained with estimated masks can be directly related to the proportion of correctly identified uncorrupted coefficients.

1. INTRODUCTION

Automatic speech recognition (ASR) performance degrades substantially when speech is corrupted by background noises that were not seen during training. Missing Data Techniques (MDT) [1, 2] constitute a powerful way to mitigate the impact of both stationary and non-stationary noise. 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). By storing these reliability estimates in a so called *spectrographic mask*, this information can be used to treat reliable features differently from unreliable ones during decoding: Either the unreliable features can be replaced by clean speech estimates (feature vector imputation [3, 4]), or the decoder can be modified so that it can deal with the unreliable input data directly (marginalization [2]). In this paper we will only deal with imputation.

Many different techniques have been proposed to estimate spectrographic masks (cf. [5] for a comprehensive survey), ranging from SNR based estimators [6] to methods that focus on speech characteristics, e.g. harmonicity based SNR estimation [7], mask estimation by means of Bayesian classifiers [8] and masks composed of spectro-temporal fragments [9]. From experiments with signals that have been constructed by artificially adding noise to clean speech, it is well-known that estimated masks yield inferior recognition accuracies compared to an 'oracle' mask [1]. As explained in [8] the gain in recognition accuracy obtainable with a given estimated mask is hard to predict from a direct comparison with the oracle mask.

Due to continuity constraints implicitly imposed by the speech production system, speech energy is not randomly distributed over the time-frequency plane and as a consequence, a realistic mask will in general not have arbitrary granularity. Unfortunately, in most ASR approaches imputation takes place on a frame by frame (i.e. strictly local) basis [2]. This hampers exploiting the continuity over time of the mask and the speech signal. Particularly at low SNRs ($\leq$ 0 dB), it may happen that only few, if any, elements in a single acoustic vector are labeled reliable. The more features become unreliable, the more serious the risk that an individual frame contains too little information for properly dealing with unreliable coefficients. This effect will be aggravated if some of the coefficients were incorrectly labeled reliable by the used mask estimation procedure. As a consequence the acoustic scores of such frames will be affected and if there are too many frames with few isolated reliable features, recognition accuracy is bound to suffer significantly.

In this paper we propose a novel data imputation technique that does take into account a larger spectro-temporal context. The novel technique is dubbed *sparse imputation* and is based on the work in *Compressed Sensing* [10, 11]. The technique is illustrated by means of experiments using the AURORA-2 digit recognition task.

Similar to the AURORA-backend that uses whole word models we treat noisy digits as units and represent them by fixed length vectors. Following the same approach as in [12], we represent unknown digits as a linear combination of as few as possible exemplar digits taken from the clean speech part of the database. In building the optimal linear combination to represent noisy digits, we only take into consideration the features that were considered reliable in the noisy input. Next, the selected clean exemplar digits are used for reconstructing the unreliable coefficients of the noisy digits. Finally,the imputed feature vectors are processed by a conventional HMM-based ASR assuming that all features are reliable.

We investigate the performance of sparse imputation by comparing recognition accuracies with the results of a classical frame based imputation approach. Since the performance of any imputation technique hinges on the quality of the spectrographic mask, we investigate sparse imputation for two types of masks: 1) an oracle mask and 2) an estimated spectrographic mask in the form of a harmonicity mask [7]. In estimated masks the estimates can be biased towards higher false accept or higher false reject rates. Therefore, we investigate the performance with the harmonicity

[1] Oracle masks are masks in which reliability decisions are based on exact knowledge (e.g. not available in practical settings) about the extent to which each time-frequency cell is dominated by either noise or speech.

[2] In fact, the fragment decoder approach [9] in which decoder knowledge may affect the eventual choice of mask is the only exception we are aware of.

mask for three different settings, resulting in three different proportions of features considered as reliable.

2. METHOD

2.1 Speech material and classification task

In order to be able to focus on key factors that govern the success of our new data imputation technique, without being hampered by complications associated with segmentation issues, 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 [13] by segmenting the digit words in each utterance using the segmentation 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 noise types (subway, car, babble, exhibition hall) at six SNR values, SNR= $20, 15, 10, 5, 0, -5$ dB to evaluate recognition accuracy as a function of imputation method, SNR and bias in the harmonicity masks.

