Chapter 16
Missing Data Solutions for Robust Speech Recognition

Yujun Wang, Jort F. Gemmeke, Kris Demuynck, and Hugo Van hamme

16.1 Introduction

One of the major concerns when deploying speech recognition applications is the lack of robustness of the technology. Humans are robust to noise, different acoustic environments, pronunciation variation, ungrammatical sentences, incomplete utterances, filled pauses, stutters, etc. and this engenders the same expectation for automatic systems. In this contribution we discuss an approach called missing data techniques (MDT) [3, 27] to deal with one of these problems: noise robustness. Unlike many previously proposed solutions, MDT can deal with noise exhibiting rapidly changing characteristics, which is often the case in practical deployments. For example, a mobile device used in a city will pick up the noise of cars passing by, of construction sites, from car horns, of people talking or shouting, etc.

In a nutshell, MDT is based on the idea that even in noisy speech, some of the features describing the speech signal remain uncorrupted. The goal is to identify the corrupted (missing) features and to then replace them (impute) with clean speech estimates. In this contribution we describe the research carried out in the MIDAS project, which focussed on two aspects of MDT. First, we discuss an novel imputation method to derive clean speech estimates of the corrupted noise speech features, a method dubbed Sparse Imputation. This method models speech as a linear combination of *exemplars*, segments of speech, rather than modelling speech using a statistical model. Second, we describe how a state-of-the-art large vocabulary automatic speech recognition (ASR) system based on the prevailing hidden Markov model (HMM) can be made noise robust using conventional MDT. Unlike many publications on noise robust ASR, which only report results on

Y. Wang · J.F. Gemmeke · K. Demuynck · H. Van hamme (✉)
ESAT Department, Katholieke Universiteit, Leuven, Belgium
e-mail: yujun.wang@esat.kuleuven.be; jort.gemmeke@esat.kuleuven.be; kris.demuynck@esat.kuleuven.be; hugo.vanhamme@esat.kuleuven.be

P. Spyns and J. Odijk (eds.), *Essential Speech and Language Technology for Dutch*,
Theory and Applications of Natural Language Processing,
DOI 10.1007/978-3-642-30910-6_16, © The Author(s) 2013

artificially corrupted noisy speech, this chapter also describes results on noisy speech recorded in realistic environments.

The rest of the chapter is organised as follows. In Sect. 16.2 we briefly introduce MDT. In Sect. 16.3 we describe the sparse imputation method and the AURORA-2 and Finnish SPEECON [21] databases used for evaluations, and in Sect. 16.4 we describe and discuss the recognition accuracies that were obtained. In Sect. 16.5 we describe the large-vocabulary ASR system, the MDT method employed, and the material from the Flemish SPEECON [21] and SpeechDat-Car [30] databases that were used. In Sect. 16.6 we investigate the performance of the resulting system, both in terms of speech recognition accuracy as well as in terms of speed of program execution. We conclude with a discussion and present our plans for future work in Sect. 16.7.

16.2 Missing Data Techniques

In ASR, the basic representation of speech is a spectro-temporal distribution of acoustic power, a *spectrogram*. The spectrogram typically consist of 20–25 band-pass filters equally spaced on a Mel-frequency scale, and is typically sampled at 8 or 10 ms intervals (a frame). In noise-free conditions, the value of each time-frequency cell in this two-dimensional matrix is determined only by the speech signal. In noisy conditions, the value in each cell represents a combination of speech and background noise power. To mimic human hearing, a logarithmic compression of the power scale is employed.

In the spectrogram of noisy speech, MDT distinguishes time-frequency cells that predominantly contain speech or noise energy by introducing a missing data mask. The elements of that mask are either 1, meaning that the corresponding element of the noisy speech spectrogram is dominated by speech (‘reliable’) or 0, meaning that it is dominated by noise (‘unreliable’ c.q. ‘missing’). Assuming that only additive noise corrupted the clean speech, the power spectrogram of noisy speech can be approximately described as the sum of the individual power spectrograms of clean speech and noise. As a consequence, in the logarithmic domain, the reliable noisy speech features remain approximately uncorrupted [27] and can be used directly as estimates of the clean speech features. It is the goal of the imputation method to replace (‘impute’) the unreliable features by clean speech estimates.

After imputation in the Mel-spectral domain, the imputed spectra can be converted to features such as Mel-frequency cepstral coefficients (MFCC). Then, delta and delta-delta derivative features (used in all experiments described in this chapter) can be derived from these. If the clean speech and noise signals or their spectral representations are available so that we know the speech and noise power in each time-frequency cell, a so-called *oracle mask* may be constructed. In realistic situations, however, the location of reliable and unreliable components needs to be estimated. This results in an *estimated mask*. For an overview of mask estimation methods we refer the reader to [2].

16.3 Material and Methods: Sparse Imputation

16.3.1 Sparse Imputation

In this section we give a brief and informal account of the sparse imputation method. For a more formal and in-depth explanation we refer to [12, 15].

