

Automatic speech recognition using missing data techniques: Handling of real-world data

Jort F. Gemmeke, Maarten Van Segbroeck, Yujun Wang,
Bert Cranen, Hugo Van hamme

Abstract In this chapter, we investigate the performance of a missing data recognizer on real-world speech from the SPEECON and SpeechDat-Car databases. In previous work we hypothesized that in real-world speech, which is corrupted not only by environmental noise, but also by speaker, reverberation and channel effects, the ‘reliable’ features do no longer match an acoustic model trained on clean speech. In a series of experiments, we investigate the validity of this hypothesis and explore to what extent performance can be improved by combining MDT with three conventional techniques, viz. multi-condition training, de-reverberation and feature enhancement. Our results confirm our hypothesis and show that the mismatch can be reduced by multi-condition training of the acoustic models and feature enhancement, and that these effects combine to some degree. Our experiments with de-reverberation reveal that reverberation can have a major impact on recognition performance, but that MDT with a suitable missing data mask is capable of compensating both the environmental noise as well as the reverberation at once.

1 Introduction

Automatic speech recognition (ASR) performance drops rapidly when speech is corrupted with increasing levels of unfamiliar background noise (i.e., noise not seen during training) since the observed acoustic features no longer match the acoustic models. One of the most effective approaches to improving the noise robustness of a speech recognizer is to perform multi-condition training [15]: Rather than training

Jort F. Gemmeke, Bert Cranen
Dept. of Linguistics, Radboud University Nijmegen, The Netherlands., e-mail: \{j.gemmeke,
b.cranen\}@let.ru.nl

Maarten Van Segbroeck, Yujun Wang, Hugo Van hamme
ESAT Department, Katholieke Universiteit Leuven, Belgium e-mail: \{yujun.wang,
maarten.vansegbroeck, hugo.vanhamme\}@esat.kuleuven.be

acoustic models on speech from a quiet environment only, the acoustic models are trained directly on noisy speech signals. By carefully selecting the training speech to reflect the multiple acoustic conditions under which the system must operate, it is possible to minimize the mismatch between training and test/usage conditions. While often effective, recognition accuracies obtained with multi-condition training quickly deteriorate when the noisy environment deviates from the one that was used for training. Another disadvantage of multi-condition training is that the performance for truly clean speech tends to degrade.

Missing Data Techniques (MDT) [25] are a very different approach to improve noise robustness that ideally overcomes the problems with multi-condition training. MDT, first proposed in [5], build on two assumptions: The first assumption is that it is possible to estimate —prior to decoding— which spectro-temporal elements in the acoustic representation of noisy speech are reliable (i.e., dominated by speech energy) and which are unreliable (i.e., dominated by background noise). These reliability estimates are referred to as a *missing data mask*. The second assumption is that the statistics of the features which are considered as dominated by speech energy match with the statistics of clean speech training data. This assumption implies that the acoustic models of MDT recognizers can be trained using clean speech.

In the unreliable elements, the speech information is considered *missing*, and the challenge is then to do speech recognition with partially observed data. In this work, we focus on the so-called *imputation* approach [24], which handles the missing elements by replacing them with clean speech estimates. Classic imputation methods include e.g. correlation and cluster-based reconstruction [23, 25], state-dependent imputation [17] which combines front-end imputation and classifier modification, and the Gaussian-dependent method [32] which additionally allows for reconstruction in the cepstral and PROSPECT domains. The latter method is employed in this chapter.

While imputation has proven effective for increasing noise robustness in the presence of both stationary and non-stationary noise, most of the existing knowledge about the effectiveness of MDT has been acquired using databases with noisy speech that has been constructed by artificially adding noise of various types and intensities to clean speech (see e.g. [6, 23]). Using artificially corrupted data is attractive as it allows creating a missing data mask based on exact knowledge of the speech and noise power in each time-frequency cell. This facilitates comparison of different MDT approaches and allows for analysis of the influence of errors in reliability estimation.