2.2 Speech decoder

For the baseline system, we used a MATLAB implementation of a missing data recognition system described in [4]. Acoustic feature vectors consisted of mel frequency log power spectra (23 bands with center frequencies starting at 100 Hz, as well as first and second derivatives, i.e. 69 coefficients in total), which are then converted to 69 PROSPECT features [4]. Unreliable features are replaced by estimated values using maximum likelihood per Gaussian-based imputation [4]. As in [4] we trained 11 whole-word models with 16 states per word, as well as two silence words with 1 and 3 states respectively, using clean speech. The acoustic representations obtained with our sparse imputation method were recognized using this same decoder, using a spectrographic mask that considers every time-frequency cell as reliable (thus performing no additional missing data imputation). Prior to performing recognition delta and delta-delta coefficients were calculated on the restored acoustic features.

2.3 Fixed length vector representation of digits

Since the digits have different durations, and 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 representation (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 35 time frames per word i.e., the mean number of time frames per word in the training set. For the sparse imputation technique the time-frames were then concatenated to form a single fixed length observation vector. Thus each digit was represented by a $K = 23 \times 35 = 805$ dimensional vector $\boldsymbol{y}$.

A pilot study revealed that the recognition accuracies did not decrease after applying the resampling procedure. This can be understood from the nature of the back-end: while digit length may be somewhat discriminative, it is known to hardly affect the recognition results of an HMM-based decoder

2.4 Sparse representation

Following [12] we consider a test digit $\boldsymbol{y}$ to be a linear combination of exemplar digits $\boldsymbol{d}_n$, where the index n denotes a specific exemplar digit ($1 \leq n \leq N$) and N the number of exemplar digits. We write:

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

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

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

$$A = \begin{pmatrix} d_1^1 \; d_2^1 & \dots & d_{N-1}^1 \; d_N^1 \\ d_1^2 \; d_2^2 & \dots & d_{N-1}^2 \; d_N^2 \\ \vdots & & \vdots \\ d_1^K \; d_2^K & \dots & d_{N-1}^K \; d_N^K \end{pmatrix}$$

We can now express any digit $\boldsymbol{y}$ 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).

The exemplar digits were taken from the clean train set of AURORA-2 which consists of $N = 27748$ digits. However, the number of columns N in A had to be reduced in order to make classification times practical. Thus, a subset of the training set was randomly selected, i.e., no attempt was made to represent genders, regional background or digits uniformly. A pilot study showed that any basis size larger than $N = 4000$ columns yielded equivalent recognition accuracies. In this paper, we will therefore be using $N = 4000$.

2.5 l^1minimization

In order to utilize the sparse vector $\boldsymbol{x}$ to represent a digit $\boldsymbol{y}$ we need to solve the system of linear equations of Eq. 1. Typically, the number of exemplar digits 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 *underdetermined* and has no unique solution.

Research in the field of *compressed sensing* [10, 11] has shown that if $\boldsymbol{x}$ is sparse, $\boldsymbol{x}$ can be recovered *exactly* by solving:

$$\min \|\boldsymbol{x}\|_0 \text{ subject to } \boldsymbol{y} = A\boldsymbol{x}$$

with $\|.\|_0$ the l^0 norm (i.e., the number of nonzero elements). Unfortunately, this combinatorial problem is NP-hard [14] and therefore infeasible in practical applications. However, it has been shown that $\boldsymbol{x}$ can be recovered with high probability [15] by solving:

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

This l^1minimization problem can be cast as a least squares problem with a l^1penalty also referred to as the LASSO [16]:

$$\min \|A\boldsymbol{x} - y\|_2 + \lambda \|\boldsymbol{x}\|_1 \tag{2}$$

with a regularization parameter λ and a non-negativity constraint on $\boldsymbol{x}$.

2.6 Spectrographic mask

A spectrographic mask is a matrix with the same dimensions as the spectrographic representation of a digit. After the resampling procedure described in Section 2.3 its size is $I \times J$ with $I = 23$ the number of frequency bands, and $J = 35$ the number of time frames. We used two different masks to describe the reliability of time-frequency cells in the spectrographic representation of a digit: 1) an oracle mask and 2) a harmonicity mask [7].

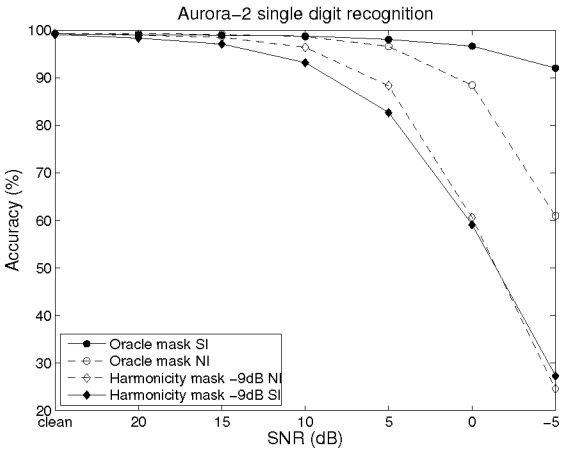

Figure 1: AURORA-2 single digit recognition accuracy. The figure shows results for both normal Missing Data Imputation (NI) as well as sparse imputation (SI) for the oracle mask and the harmonicity mask.

