Noise robust digit recognition using sparse representations

J. F. Gemmeke, B. Cranen

Centre for Language and Speech Technology (CLST)
Radboud University, P.O. Box 9103,
NL-6500 HD Nijmegen, The Netherlands
{J.Gemmeke, B.Cranen}@let.ru.nl

Abstract

Despite the use of noise robustness techniques, automatic speech recognition (ASR) systems make many more recognition errors than humans, especially in very noisy circumstances. We argue that this inferior recognition performance is largely due to the fact that in ASR speech is typically processed on a frame-by-frame basis preventing the redundancy in the speech signal to be optimally exploited. We present a novel non-parametric classification method that can handle missing data while simultaneously exploiting the dependencies between the reliable features in an entire word. We compare the new method with a state-of-the-art HMM-based speech decoder in which missing data are imputed on a frame-by-frame basis. Both methods are tested on a single digit recognition task (based on AURORA-2 data) using an oracle and an estimated harmonicity mask. We show that at an SNR of -5 dB using the reliable features of an entire word allows an accuracy of 91% (using mel-log-energy features in combination with an oracle mask), while a conventional frame-based approach achieves only 61%. Results obtained with the harmonicity mask suggest that this specific mask estimation technique is simply unable to deliver sufficient reliable features for acceptable recognition rates at these low SNRs.

Index Terms: Non-parametric model, Compressed Sensing, Missing Data Techniques

1. Introduction

Humans outperform automatic speech recognition (ASR) systems, especially when the speech is severely degraded by noise. While much progress has been made in the past few decades, at low SNRs, ASR is so little noise robust that it becomes virtually useless. In this paper we argue that this inferior performance at low SNRs is largely due to the frame-based approach that is typically employed in state-of-the-art ASR systems.

Missing Data Techniques (MDT) [1, 2] have proved a powerful means to mitigate the impact of noise on recognition accuracy. 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: One may either impute the unreliable features [3, 4], or one may employ a marginalization approach in the decoder when evaluating unreliable input data [2].

The continuity constraints imposed upon the speech signal by the speech production system are considered to constitute an important source of redundancy. In most ASR approaches, however, remarkably little effort is spent on exploiting this knowledge. Speech is mostly processed on a frame by frame (i.e. strictly local) basis. The disadvantages of this approach become particularly evident at low SNRs ($\leq$ 0 dB). In those conditions, it may happen that only few, if any, elements in an acoustic vector are labeled reliable. And obviously, the fewer reliable features remain, the more serious the risk that an individual frame contains too little information for properly dealing with the unreliable coefficients. It therefore seems logical to presume that, if too many frames with only a few reliable features exist, recognition accuracy will suffer significantly and that this problem can only be relaxed by making better use of the reliable features in neighboring frames as well.

In this paper we present an innovative classification technique recently introduced in the field of face recognition [5] that allows us to study the importance of wider time context. This technique, hereafter referred to as *sparse classification*, is an application of *compressed sensing* [6, 7] and performs classification by looking how well an observed sequence of feature vectors can be explained by a linear combination of example speech sequences from the same class. This non-parametric approach requires no training and is readily extended to handle missing data since the dimensionality of the observed speech signal does not have to be fixed in advance as opposed to when using a parametric model.

In order to investigate to what extent recognition accuracy can be improved by using reliable data from a much wider time context than just the current frame, we compare recognition results obtained with the novel sparse classification technique to those of a classical ASR engine that employs frame based imputation. We study a single digit recognition task where the digits were taken from the AURORA-2 corpus. Since the performance of any MDT technique hinges on the quality of the spectrographic mask, we investigate sparse classification for two types of masks: 1) an oracle mask[1] and 2) an estimated mask in the form of a harmonicity mask [8].

2. Method

2.1. Speech data and classification task

The classification method described in the following sections works on observation vectors of fixed size. This makes it not directly applicable to running speech. In this paper, we therefore restrict ourselves to a single-digit recognition/classification task. The single-digit utterances were created by copying all individual digits from the AURORA-2 corpus. The segment

[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

boundaries were obtained via a forced alignment of the clean speech utterances with the reference transcriptions. We used only 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 classification method and SNR.