Real-world recordings, however, are generally not only corrupted by background noise, but can also be affected by room acoustics. Moreover, real-world recordings are more likely to introduce a mismatch between the observed speech and the speech on which the recognizer is trained, due to microphone characteristics and speaker specific behavior such as lip-noises and Lombard effect. Very few reports exist that describe the effectiveness of single-channel MDT recognition on real-world recordings (notable exceptions are [13, 19, 27]). In previous research we have used the SPEECON [16] and SpeechDat-Car [30] databases for that purpose. The SPEECON and SpeechDat-Car databases are recorded in realistic environments

such as in an entertainment room, office, public hall and car. The databases contain simultaneous recordings from four microphones placed at different distances from the speaker, one of them being a close-talk microphone. Thus, SPEECON and SpeechDat-Car make it possible to investigate the impact of different degrees of natural distortions (background noise and reverberation) on the performance of ASR systems. Specifically, since the close-talk microphone could be considered as an approximation of ‘clean speech’, these corpora make it possible to investigate the performance of MDT on real-world speech using an approach similar to what has proven so effective with artificially corrupted databases.

We have found that a MDT recognizer that is trained with speech from the close-talk microphone is not very robust against the distortions that are present in the speech recorded with the three other (far-talk) microphones [13]. Moreover, even when using information from all available channels to estimate a ‘cheating’ missing data mask, the so-called *semi-oracle mask*, we obtained much lower accuracies than previously obtained on similar recognition tasks (such as AURORA-4 [22]) with artificiality corrupted speech [35]. We hypothesize that this is due to a violation of the second assumption underlying MDT, namely that the statistics of the features that are not dominated by background noise match with the statistics of the features from the close-talk microphone. Experiments with artificially added noise all but guarantee that the second assumption holds true: If in some spectro-temporal element the speech energy is higher than the noise energy, the observed signal will fit the distribution of the clean training data. With real-world recordings, however, the speech in the other recording channels is not only affected by additive noise, but also by microphone characteristics and reverberation. This has the effect that the ‘reliable’ features, while dominated by speech energy, still mismatch the trained speech features. As a result, imputation and recognition accuracy are bound to suffer.

In this chapter, we test this hypothesis and explore whether recognition accuracy can be improved by combining MDT with three conventional techniques, multi-condition training, de-reverberation and spectral subtraction. First, we extend the MDT approach, in which the recognizer is trained on close-talk channel ‘clean’ speech, by using acoustic models that are trained on multi-condition training material from all recording channels. In doing so, we assume that the proven techniques for estimating missing data masks and for imputing missing data can also be applied to real-world speech. Second, the availability of four parallel channels in the SPEECON and SpeechDat-Car databases makes it possible to detect strong reverberation and to create a new kind of ‘cheating’ missing data mask that labels time-frequency cells dominated by reverberation as ‘unreliable’. This missing data mask, reminiscent of the de-reverberation technique used in [7, 21], allows us to investigate the impact of reverberation on the recognition accuracy and to explore whether a combination of de-reverberation and MDT will improve performance. Third, we investigate the performance obtainable with feature enhancement techniques on real-world recordings and whether feature enhancement can be combined with MDT, either to improve missing data mask estimation or to replace multi-condition training as a means for diminishing the hypothesized mismatch between training and test/use conditions.

The rest of the chapter is organized as follows. In section 2 we introduce MDT and the imputation method used in the chapter. In section 3 we describe the isolated word recognition task used in our experiments. In section 4 we describe the missing data mask estimation techniques used for MDT and the decoding architecture used in later sections. In section 5 we present the experiments with clean and multi-condition acoustic models. In section 6 we investigate the combination of MDT and de-reverberation and in section 7 we investigate use of feature enhancement in combination with MDT. Finally, we have a general discussion and present our conclusions in section 8.