In the sparse imputation approach, speech signals are represented as a linear combination of example (clean) speech signals. This linear combination of *exemplars* is sparse, meaning that only a few exemplars should suffice to model the speech signal with sufficient accuracy. The collection of clean speech exemplars used to represent speech signals is called the exemplar dictionary, and is randomly extracted from a training database.

The observed noisy speech signals are processed using overlapping windows, each consisting of a spectrogram spanning 5–30 frames. For this research neighbouring windows were shifted by a single frame. Sparse imputation works in two steps. First, for each observed noisy speech window, a maximally sparse linear combination of exemplars from the dictionary is sought using only the reliable features of the noisy speech and the corresponding features of the exemplars. Then, given this sparse representation of the speech, a clean speech estimate of the unreliable features is made by reconstruction using only those features in the clean speech dictionary that correspond to the locations of the unreliable features of the noisy speech. Applying this procedure for every window position, the clean speech estimates for overlapping windows are combined through averaging.

16.3.2 Databases

The main experiments with sparse imputation have been carried out using the AURORA-2 connected digit database [18] and the Finnish SPEECON large vocabulary database [21]. The use of these databases in [15] and [12] are briefly introduced below. For evaluations of sparse imputation on other tasks and databases, we refer the reader to [9, 10, 23].

In [15], a digit classification task was evaluated using material from the AURORA-2 corpus. The AURORA-2 corpus contains utterances with the digits ‘zero’ through ‘nine’ and ‘oh’, and one to seven digits per utterance. The isolated-digit speech data was created by extracting individual digits using segmentations obtained by a forced alignment of the clean speech utterances with the reference transcription. The clean speech training set of AURORA-2 consists of $27,748$ digits. The test digits were extracted from test set A, which comprises 4 clean and 24 noisy subsets. The noisy subsets are composed of four noise types (subway, car, babble, exhibition hall) artificially mixed at six SNR values, SNR$= 20, 15, 10, 5, 0, -5\,$dB. Every SNR subset consisted of 3,257, 3,308, 3,353 and 3,241 digits per noise type, respectively.

In [12], material from the Finnish SPEECON large vocabulary database was used. The artificially corrupted read speech was constructed by mixing headset-recorded clean speech utterances with a randomly selected sample of the babble noise from the NOISEX-92 database [37] at four SNR values, SNR$= 15, 10, 5, 0$ dB. The training data consists of 30 h of clean speech recorded with a headset in quiet conditions, spoken by 293 speakers. The test set contains 115 min of speech in 1,093 utterances, spoken by 40 speakers.

16.4 Experiments: Sparse Imputation

In this section we give an overview of the most important results obtained with sparse imputation as reported in [12, 15]. In Sect. 16.4.1 we give a summary of the experimental setup and in Sect. 16.4.2 we discuss the obtained results.

16.4.1 Experimental Setup

16.4.1.1 Digit Classification

For the digit classification task described above, only a single sparse representation was used to represent the entire digit. In other words, only a single window was used, and each digit was time-normalised using linear interpolation to have a fixed length of 35 (8 ms) frames. With each spectrogram consisting of 23 Mel-frequency bands, each digit was thus described by $23 \cdot 35 = 805$ features. The exemplar dictionary consisted of $4,000$ exemplars randomly extracted from the digits in the training database.

Recognition was done using a MATLAB-based ASR engine that can optionally perform missing data imputation using Gaussian-dependent imputation (cf. Sect. 16.5.1) [32]. After applying sparse imputation in the mel-spectral domain, recognition was carried out using PROSPECT features [32]. This technique is described in more detail in Sect. 16.5.1. For further comparison, the cluster-based imputation technique proposed in [26] was used. Two missing data mask methods were used, the oracle mask described in Sect. 16.2 and an estimated mask. In brief, the estimated mask combines a harmonic decomposition and an SNR estimate to label features mostly dominated by harmonic energy and/or with a high SNR as reliable [33].

16.4.1.2 Large Vocabulary Task

For the large vocabulary recognition task using SPEECON, we focus on the results obtained using sparse imputation with spectrograms spanning 20 (8 ms) frames.

With each spectrogram consisting of 21 Mel-frequency bands, each window was thus described by $21 \cdot 20 = 420$ features. The exemplar dictionary consisted of $8,000$ spectrograms randomly extracted from the clean speech in the training database and thus contains anything from whole words, parts of words, word-word transitions to silence and silence-word transitions.

Recognition was done using the large vocabulary continuous speech recognition system developed at the Aalto University School of Science [19]. After imputation in the mel-spectral domain, recognition was carried out using MFCC features. For comparison, the cluster-based imputation technique proposed in [26] was used. The Gaussian-dependent imputation technique used in the other experiments in this chapter was not used, since that method requires recogniser modifications that have not been applied to the Finnish ASR engine used in this experiment. Two missing data mask methods were used, the oracle mask described above and an estimated mask. Unlike the mask estimation method described above, the estimated mask does not employ harmonicty and is constructed using only local SNR estimates obtained from comparing the noisy speech to a static noise estimate calculated during speech pauses [28]. All parameters were optimised using the development data. The speech recognition performance is measured in letter error rates (LER) because the words in Finnish are often very long and consist of several morphemes.