2.2. Baseline speech decoder

As our baseline system, we used a MATLAB implementation of a missing data recognition system described in [4]. Acoustic feature vectors consisted of 69 PROSPECT features [4], constructed from 23 mel frequency log power spectra as well as their first and second derivatives. Features that are labeled as unreliable (by some externally provided spectrographic mask) are replaced by estimated values using maximum likelihood per Gaussian-based imputation [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.

2.3. Fixed length vector representation of digits

To obtain a fixed length feature vector for each digit, as required by the sparse representation method in section 2.4, we converted the variable number of acoustic vectors making up a word unit (originally at a fixed frame rate of 100 Hz) to a time normalized version (with a fixed number of acoustic vectors at a variable frame rate). A spline interpolation was applied to the individual time tracks of all mel frequency log-energy coefficients after which they were re-sampled so that 35 acoustic vectors per digit resulted (i.e., the mean number of 10 ms time frames per word in the training set). Next, these time-frames were concatenated to form a single, fixed length observation vector $\boldsymbol{y}$ per digit with a dimension $K = 23 \cdot 35 = 805$.

2.4. Sparse representation

Following [5] we consider a test digit $\boldsymbol{y}$ to be a linear combination of exemplar digits $\boldsymbol{d}_{i,n}$, where the first index $(1 \leq i \leq I)$ denotes one of the $I = 11$ digit classes $\{$*one, two, . . . nine, zero, oh*$\}$ and the second index $(1 \leq n \leq N_i)$ a specific exemplar digit of class i with N_i the number of exemplar digits in each class. We write:

$$\boldsymbol{y} = \sum_{i=1}^{I} \sum_{n=1}^{N_i} \alpha_{i,n} \boldsymbol{d}_{i,n}$$

with weights $\alpha_{i,n} \in \mathbb{R}$.

Denoting the k^{th} vector element of $\boldsymbol{d}_{i,n}$ by $d^k_{i,n}$, and recalling that each digit in the example set is represented by a K-dimensional vector, we write our set of example digits as a $K \times N$ dimensional matrix A with $(N = N_1 + N_2 + \ldots + N_I)$:

$$A = \begin{pmatrix} \overbrace{\begin{matrix} d^1_{1,1} \ldots d^1_{1,N_1} \\ d^2_{1,1} \ldots d^2_{1,N_1} \\ \vdots \\ d^K_{1,1} \ldots d^K_{1,N_1} \end{matrix}}^{\substack{N_1 \text{ exemplars} \\ \text{of digit } 1}} & \begin{matrix} \ldots \\ \ldots \\ \\ \ldots \end{matrix} & \overbrace{\begin{matrix} d^1_{I,1} \ldots d^1_{I,N_I} \\ d^2_{I,1} \ldots d^2_{I,N_I} \\ \vdots \\ d^K_{I,1} \ldots d^K_{I,N_I} \end{matrix}}^{\substack{N_I \text{ exemplars} \\ \text{of digit } I}} \end{pmatrix}$$

Thus, we can express any digit $\boldsymbol{y}$ as

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

with $\boldsymbol{x}$ an N-dimensional vector that ideally will be sparsely represented in A as $\boldsymbol{x} = [0 \ldots 0 \ \alpha_{i,1}\alpha_{i,2} \ldots \alpha_{i,N_i} \ 0 \ldots 0]^T$ (i.e., most coefficients not associated with class i are zero)

The exemplar digits were taken from the clean train set of AURORA-2 which consists of $N = 27748$ digits. Since a matrix A with 27748 columns makes classification times impractical, we constructed a matrix of reduced size by randomly selecting a subset of the training set. We assume that the statistics of this subset will be approximately the same as of the original training set; no measures were taken to enforce additional balance for potentially important factors like gender, regional background, digit class, etc. A pilot study showed that any basis size larger than $N = 4000$ columns yielded only slightly better results. In this paper, we therefore use $N = 4000$.

2.5. l^1minimization

In order to represent a digit $\boldsymbol{y}$ by the sparse vector $\boldsymbol{x}$ one needs 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 $(N \gg K)$. Thus, the system of linear equations in Eq. 1 is *underdetermined* and has no unique solution.