2 MDT ASR

2.1 Missing data Techniques

In this section, we briefly review the MDT framework [5, 25]. In ASR, the basic representation of speech is a spectro-temporal distribution of acoustic power, a *spectrogram*. 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, often a MEL-frequency scale and logarithmic compression of the power scale are employed. We denote the (MEL-frequency) log-power spectrograms of noisy speech as \mathbf{Y} , of clean speech as \mathbf{S} , and of the background noise as \mathbf{N} . Elements of \mathbf{Y} that predominantly contain speech or noise energy are distinguished by introducing a missing data mask \mathbf{M} . The elements of a mask \mathbf{M} are either 1, meaning that the corresponding element of \mathbf{Y} is dominated by speech ('reliable') or 0, meaning that it is dominated by noise ('unreliable' c.q. 'missing'). Thus, we write:

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

with \mathbf{M} , \mathbf{Y} , \mathbf{S} , and \mathbf{N} two-dimensional matrices of size $K \times T$, with frequency-band index k , $1 \leq k \leq K$ and time-frame index t , $1 \leq t \leq T$. θ denotes a constant SNR-threshold.

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 [25] and can be used directly as estimates of the clean speech features.

In the real-world speech recorded by multiple microphones considered in this chapter it is questionable whether features that are labeled reliable with such a procedure remain approximately uncorrupted. Most speaker effects (like the Lombard effect) will show up equally in all recording channels. Environmental noise, channel

effects and reverberation, however, are likely to affect the different channel recordings differently. A fundamental problem is thus the definition of ‘clean’ speech and ‘noise’ underlying (1). Even if a close-talk microphone signal is used for training the acoustic models as an approximation of ‘clean’ speech, as was done in previous work [13], the ‘noise’ in the far-talk channels actually constitutes not only the environmental noise, but also extra feature variation due to the way in which channel characteristics and reverberation have affected the speech energy. Conventional mask estimation techniques, however, make the distinction between features dominated by speech or background noise by searching for spectro-temporal elements that have the characteristics of speech. As a result, the resulting ‘reliable’ features retain any channel and reverberation effects. As a consequence, ‘reliable’ features that are determined in the conventional way are likely to mismatch the statistics of the features in the close-talk channel used for training.

In sections 5 and 7 we explore the impact of reducing this mismatch between the acoustic model and the ‘reliable’ features. In section 6 we take the opposite approach and see if we can reduce a part of the mismatch by considering the reverberated speech features as ‘noise’ and modify the missing data mask accordingly.

2.2 Gaussian-dependent imputation

Originally, MDT was formulated in the log spectral domain [5]. Here, speech is represented by the log-energy outputs of a filter bank and modeled by a Gaussians 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 [6].

Later, it was found the method could be improved by using state- [17] or even Gaussian-dependent [31] clean speech estimates. In these approaches, the unreliable features are imputed during decoding and effectively depend on the hypothesized state identity. However, filter bank outputs are highly correlated and poorly modeled with a GMM with a diagonal covariance. This is the reason why conventional (non-MDT) speech recognizers employ cepstral features, obtained by applying a de-correlating Discrete Cosine Transformation (DCT) on the spectral features.

In [31] a technique was proposed to do Gaussian-dependent (bounded) imputation in the cepstral domain. The drawback of that technique was the increased computational cost, because the imputation of the clean speech was done by solving a Non-negative Least Square (NNLSQ) problem. The Gaussian-dependent imputation approach used in this chapter [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.

3 Real-world Data: the SPEECON and SpeechDat-Car databases

In the research reported in this chapter we used the SPEECON [16] and the SpeechDat-Car [30] databases. These databases contain speech recorded in realistic environments with multiple microphones. There are four recording environments: office, public hall, entertainment room and car. The office, public hall and entertainment room material stems from the SPEECON database and contains multichannel data with each channel corresponding to a different microphone position: channel #1 is a headset microphone, #2 a lavalier microphone and #3 and #4 are medium and far distance microphones placed at 0.5 to 3 meter from the speaker. The car environment contains material from both the SPEECON and SpeechDat-Car databases, with again channel #1 a headset microphone and #2 a lavalier microphone, while the channel #3 microphone was placed behind the rear-view mirror and the channel #4 microphone was placed near the rear window (SpeechDat-Car) or near the rear-view mirror (SPEECON). The speech material is recorded with a 16 kHz sampling rate.