16.4.2 Results

In the experiments described here, the aim is to evaluate the effectiveness of sparse imputation compared to other imputation methods. To that end we compare classification accuracy or recognition accuracy as a function of SNR as obtained with various methods.

In Fig. 16.1 we compare the performance obtained with sparse imputation, cluster-based and Gaussian-dependent imputation on the digit classification task. For estimated masks, we can observe that sparse imputation performed comparably or somewhat worse than Gaussian-dependent imputation but better than cluster-based imputation. For oracle masks, sparse imputation outperforms both Gaussian-dependent and cluster-based imputation by a large margin at SNRs <15 dB.

In Fig. 16.2 we compare the performance obtained with sparse imputation, cluster-based and the baseline recogniser on the SPEECON large vocabulary recognition task. We can observe that sparse imputation performs much better than cluster-based imputation if an oracle mask was used. When using an estimated mask, sparse imputation performed better than cluster-based imputation at lower SNRs, and comparably at higher SNRs. These findings were confirmed in experiments on noisy speech recorded in real-world car and public environments [12].

From these results it is already apparent that advances in mask estimation quality are necessary for further advances in noise robustness, especially for sparse imputation. We will revisit this issue in Sect. 16.7.

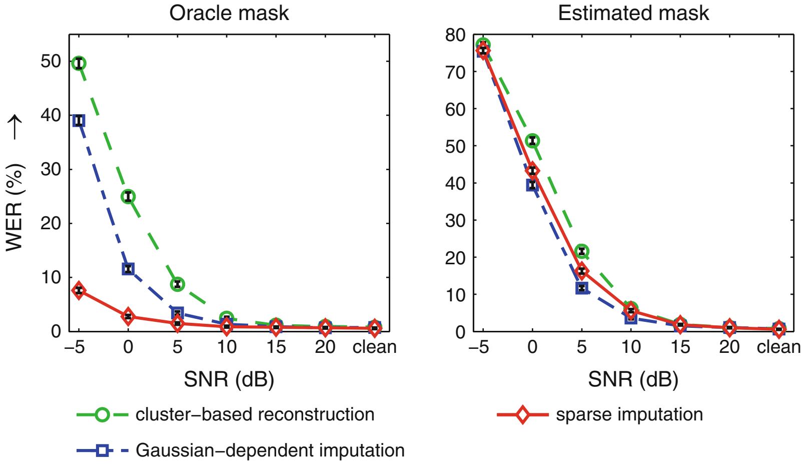

Fig. 16.1 Word error rates (WER) obtained on AURORA-2 isolated digits database with cluster-based, Gaussian-dependent, and sparse imputation. The *left panel* shows the results obtained using oracle masks and the *right panel* shows the results obtained using estimated masks. The horizontal axis describes the SNR at which the clean speech is mixed with the background noise and the vertical axis describes the WER averaged over the four noise types: subway, car, babble, and exhibition hall noise. The vertical bars around data points indicate the 95 % confidence intervals, assuming a binomial distribution

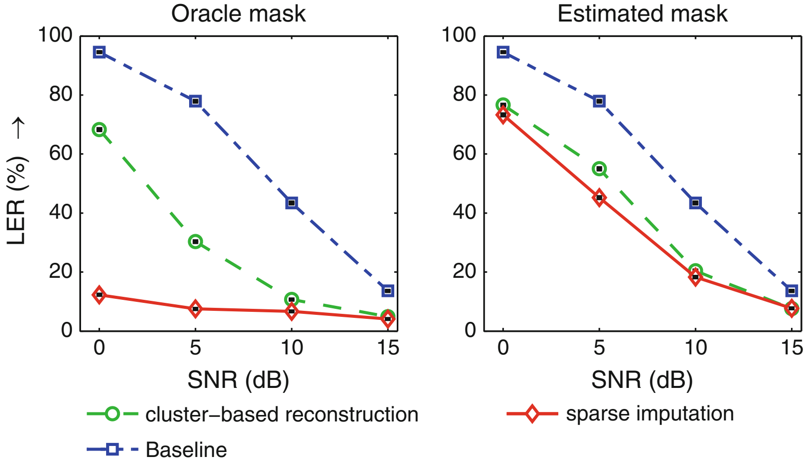

Fig. 16.2 Letter error rates (LER) obtained on the Finnish SPEECON database with cluster-based imputation and sparse imputation. The *left panel* shows the results obtained using oracle masks and the *right panel* shows the results obtained using estimated masks. The horizontal axis describes the SNR at which the clean speech is artificially mixed with babble noise. The vertical bars around data points indicate the 95 % confidence intervals, assuming a binomial distribution

16.5 Material and Methods: Gaussian-Dependent Imputation

16.5.1 Gaussian-Dependent Imputation

Originally, MDT was formulated in the log spectral domain [3]. Here, speech is represented by the log-energy outputs of a filter bank and modelled by a Gaussian Mixture Model (GMM) with diagonal covariance. In the *imputation* approach to MDT, the GMM is then used to reconstruct clean speech estimates for the unreliable features. When doing *bounded imputation*, the unreliable features are not discarded but used as an upper bound on the log-power of the clean speech estimate [4].