Research in the field of *compressed sensing* [6, 7] has shown that if $\boldsymbol{x}$ is sparse, $\boldsymbol{x}$ can be recovered 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 is 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 combinatorial problem is an NP-hard [9] while l^1minimization can be done efficiently in polynomial time. Since in practice it may be impossible to express a digit exactly as a superposition of exemplar digits, 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 $||\boldsymbol{e}||_2 < \epsilon$.

2.6. Spectrographic mask

A spectrographic mask is a matrix with the same dimensions as the spectrographic representation of a digit. After the the resampling procedure described in Section 2.3 its size is $F \times J$ with $F = 23$ 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: 1) an oracle mask and 2) a harmonicity mask [8].

The oracle mask was computed on the resampled spectrographic representations as follows:

$$M(f,j) = \begin{cases} 1 \overset{def}{=} \text{reliable} & S(f,j) \geq (N(f,j) - \theta) \\ 0 \overset{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \tag{4}$$

with f $(1 \leq f \leq F)$ denoting frequency band and j $(1 \leq j \leq J)$ time frame. We used a fixed threshold $\theta = 3$ dB.

For the computation of the harmonicity mask the noisy speech signal is first decomposed into a harmonic and a random part. Next the local energy of speech and noise are estimated by thresholding the ratio between the harmonic and random part analogously to Eq. 4. In [8] it was determined that a threshold of $\theta = -9$ dB was optimal for AURORA-2.

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 [11]. For use in the sparse classification framework, we need to obtain a spectrographic mask with proper time normalization. To that end, we applied the resampling procedure described in Section 2.3 directly on the harmonicity mask and applied a threshold to convert the interpolated values into a binary mask. Next the mask M is reshaped into a $K = 805$-dimensional vector $\boldsymbol{m}$ by concatenating subsequent time frames (cf. Sec. 2.3).

2.7. Sparse classification (SC)

Given an observation vector $\boldsymbol{y}$ (representing an entire digit), we denote its reliable coefficients (marked by ones in the mask vector $\boldsymbol{m}$) by $\boldsymbol{y}_r$. Analogously, we denote the unreliable coefficients (marked by zeros in the mask vector $\boldsymbol{m}$) by $\boldsymbol{y}_u$.

Making no assumptions about the unreliable coefficients $\boldsymbol{y_u}$, we perform classification by comparing the *support* of $\boldsymbol{y_r}$ in parts of A associated with different classes i. In other words, we compare how well the various parts of $\boldsymbol{x}$ associated with different classes i can reproduce $\boldsymbol{y_r}$. The reproduction error is called the *residual*. The residual of class i is calculated by setting the coefficients of $\boldsymbol{x}$ not associated with i to zero while keeping the coefficients associated with i unchanged. Thus the residual is:

$$r_i(\boldsymbol{y_r}) = ||\boldsymbol{y_r} - A\delta_i(\boldsymbol{x})||_2 \qquad (5)$$

with $\delta_i(\boldsymbol{x})$, the vector selecting only the columns of A that correspond to class i.

The class c that is assigned to a observed digit $\boldsymbol{y}$ is the one that gives rise to the smallest residual:

$$c = \underset{i}{\text{argmin}}\, r_i(\boldsymbol{y_r}). \qquad (6)$$

Obviously, no classification is possible if the number of reliable coefficients in $\boldsymbol{y}$ denoted by $K_r = \dim(\boldsymbol{y}_r)$ equals zero. In practice, finding a sparse representation $\boldsymbol{x}$ will be unlikely below some threshold $K_r < \delta$. However, while for some problems the value of δ can be theoretically derived [6, 7, 12], it is not trivial to estimate bounds on the value of δ, if only because we cannot predict the sparsity of $\boldsymbol{x}$ obtained in Eq. 3. Hence, we decided to always perform classification using $\boldsymbol{y_r}$ except if $K_r = 0$ in which case we use $\boldsymbol{y}$.

The method was implemented in MATLAB. The l^1-minimization was carried out using the `l1_ls` package described in [13]. We used $\epsilon = 0.01$ in our experiments.