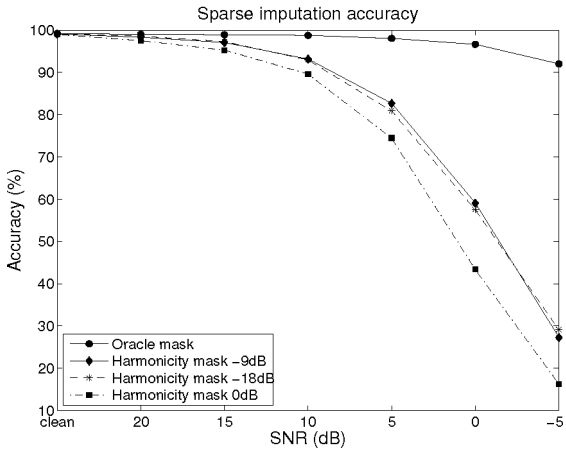

Figure 2: AURORA-2 single digit recognition accuracy. The figure shows results for sparse imputation for the oracle mask and a harmonicity mask at three threshold levels -18, -9 and 0 dB.

The oracle mask was computed on resampled spectrographic representations of the noise N and clean speech S as follows:

$$M(i,j) = \begin{cases} 1 \stackrel{def}{=} \text{reliable} & S(i,j) \geq (N(i,j) - \theta) \\ 0 \stackrel{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \quad (3)$$

with frequency band i ($1 \leq i \leq I$) and time frame j ($1 \leq i \leq J$). We used a fixed threshold $\theta = 3\,\text{dB}$.

For the computation of the harmonicity mask the noisy speech signal is first decomposed into a harmonic and a random part using the procedure in [7]. Next, the local energy of speech and noise are estimated by thresholding the ratio between the harmonic and random part analogously to Eq.3. In [7] it was determined that a threshold of $\theta = -9\,\text{dB}$ was optimal for AURORA-2. Since this threshold value influences the number of time-frequency cells labelled reliable as well as the number of cells incorrectly labeled reliable, we experimented with a large number of threshold values in the range $[0, 18]$. In this paper we show illustrative results for three different thresholds, viz. 0, -9 and $-18\,\text{dB}$. The harmonicity mask is created directly from the raw acoustic data. In order to obtain a spectrographic mask with proper time normalization we therefore applied the resampling procedure described in Section 2.3 directly on the harmonicity mask. Next, we applied thresholding to convert the resampled mask to a binary mask.

For use in the sparse imputation framework, we reshape the mask M to form a $K = 805$-dimensional vector $\boldsymbol{m}$ by concatenating subsequent time frames as described in 2.3. Since the baseline MDT decoder employs delta and delta-delta coefficients imputation, we construct a spectrographic mask for these coefficients using the procedure described in [17].

2.7 Sparse imputation

Given an observation vector $\boldsymbol{y}$ (representing an entire digit), we denote $\boldsymbol{y}_r$ as 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 [18] so that we can write:

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

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 dominated by clean speech, we solve Eq. 2 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:

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

yielding a new observation vector $\hat{\boldsymbol{y}}$. We denote the number of reliable coefficients in $\boldsymbol{y}$ by $K_r = \dim(\boldsymbol{y}_r)$. Obviously, no restoration of the unreliable coefficients in $\boldsymbol{y}$ is possible if $K_r = 0$. In practice, restoration of the unreliable coefficients will be unlikely below some threshold $K_r < \delta$. However, while for some problems the value of δ can be theoretically derived [10, 11, 18], it is not trivial to estimate bounds on the value of δ, for example because we cannot predict the sparsity of $\boldsymbol{x}$ obtained in Eq. 4. Hence, in our implementation, we do not perform sparse imputation if $K_r = 0$ but otherwise ignore the possible unlikeliness of the successful restoration of $\boldsymbol{y}$.

In order to perform recognition we restore the original ordering and reshape $\hat{\boldsymbol{y}}$ of Eq. 5 to a spectrographic representation with dimensions 23×35. The method was implemented in MATLAB. The l^1minimization was carried out using the `SolveLasso` package described in [19] and implemented as part of the `SparseLab` toolbox which can be obtained from www.sparselab.stanford.edu.