Later, it was found the method could be improved by using state-dependent [22] or even Gaussian-dependent [31] clean speech estimates. In these approaches, the unreliable features are imputed during decoding and effectively depend on the hypothesised state identity. However, filter bank outputs are highly correlated and poorly modelled with a GMM with a diagonal covariance. This is the reason why conventional (non-MDT) speech recognisers employ cepstral features, obtained by applying a de-correlating Discrete Cosine Transformation (DCT) on the spectral features.

In [31] it was proposed to do cepstral-domain Gaussian-dependent (bounded) imputation by solving a non-negative least squares (NNLSQ) problem. The method proposed in [32] refines that approach by replacing the DCT used in the generation of cepstra by another data-independent linear transformation that results in computational gains while solving the NNLSQ problem. The resulting PROSPECT features are, just like cepstral coefficients, largely uncorrelated and therefore allow to retain the high accuracy at high SNRs as well as the good performance at lower SNRs obtained with Gaussian-dependent imputation.

16.5.1.1 Multi-candidate MDT

In this chapter we use a faster approach that does not solve the imputation problem for every backend Gaussian (BG), the Gaussians of the HMM acoustic model, but only for a small set of Gaussians using a technique called Multi-Candidate(MC) MDT [38]. In MC MDT, a reduced set of Cluster Gaussians (CG) are established on top of the BGs, with the number of CGs one to two orders of magnitude smaller than the number of BGs. Instead of solving the imputation problem for each BG, candidate solutions are selected from the CGs through MDT imputation. The candidate that maximises the likelihood of the BG is retained as the BG-dependent prediction of the clean speech. In other words, the MDT imputation problem is solved approximately for the BG by constraining possible solutions to a set proposed by the CGs. Computational gains in CG imputation are again obtained by a PROSPECT formulation. The imputed clean filter-bank energies are then transformed to the preferred feature representation of the BGs. This means that the backend acoustic model of a non-MDT system can be used, which constitutes a great advantage when building MDT systems.

However, since there may be hundreds of CGs, it is not feasible to evaluate each BG on each candidate solutions. Therefore, for every BG, we construct a short-list of CGs that were most successful in producing a winning candidate on a forced alignment of clean training data. The length of this short-list controls the trade-off between computational effort and obtained robustness. Experiments have shown that retaining only a handful of CGs does not lead to loss of accuracy.

During recognition, Gaussian selection is combined with MC MDT. Gaussian selection is motivated by the observation that only a small (frame dependent) portion of Gaussians dominate the likelihoods of the HMM states, and are therefore worth evaluating. The likelihood of a CG evaluated at its imputed value is used to select only the CGs that describe the frame of data sufficiently well. The unlikely CGs are not allowed to propose candidates for evaluation by the BG, which leads to the desired result that unlikely BGs are not evaluated. The proposed Gaussian selection method differs from traditional Gaussian selection methods [1, 6] in that it uses MDT to select relevant clusters. This is advantageous since data that is not close to the Gaussian means because it is severely corrupted by noise can still activate the appropriate Gaussian that models the underlying clean speech.

16.5.1.2 Mask Estimation

To estimate the missing data mask, we use a Vector Quantisation (VQ) strategy that is closely related the method employed in [36]. The key idea is to estimate masks by making only weak assumptions about the noise, while relying on a strong model for the speech, captured in a codebook. The harmonicity found in voiced speech is exploited through a harmonic decomposition as proposed in [33], which decomposes the signal in two parts: the periodic signal part consists of the harmonics at pitch multiples and the remaining spectral energy is considered the aperiodic part.

We first construct a codebook of clean speech by clustering stacked features containing framed spectral representations of the periodic and aperiodic decomposition of the clean speech. Since the codebook only represents a model for the human voice, decoding of non-speech (or noise) frames will lead to incorrect codebook matching and misclassification of mask elements. Therefore, a second, much smaller codebook, is used for non-speech frames. A *Voice Activity Detector* (VAD), segments speech from non-speech frames in order to select the appropriate VQ codebook. To compensate for linear channel distortions, the VQ-system self-adjusts the codebook to the channel during recognition.

The input for the VQ-system consists of three components. The first two components, the spectral representation of the periodic and aperiodic decomposition of the noisy speech, match the content of the codebook entries. The third input component is an estimate of the mean of the noise spectrum. The VQ-system compares the first two components with the content of each codebook entry, given that noise must be added to the noise-free codebook entries. The instantaneous noise is assumed to be drawn from a distribution with the given noise mean (the third input) and is assumed to be smooth, meaning that it has no periodic structure so that the instantaneous noise periodic and aperiodic parts are close to identical. The smoothness assumption is reasonable for many noise types including speech babble, but it may be violated for a single interfering speaker or for some types of music.