2.8. Dimensionality Reduction

Because solving Eq. 3 is computationally demanding, we reduce the dimensionality of the vector y_r using Random Projections as in [5] if the dimensionality of K_r is above a threshold dimension D. Using $D = 100$, we use a transformation matrix R with dimensions $K_r \times D$ to reduce the dimensionality from K_r to D if $K_r > D$. Eq. 3 then becomes:

$$min||\boldsymbol{x}||_1 \text{ subject to } ||R\boldsymbol{y_r} - RA_r\boldsymbol{x}||_2 \leq \epsilon$$

with matrix R populated with values randomly drawn from a Gaussian distribution with zero mean and unit variance. Since K_r varies per digit, a new R was constructed using a pseudo-random number generator for every digit with $K_r > D$.

3. Results

Figure 1 shows the recognition accuracies (averages over the four noise types) for the single-digits from test set A. Normal classification (NC) results for the baseline missing data recognizer are depicted with dashed lines; those for the sparse classification (SC) method with solid lines. Using the oracle mask, the sparse classification technique substantially outperforms the baseline recognizer for SNR $<$ 5 dB: At SNR=-5 dB a recognition accuracy 91% is obtained as opposed to 61%. At higher SNRs the baseline recognizer achieves consistently 1% higher recognition accuracies. At the same time, accuracies using an estimated (harmonicity) mask with sparse classification are lower than the baseline recognizer (up to 10% at SNR 5 dB).

Figure 2 shows the percentage of reliable time-frequency cells in a spectrographic mask. The oracle mask classifies a much larger proportion ($20 - 30\%$) of time-frequency cells as reliable when compared to the harmonicity mask. Also note that the bars that depicting the proportion of time-frequency cells that are labeled reliable by the harmonicity mask comprise a portion that, using the oracle mask as golden standard, are incorrectly labeled as reliable (dubbed false reliables).

4. Discussion

The recognition accuracy of 91% at SNR = -5 dB obtained with the SC-method using an oracle mask shows that at very low SNRs enough information about the speech signal is preserved to successfully perform classification solely on the basis of reliable time-frequency cells, even when the acoustic representation consists of classical mel frequency filterbands. Comparing this to the 61% recognition accuracy of the baseline decoder, it is clear that the information contained in reliable features is only partially utilized when doing missing data recognition on a frame-by-frame basis.

The drop in accuracy at the lowest SNR's from 99% to 91% is mainly due to digits which have very few or perhaps no reliable features. Apparently, it occasionally happens that simply insufficient reliable data is left to base recognition on. This effect is not unexpected and a similar drop in recognition accuracy can be observed for human subjects at negative SNRs.

Using an oracle mask, SC performs slightly worse than the NC of the baseline recognizer for SNR $\geq$ 5 dB. This indicates that the SC method does not generalize to observed digits as well as the HMM-based approach. There may be several reasons for this. Possibly, the basis size is too small and the accuracy gap could be closed by using a larger basis. Also, the dimensionality reduction presented in Sec. 2.8, while greatly reducing the computational cost, might have a slight, adverse effect on recognition accuracy. Finally, the baseline recognizer uses MFCC-like (i.e. PROSPECT) features while our classification method works directly on mel log-energy coefficients. Future research is needed to what extent these factors play a significant role.

Using the SC method with an estimated spectrographic mask, i.e. the harmonicity mask, we obtain recognition accuracies that are substantially lower than those with an oracle mask. The fact, that the SC method performs even worse than the NC of the baseline recognizer can only mean that the amount of reliable features identified as such by the harmonicity mask is too small to warrant a proper recognition.

Comparing the accuracies obtained with SC in combination with the oracle mask on the one hand and those obtained with NC in combination with the harmonicity mask on the other (cf. Fig. 1), one may observe that the values of the first curve attain approximately the same values as the latter, but at SNRs which are 10 dB higher Fig. 2 reveals that a similar relation holds for the percentage of reliable cells (the underdeterminedness): The percentage of reliable cells found with the harmonicity mask is roughly the same as with an oracle mask at noise levels which are 10 dB apart. Assuming that the fraction of reliable features per word is determining the maximum achievable recognition accuracy, this suggests that the baseline recognizer already has reached a ceiling and that the low accuracies at lower SNRs must be attributed to the fact that there are simply not enough reliable coefficients left.