3. RESULTS

Figure 1 shows the recognition accuracy on the AURORA-2 corpus single-digit task. The accuracies reported here are the averages obtained for the four noise types in test set A. The results show recognition accuracies using the oracle mask and the estimated harmonicity mask (with $\theta = -9\,\text{dB}$ as described in [7]) for the baseline missing data recognizer, as well as the sparse imputation front-end. For the low SNRs and

the oracle masks the sparse imputation technique substantially outperforms the baseline imputation technique, with a recognition accuracies of 92% and 61% at SNR=−5 dB. Accuracies using an estimated (harmonicity) mask with sparse imputation at higher SNR's (> 0 dB) are lower than when doing standard imputation (at most 6% at SNR=5 dB). At SNR=0 dB the results are competitive while at SNR=−5 dB the sparse imputation technique performs better than the baseline.

Figure 2 shows the recognition accuracies of different thresholds for the harmonicity mask when used in combination with the sparse imputation frontend. The best overall accuracies are obtained using the -9 dB threshold, while a lower (-18 dB) threshold value results in slightly better performance at SNR=−5 dB; a higher (0 dB) threshold value affects recognition accuracies for all SNRs below 20 dB.

Figure 3 shows the percentage of reliable time-frequency cells in a spectrographic mask according to the mask estimation procedures. The oracle mask classifies the largest proportion of time-frequency cells as reliable, followed by the harmonicity mask at threshold −18 dB. Lower numbers of reliable time-frequency cells are obtained at thresholds levels −9 and 0 dB. The percentage of reliable cells mostly linear with respect to the SNR, except for the slight asymptotic behavior at SNRs below zero. Additionally, Figure 3 shows the number of unreliable time-frequency cells incorrectly labeled reliable (dubbed false reliables), expressed as percentage of the number of reliable cells, using the oracle mask as golden standard. The figure shows that the highest percentage of false reliables is obtained at threshold value −18 dB, followed by −9 and 0 dB.

4. DISCUSSION

The recognition accuracy of the sparse imputation method with the oracle mask, 92% at SNR=−5 dB shows that the speech signal contains enough information to restore the unreliable time-frequency cells, even at negative SNRs. Comparing this to the 61% recognition accuracy of the baseline decoder, it is clear that this information is not fully employed when doing imputation on a frame-by-frame basis. The success of the sparse imputation technique suggests that in general the time-frequency cells marked as reliable with the oracle mask suffice for finding a sparse representation in the clean example digits that allows us to reconstruct the features marked unreliable. The drop in accuracy at lower SNRs (although only from 100% to 92%) is mainly due to digits which have very few, if any, reliable cells in the entire mask. This corresponds to the drop in recognition performance of human subjects at negative SNRs [20], probably because of the same reason: not enough reliable information is left.

Using the sparse imputation method with the harmonicity mask, an estimated mask, we obtain recognition accuracies lower than the baseline imputation method at SNRs ≥ 0 dB. A closer look at Fig. 3 reveals that this may be due to the reduced number of reliable features. For example, the percentage of reliable cells (the underdeterminedness) of the harmonicity mask with threshold -9 dB at SNR=10 dB is roughly equal to the percentage found at SNR=0 dB of the oracle mask. At the same time, the recognition accuracy of that harmonicity mask at SNR=10 dB is equal to accuracies obtained with the oracle mask at SNR=0 dB. This same relation between percentage of reliable cells and recognition accuracy across different masking methods is found at other SNR values. It seems likely that there are simply not enough reliable coefficients left ($K_r < \delta$) at the threshold of $\theta = -9\,\text{dB}$ resulting in the low accuracies. However, while lowering this threshold of the harmonicity mask increases the number of cells labeled as reliable this leads to slightly lower recognition accuracies as shown in Fig. 2. This is due to an increase in labelling errors: cells labeled reliable while being unreliable (false reliables). The dependency of these errors as function of SNR and threshold value is also shown in Fig. 3. These unreliable cells introduce errors in the estimation of the sparse representation, in turn leading to imputation errors. The opposite effect, reducing the number of false reliables by calling less cells reliable through higher threshold (0 dB) also has an adverse effect on recognition accuracy. It is obvious that for a given mask technique there is a tradeoff between the number of reliable cells on the one hand and the number of false reliables on the other hand. In practice, finding the optimum between true and false reliables will require an iterative search. It interesting however, that at SNR=−5 dB the sparse imputation method outperforms the baseline method using the estimated mask. This suggests that while the sparse imputation method suffers more from either a reduced number of reliable features or a high amount of false reliables at SNRs ≥ 0 than the baseline method, this behavior is reversed at low SNRs.