The two noise assumptions allow a closed form distance metric for comparing the noise free VQ-entries with the noisy input and, as a side effect, also returns the estimated instantaneous noise [36]. A speech estimate is obtained by summing the periodic and aperiodic both parts of the codebook entry. Once the best matching codebook entry is found, the spectrographic VQ-based mask is estimated by thresholding the ratio of speech and noise estimates.

The noise tracker (third input to the VQ-system) combines two techniques. First, a short-term spectral estimate of the aperiodic noise is obtained from minimum statistics [24] on the aperiodic component of the noisy signal over a sub-second window. This system is well suited for rapid changing noise types with no periodic structure. A disadvantage of this approach is that the tracker also triggers on long fricatives and fails on periodic noise types. Whereas the experiments in previous publications used only this method, in this work we added a noise tracker developed for another noise robust technique present in SPRAAK called *noise normalisation* [7]. This second noise tracker looks over a longer 1.5 s window and uses ordered statistics instead of minimum statistics to obtain more robust and accurate noise estimates. By combining the two noise trackers, good behaviour on both stationary, non-stationary and periodic noise types is obtained.

16.5.2 Real-World Data: The SPEECON and SpeechDat-Car Databases

In the research reported in this part of the chapter we use material from the Flemish SPEECON [21] and the SpeechDat-Car [30] databases. These databases contain speech recorded in realistic environments with multiple microphones. In total, there are four recording environments: office, public hall, entertainment room and car. All speech material was simultaneously recorded with four different microphones (channels) at increasing distances, resulting in utterances corrupted by varying levels of noise. We used the method described in [16] to obtain SNR estimates of all utterances.

The multi-condition training data consists of 231, 849 utterances spoken in 205 h of speech. The speech is taken from the office, public hall and car noise environments, with most of the data (168 h) coming from the office environment. It contains a clean data portion of 61, 940 utterances from channel #1 (closetalk microphone) data with an estimated SNR range of 15–50 dB. Additionally, the multi-condition set contains all utterances from channels #2, #3 and #4 which have an estimated SNR of 10 dB and higher, containing 54,381, 53,248 and 31,975 utterances, respectively.

For the test set, we use material pertaining to a connected digit recognition task. The utterances contain the ten digits ‘zero’ through ‘nine’, with between one and ten digits per utterance. The 6,218 utterances (containing 25,737 digits) of the test set are divided in 6 SNR subsets in the 0–30 dB range with a 5 dB bin width. The SNR bins do not contain equal numbers of utterances from the four channels: Generally speaking, the highest SNR bins mostly contain utterances from channel #1, while the lowest SNR bins mostly contains channel #4 speech.

16.6 Experiments: Gaussian-Dependent Imputation

16.6.1 Experimental setup

16.6.1.1 Speech Recogniser and Acoustic Models

The implementation of MDT does not require a complete overhaul of the software architecture of a speech recogniser. We extended the code of the SPRAAK-recogniser (described in Chap. 6, p. 95) to include a missing data mask estimator (cf. Sect. 16.5.1.2) and to evaluate the acoustic model according to the principles described in Sect. 16.5.1. Below, we will successively describe the configuration in which SPRAAK was used, how its acoustic models were created, how the baseline so-called PROSPECT models were created to benchmark the proposed speed-ups and how the data structures for the multi-candidate MDT were obtained.

The acoustic feature vectors consisted of MEL-frequency log power spectra: 22 frequency bands with centre frequencies starting at 200 Hz. The spectra were created by framing the 16 kHz signal with a Hamming window with a window size 25 ms and a frame shift of 10 ms. The decoder also uses the first and second time derivative of these features, resulting in a 66-dimensional feature vector. This vector is transformed linearly by using Mutual Information Discriminant Analysis (MIDA) linear transformation [5]. During training, mean normalisation is applied to the features. During decoding, the features are normalised by a more sophisticated technique which is compatible with MDT and which works by updating an initial channel estimate through maximisation of the log-likelihood of the best-scoring state sequence of a recognised utterance [35].

The training of the multi-condition context-dependent acoustic models on a set of 46 phones plus four filler models and a silence model follows the standard training scripts of SPRAAK and leads to 4,476 states tied through a phonetic decision tree and uses a pool of 32,747 Gaussians.

16.6.1.2 Imputation

A set of 700 cluster Gaussians required for the MC-MDT acoustic model was obtained by pruning back the phonetic tree to 700 leaves, each modelled with a single PROSPECT Gaussian trained on the respective training sets. The cluster Gaussians and backend Gaussians are associated in a table which retains only the most frequent co-occurrence of the most likely cluster Gaussian and the most likely backend Gaussian in Viterbi alignment of the training data. The SPRAAK toolkit was extended with adequate tools to perform these operations. The association table was then pruned to allow maximally five cluster Gaussians per back-end Gaussian. The average number of cluster Gaussians per back-end Gaussian is 3.6.