The use of these databases is a middle ground between the artificially corrupted speech as found in for example the AURORA databases [14, 22] on the one hand, and the complex real-world conditions on the other.

3.1 *Isolated word test set*

For our recognition experiments, we used a subset of the isolated word data in the Flemish part of the SPEECON and the SpeechDat-Car databases. These isolated word data contains command words, nouns and verbs. We constructed a test set containing a balanced mixture of SNR conditions. Using the SNR estimates obtained in [13] we created 6 SNR subsets, each with a 5 dB bin width, spanning a 0 dB to 30 dB range. The SNR subsets were filled by randomly selecting 700 utterances per SNR subset, ensuring a uniform word occurrence. 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 contain channel #4 speech.

The resulting test set contains 16,535 utterances¹, with 565 unique words, 54 minutes of speech embedded in 13 hours of audio signal. The test set is spoken by 232 speakers, 115 male and 117 female.

¹ The observant reader will have noticed that the total number of words does not add up to $4 \times 6 \times 700 = 16,800$. This is because one subset (the entertainment room environment in the $[0 - 5]$ dB SNR bin) only contains 435 utterances rather than 700 due to data scarcity.

3.2 Training sets

The clean training set contains 40 hours of speech embedded in 63.5 hours of signal. Among the utterances used for training are command words and read sentences. All the 61,940 utterances in this set are from channel #1 data, with an estimated SNR range of 15 to 50 dB. The clean training set was spoken by 191 speakers, 82 of them are male and 109 female. There is no overlap between speakers in the test and training sets.

The multi-condition train set contains 127 hours of speech embedded in 205 hours of signal, 231,849 utterances in total. Beside all channel #1 data included in the clean training set, the multi-condition set contains all utterances from channels #2, #3 and #4 which have an estimated SNR of 10 dB and higher. The 10 dB cut-off is necessary to prevent frame/state alignment issues during training and to ensure the acoustic models trained on this data remain sufficiently discriminative. The multi-condition training set thus contains an additional 55 hours of channel #2 data (54,381 utterances), 54 hours of channel #3 data (53,248 utterances) and 32 hours of channel #4 data (31,975 utterances). While the training sets differ in size, they do not differ in terms of speech-related observations since the data stems from multi-channel recordings.

The speech in the training set is taken from three noise environments: office, public hall, and car. The channel #1 speech used in the clean training set, 63.5 hours of signal in total, is composed of 47 hours of signal from the office environment, 2.2 hours from the public hall environment and 14.3 hours from the car environment. For the multi-condition model, the 205 hours of signal originate from 168 hours of signal recorded in the office environment, 6 hours from the public hall environment and 31 hours from the car environment.

4 Experimental setup

4.1 Mask estimation

4.1.1 Semi-oracle masks

With the recordings used in this paper, the underlying clean speech and noise are not known exactly. As a consequence, oracle masks useful for obtaining an estimate of the upper bound on recognition performance with MDT, cannot be computed. We can, however, use the multi-channel data to estimate the so-called *semi-oracle mask*. In order to calculate this mask, we use the channel #1 data, which is obtained from a headset microphone, as an estimate for the underlying clean speech in the other channels. In order to compensate for the delay and microphone differences between channel #1 and the other channels, we use an acoustic echo canceler (AEC) to predict the clean speech component.

By minimizing the (energy) difference between a filtered version of the channel #1 signal and a far-talk microphone signal, the AEC estimates a Finite Impulse Response (FIR) which can be considered as the best possible estimate of the transmission path from the close-talk microphone to the far-talk microphone. Thus, the remaining differences between the filtered far-talk channel and the unfiltered original can be attributed to the noise in the far-talk channel and can serve as a noise estimate. By thresholding the difference between the speech and noise estimates using (1) we obtain the semi-oracle mask.

For the AEC, we used the PEM-AFROW algorithm [38], using second order pre-whitening filters and a 25 ms FIR filter. Because we cannot guarantee that the distance between the speaker and the microphone is constant for all utterances in a session, the filters are re-estimated for every utterance and multiple iterations over the same utterance are used to improve the convergence. Since this is a ‘cheating’ missing data mask, we manually selected the optimal mask threshold after recognition over a large interval of threshold values for each recording environment and each acoustic model.