In the classical frame-by-frame missing data framework the differences in recognition accuracy between oracle and estimated mask cannot be expressed simply as a function of the number of differing time-frequency cells [8]. This is due to a non-uniform importance of reliable frequency cells in the spectrographic mask. In the current sparse imputation framework this effect is reduced thanks to the wide time context: our results seem to indicate that study of the mask underdeterminedness and the number of false reliables with respect to the oracle mask can be predictive for the expected performance. Additionally, the excellent recognition accuracies obtained using an oracle mask indicate that much higher accuracies can be obtained when more advanced mask estimation methods are combined with an imputation method that uses a wider context.

5. FUTURE WORK

The current implementation of the sparse imputation technique only works with fixed length feature representations. In order to be be used as a general front-end for ASR systems the method needs to be extended to work in a continuous time setting. A possible approach would be to use a sliding (overlapping) time-window using several neighboring time frames as generally used in frame-based Support Vector Machine and Neural Net classification tasks. The basis is then formed by a random sample of the clean speech training database using fixed length time-windows. While the computational complexity of such an approach is larger than for the fixed-length representations presented in this work, it is only linear in the number of overlapping frames.

6. CONCLUSIONS

We introduced a missing data imputation method which works by finding a sparse representation of the noisy speech signal, using only the reliable information of the speech signal as labeled 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 the missing (unreliable) coefficients of the speech signal after which classic speech recognition can take place. The recognition accuracy of 92% at SNR=-5 dB obtained using an oracle mask, an increase of 31% percent absolute over a state-of-the art missing data speech recognizer using frame by frame imputation, showed that even at very low SNRs enough information about the speech signal is preserved to successfully perform imputation solely on the basis of reliable time-frequency cells provided enough time-context is used.

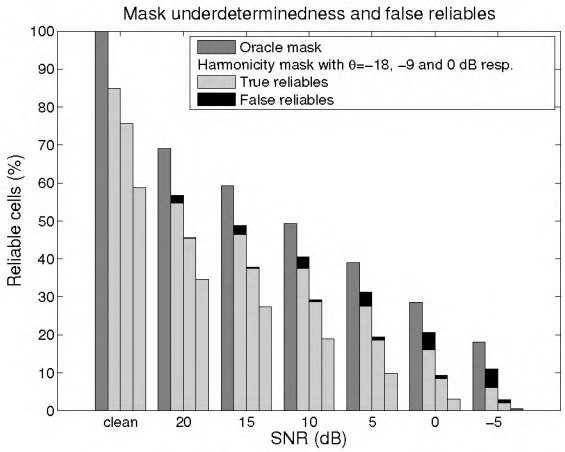

Figure 3: Percentage of reliable time-frequency cells . The figure shows results for the oracle mask as well as three threshold (0, -9 and -18) values for harmonicity masks. Additionally, the figure shows the percentage of false reliables in the harmonicity mask: the number of time-frequency cells labelled reliable while being unreliable according to the oracle mask.

.

The sparse imputation method using an estimated harmonicity mask also performed better than baseline at SNR=−5 dB. The lower accuracies at higher SNRs were shown to relate directly to the number of reliable coefficients: recognition accuracies using the estimated mask were similar to oracle mask recognition accuracies with the same number of reliable coefficients. We showed that there is a tradeoff between the number of coefficients labeled reliable by the estimated mask and the number of false reliable coefficients. We suggest therefore that the recognition accuracy of the sparse imputation method obtained with estimated masks is predictable from a comparison with the oracle mask. Future work will focus on the application to continuous time ASR.

7. ACKNOWLEDGMENTS

The research of Jort Gemmeke was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program. The project partners are the universities of Leuven, Nijmegen and the company Nuance.

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] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, Camegie Mellon University, 2000.

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

[5] 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.

[6] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: An integrated study," in *Proceedings of Eurospeech*, 1999, pp. 2407–2410.

[7] 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.

[8] 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.

[9] J. Barker, M. Cooke, and D. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, pp. 5–25, 2005.

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

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

[12] 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.

[13] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of ISCA ASR2000 Workshop, Paris, France*, 2000, pp. 181–188.

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

[15] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[17] H. Van hamme, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proceedings of IEEE ICASSP*, 2006.

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

[19] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[20] R. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, pp. 1–15, 1997.