The VQ-codebook used in mask estimation was trained on features extracted from the close-talk channel SPEECON training database. The number of codebook entries was 500 for speech and 20 for silence. Recognition tests on the complete test set using a large interval of threshold values revealed that the threshold setting was not very sensitive. The (optimal) results presented in this work were obtained with 8 dB. Missing data masks for the derivative features were created by taking the first and second derivative of the missing data mask [34].

16.6.1.3 VOCON

The VOCON 3200 ASR engine is a small-footprint engine, using MFCC based features and HMM models. It contains techniques to cope with stationary or slowly varying background noise. Its training data includes in-car recorded samples, i.e., it uses the multi-condition training approach in tandem with noise reduction techniques. The VOCON recogniser uses whole-word models to model digits, whereas the MDT system uses triphones.

16.6.2 Results

In Fig. 16.3 we compare the performance obtained with the SPRAAK baseline system (the SPRAAK system described in Sect. 16.6.1.1, without employing MDT), the SPRAAK MDT system and the VOCON recogniser. We can observe that the use of MDT in the SPRAAK recogniser reduces the WER substantially in all noise environments and at all SNRs. The only exception is the 0–5 dB SNR bin in the office noise environment, but here the difference with the SPRAAK baseline is not significant. When comparing the SPRAAK recognisers with the VOCON recogniser, we observe that the VOCON recogniser typically performs better at the lowest SNRs, but at the cost of a higher WER at higher SNRs.

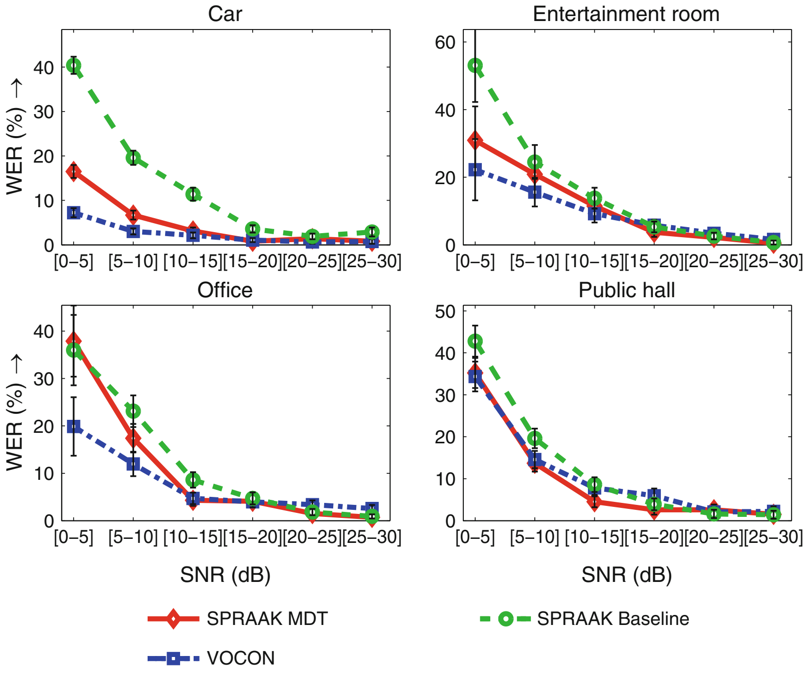

Fig. 16.3 Word error rates (WER) obtained on the Flemish SPEECON database on a connected digit recognition task, comparing the SPRAAK recogniser using imputation (MDT), the SPRAAK baseline and the VOCON recogniser. Four noise environments are shown, viz. car, entertainment room, office, and public hall. The horizontal axis describes the estimated SNR of the noisy speech. The vertical bars around data points indicate the 95 % confidence intervals, assuming a binomial distribution

Table 16.1 Timing experiments using speech from the 25–30 dB SNR bin

	CPU time (ms/frame)						
	Mask	BG	CG	Beam search	Total	BGs calculated (%)	WER (%)
SPRAAK baseline	0.0	0.9	0.2	1.8	2.9	5	1.47
SPRAAK MDT	1.1	4.1	4.2	1.8	11.2	14	0.88

In Tables 16.1 and 16.2 we show the results of a timing experiment on ‘clean’ speech (25–30 dB SNR) and noisy speech (10–15 dB SNR), respectively. The timings are obtained by recognition of 10 randomly selected sentences per noise environment (40 in total), which together contain 22,761 frames for the clean speech and 16,147 frames for the noisy speech. We can observe that the use of MDT in SPRAAK is approximately four times slower than the SPRAAK baseline in clean

Table 16.2 Timing experiments using speech from the 10–15 dB SNR bin

	CPU time (ms/frame)						
	Mask	BG	CG	Beam search	Total	BGs calculated (%)	WER (%)
SPRAAK baseline	0.0	2.7	0.2	1.9	4.8	13	10.0
SPRAAK MDT	1.1	7.2	4.7	1.8	21.1	27	5.85

conditions, but only two times slower in noisy conditions. As can be seen from the average WERs in the two tables, the use of MDT approximately halves the WER, even in the cleaner conditions.