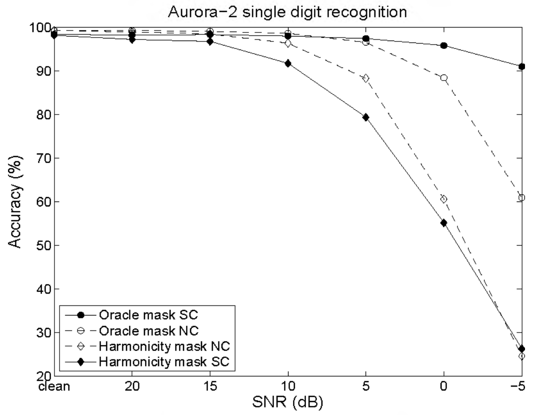

Figure 1: AURORA-2 single digit recognition accuracy. The figure shows results for both normal classification (NC) and sparse classification (SC) for the oracle mask and the harmonicity mask

A possible explanation for the fact that the SC method performs worse than NC is that the baseline decoder uses information not available to the SC method. In contrast to the SC method which makes no assumptions about the unreliable features, the baseline recognizer assumes that the energy of the speech signal in these time-frequency cells cannot exceed the observed energy. Without these bounding constraints, the SC method might also be more sensitive to false reliables which, as shown in Fig. 2, also become more numerous as noise levels increase. While not applied in the current study, the sparse classification method could easily be extended with a similar constraint.

The current implementation of the SC technique only works with fixed length feature representations. While already interesting in its own right for command&control applications, in order to be used in a general ASR system, 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 commonly used in frame-based Support Vector Machine and Neural Net classification tasks. Another approach would be to define a larger basis using time-shifted copies of the original basis [14]. The practical applicability and computational feasibility of either method is left as future work.

5. Conclusions

We introduced a non-parametric missing data classification 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 method exploits the redundancy of the speech signal in the time-frequency domain by expressing entire words as a linear combination of exemplar speech signals. We showed the potential of the method by achieving recognition accuracies on AURORA-2 digits of 91% at SNR -5 dB using an oracle mask, an increase of 30% percent absolute over a state-of-the art missing data speech recognizer. These findings show that much progress can still be made using conventional features by adapting the decoding algorithms so that the redundancy in the time domain is properly exploited. However, the recognition accuracies obtained with the harmonicity mask also indicate that this mask estimation technique might simply deliver too few reliable time-frequency cells to enable truly noise robust recognition. In order to be able to really profit from these insights future work will need to focus on techniques that exploit the redundancy properties of speech already during the mask estimation procedure.

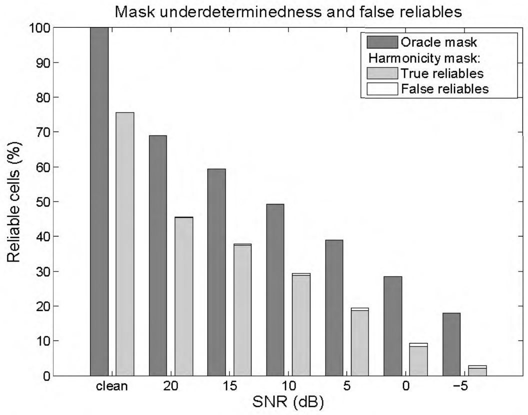

Figure 2: Percentage of reliable time-frequency cells for the oracle and harmonicity mask. The white areas indicate the portions that, according to the oracle mask, are falsely labeled reliable.

6. Acknowledgements

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.

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] 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] 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.

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

[7] E. J. Candes, “Compressive sampling,” in *Proceedings of the International Congress of Mathematicians*, 2006.

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

[12] Y. Zhang, “When is missing data recoverable?” *Technical Report*, 2006.

[13] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, “A method for large-scale l1-regularized least squares,” *IEEE Journal on Selected Topics in Signal Processing*, vol. 4, pp. 606–617, dec 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.