16.7 Discussion and Conclusions

From the results obtained with sparse imputation, one can draw two conclusions. On the one hand, the sparse imputation method achieved impressive reductions in WER and LER when used in combination with an oracle mask. On the other hand, although sparse imputation performs better than cluster-based imputation when using estimated masks, it does not perform better than Gaussian-dependent imputation. This means that for sparse imputation to reach its full potential, advances in mask estimation techniques are necessary. Unfortunately, despite a decade of research on missing data techniques the gap between estimated masks and oracle masks remains [8].

From the results obtained with the SPRAAK recogniser employing MDT, we observed a substantial improvement in noise robustness. Although at the cost of two to four times lower execution speed, the WER halved even in the cleaner conditions. Moreover, it was reported in [38] that the proposed MDT technique is not significantly slower than the baseline SPRAAK recogniser when applied on a large vocabulary task. In comparison to the VOCON recogniser, the SPRAAK recogniser typically performs better at moderate-to-high SNRs. The noise robustness of the VOCON recogniser at low SNRs can probably be attributed to its use of whole-word models.

With respect to sparse imputation, various improvements have been proposed recently, such as the use of *probabilistic masks* [11], the use of observation uncertainties to make the recogniser aware of errors in estimating clean speech features [14], and the use of additional constraints when finding a sparse representation [29]. Finally, in the course of the MIDAS project a novel speech recognition method was proposed which explicitly models noisy speech as a sparse linear combination of speech and noise exemplars, thus bypassing the need for a missing data mask. Although only evaluated on small vocabulary tasks, the results are promising [13, 17, 20], e.g., achieving a 37.6 % WER at SNR $= -5$ dB on AURORA-2.

Future work concerning the noise robust SPRAAK recogniser will focus on improving mask estimation quality in two ways. First, while it has been shown MDT can be used to combat reverberation [16, 25], to date no method has been presented that enables the estimation of reverberation-dominated features in noisy environments. Second, future work will address the poor performance of current mask estimation methods on speech corrupted by background music, a prevailing problem in searching audion archives.

References

1. Bocchieri, E.: Vector quantization for efficient computation of continuous density likelihoods. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, vol. 2, Minneapolis, Minnesota, USA, pp. 692–695 (1993)
2. Cerisara, C., Demange, S., Haton, J.P.: On noise masking for automatic missing data speech recognition: A survey and discussion. Comput. Speech Lang. **21**(3), 443–457 (2007)
3. Cooke, M., Green, P., Crawford, M.: Handling missing data in speech recognition. In: Proceedings of the International Conference on Spoken Language Processing, Yokohama, Japan, pp. 1555–1558 (1994)
4. Cooke, M., Green, P., Josifovski, L., Vizinho, A.: Robust automatic speech recognition with missing and unreliable acoustic data. Speech Commun. **34**(3), 267–285 (2001)
5. Demuynck, K., Duchateau, J., Compernolle, D.V.: Optimal feature sub-space selection based on discriminant analysis. In: Proceedings of the European Conference on Speech Communication and Technology, vol. 3, Budapest, Hungary, pp. 1311–1314 (1999)
6. Demuynck, K., Duchateau, J., Van Compernolle, D.: Reduced semi-continuous models for large vocabulary continuous speech recognition in Dutch. In: Proc. the International Conference on Spoken Language Processing, vol. IV, Philadelphia, USA, pp. 2289–2292 (1996)
7. Demuynck, K., Zhang, X., Van Compernolle, D., Van hamme, H.: Feature versus model based noise robustness. In: Proc. INTERSPEECH, Makuhari, Japan, pp. 721–724 (2010)
8. Gemmeke, J.F.: Noise robust ASR: missing data techniques and beyond. Ph.D. Thesis, Radboud Universiteit Nijmegen, The Netherlands (2011)
9. Gemmeke, J.F., Cranen, B.: Noise reduction through compressed sensing. In: Proceedings of the INTERSPEECH, Brisbane, Australia, pp. 1785–1788 (2008)
10. Gemmeke, J.F., Cranen, B.: Missing data imputation using compressive sensing techniques for connected digit recognition. In: Proceedings of the International Conference on Digital Signal Processing, Santorini, Greece, pp. 1–8 (2009)
11. Gemmeke, J.F., Cranen, B.: Sparse imputation for noise robust speech recognition using soft masks. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Taipei, Taiwan, pp. 4645–4648 (2009)
12. Gemmeke, J.F., Cranen, B., Remes, U.: Sparse imputation for large vocabulary noise robust ASR. Comput. Speech Lang. **25**(2), 462–479 (2011)
13. Gemmeke, J.F., Hurmalainen, A., Virtanen, T., Sun, Y.: Toward a practical implementation of exemplar-based noise robust ASR. In: Proceedings of the EUSIPCO, Barcelona, Spain, pp. 1490–1494 (2011)
14. Gemmeke, J.F., Remes, U., Palomäki, K.J.: Observation uncertainty measures for sparse imputation. In: Proceedings of the Interspeech, Makuhari, Japan, pp. 2262–2265 (2010)
15. Gemmeke, J.F., Van hamme, H., Cranen, B., Boves, L.: Compressive sensing for missing data imputation in noise robust speech recognition. IEEE J Sel. Top. Signal Process. **4**(2), 272–287 (2010)

16. Gemmeke, J.F., Van Segbroeck, M., Wang, Y., Cranen, B., Van hamme, H.: Automatic speech recognition using missing data techniques: handling of real-world data. In: Kolossa, D., Haeb-Umbach R. (eds.) Robust Speech Recognition of Uncertain or Missing Data, pp. 157–185. Springer Verlag, Berlin-Heidelberg (Germany) (2011)
17. Gemmeke, J.F., Virtanen, T., Hurmalainen, A.: Exemplar-based sparse representations for noise robust automatic speech recognition. IEEE Trans. Audio Speech Lang. process. **19**(7), 2067–2080 (2011)
18. Hirsch, H., Pearce, D.: The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proceedings of the ISCA Tutorial and Research Workshop ASR2000, Paris, France, pp. 181–188 (2000)
19. Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, J.: Unlimited vocabulary speech recognition with morph language models applied to Finnish. Comput. Speech Lang. **20**(4), 515–541 (2006)
20. Hurmalainen, A., Mahkonen, K., Gemmeke, J.F., Virtanen, T.: Exemplar-based recognition of speech in highly variable noise. In: International Workshop on Machine Listening in Multisource Environments, Florence, Italy (2011)
21. Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F., Kiessling, A.: Speecon – speech databases for consumer devices: Database specification and validation. In: Proceedings of the of LREC, Las Palmas, Spain, pp. 329–333 (2002)
22. Josifovski, L., Cooke, M., Green, P., Vizinho, A.: State based imputation of missing data for robust speech recognition and speech enhancement. In: Proceedings of the EUROSPEECH, Budapest, Hungary, pp. 2837–2840 (1999)
23. Kallasjoki, H., Keronen, S., Brown, G.J., Gemmeke, J.F., Remes, U., Palomäki, K.J.: Mask estimation and sparse imputation for missing data speech recognition in multisource reverberant environments. In: International Workshop on Machine Listening in Multisource Environments, Florence, Italy (2011)
24. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. **9**, 504–512 (2001)
25. Palomäki, K.J., Brown, G.J., Barker, J.: Techniques for handling convolutional distortion with "missing data" automatic speech recognition. Speech Commun. **43**, 123–142 (2004)
26. Raj, B., Seltzer, M.L., Stern, R.M.: Reconstruction of missing features for robust speech recognition. Speech Commun. **43**(4), 275–296 (2004)
27. Raj, B., Stern, R.M.: Missing-feature approaches in speech recognition. IEEE Signal Process. Mag. **22**(5), 101–116 (2005)
28. Remes, U., Palomäki, K.J., Kurimo, M.: Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition. In: Proceedings of the EUSIPCO, Lausanne, Switzerland (2008)
29. Tan, Q.F., Georgiou, P.G., Narayanan, S.S.: Enhanced sparse imputation techniques for a robust speech recognition front-end. IEEE Trans Audio Speech Lang. Process. **19**(8), 2418–2429 (2011)
30. van den Heuvel, H., Boudy, J., Comeyne, R., Communications, M.N.: The speechdat-car multilingual speech databases for in-car applications. In: Proceedings of the European Conference on Speech Communication and Technology, Budapest, Hungary, pp. 2279–2282 (1999)
31. Van hamme, H.: Robust speech recognition using missing feature theory in the cepstral or LDA domain. In: Proceedings of the EUROSPEECH, Geneva, Switzerland, pp. 3089–3092 (2003)
32. Van hamme, H.: PROSPECT features and their application to missing data techniques for robust speech recognition. In: Proceedings of the INTERSPEECH, Jeju Island, Korea, pp. 101–104 (2004)
33. Van hamme, H.: Robust speech recognition using cepstral domain missing data techniques and noisy masks. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Montreal, Quebec, Canada, pp. 213–216 (2004)

34. Van hamme, H.: Handling time-derivative features in a missing data framework for robust automatic speech recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Toulouse, France (2006)
35. Van Segbroeck, M., Van hamme, H.: Handling convolutional noise in missing data automatic speech recognition. In: Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Toulouse, France, pp. 2562–2565 (2006)
36. Van Segbroeck, M., Van hamme, H.: Vector-Quantization based mask estimation for missing data automatic speech recognition. In: Proceedings of the INTERSPEECH, Antwerp, Belgium, pp. 910–913. (2007)
37. Varga, A., Steeneken, H.: Assessment for automatic speech recognition: II. NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems. Speech Commun. **12**(3), 247–51 (1993)
38. Wang, Y., Van hamme, H.: Multi-candidate missing data imputation for robust speech recognition. EURASIP Journal on Audio, Speech, and Music Processing, No. 17, doi:10.1186/1687-4722-2012-17, May 2012