4.1.2 Vector Quantization Masks

As a first approach to estimate spectrographic masks from a single recording channel, we employ the Vector Quantization (VQ) strategy proposed in [37]. Here, the key idea is to estimate masks by making only weak assumptions about the noise, while relying on a strong model for the speech. The speech model is expressed as a set of codewords (a codebook) containing the periodic and aperiodic part of training speech. The periodic part consists of the harmonics at pitch multiples and the remaining spectral energy is considered the aperiodic part. Both parts are obtained using the harmonic decomposition method described in [33].

During decoding, we apply the harmonic decomposition to the observed speech. We then use the periodic and aperiodic part of the observed speech to recover a clean speech estimate from the set of stored codewords by minimizing a cost function that is robust against additive noise corruptions. The aperiodic part of the observation is used to provide a noise estimate by taking its long-term minimum as in [20]. Finally, the spectrographic VQ-based mask is estimated by thresholding the ratio of speech and noise power estimates using (1). To compensate for linear channel distortions, the VQ-system self-adjusts the codebook to the channel during recognition.

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 *Voice Activity Detector* (VAD), segments speech from non-speech frames in order to restrict mask estimation to frames containing (noisy) speech. For a frame labeled as non-speech, all mask values are set to zero, indicating that all components are unreliable.

The VQ-codebook was trained on features extracted from the close-talk channel SPEECON training database. The number of codebook entries was 500. The VAD was inspired by the integrated bi-spectrum method described in [26]. 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 $\theta = 8$ dB.

4.1.3 SVM masks

A different approach to mask estimation is to use machine learning to classify each feature as either reliable or unreliable. A machine learning algorithm can be used to associate noisy speech features with reliability scores that are obtained from suitable training material. Such training material must necessarily consist of oracle masks and therefore requires the use of artificially corrupted clean speech for training.

In [28] it was proposed to use a Bayesian classification approach for mask estimation. In this work, we use Support Vector Machine (SVM) classifiers, a machine learning algorithm which is known for its excellent performance on binary classification tasks and generalization power when trained on relatively small data sets [3]. From the machine learning perspective, mask estimation is a multi-class classification problem with 2^K classes. Since such high-dimensional multi-class classification is infeasible, we assume that the reliability estimates are independent between frequency bands and trained a separate SVM classifier for each of the K MEL-frequency bands.

Each classifier used the same set of single-frame-based $7 \times K + 1$ -dimensional features consisting of: the K -dimensional noisy speech features themselves, the harmonic and aperiodic part and long-term energy noise estimate described in section 4.1.2, the gain factor described in [33], the ‘Sub-band Energy to Sub-band Noise Floor Ratio’ and ‘Flatness’ features derived from the noisy MEL-spectral features described in [28], and finally a single VAD feature. The training material was taken from another corpus, AURORA-4 [22], which contains artificially noisified Wall Street Journal (WSJ) utterances.

SVMs were trained using LIBSVM [4] on 75000 frames (amounting to 12.5 minutes of audio signal) randomly extracted from the AURORA-4 multi-condition training set. Reliability labels used in training were obtained from the oracle mask, derived by using the (available) clean speech and noise sources in (1) with $\theta = -3$ dB (cf. [37]). We used an RBF-kernel and hyper-parameters were optimized by doing 5-fold cross validation on the training set.

4.2 Recognizer setup

4.2.1 Recognizer

The MDT-based recognizer was built by adding the required MDT modifications to the speaker-independent large vocabulary continuous speech recognition (LVCSR) system that has been developed by the ESAT speech group of the K.U.Leuven;

cf. [1] for a detailed description of the system. This recognizer was chosen because of its fast experiment turn-around time and good baseline accuracy. Decoding is done with a time-synchronous beam search algorithm.

The recognition performance will be expressed in terms of the word error rate (WER), which is defined as the number of word errors, i.e. insertions, deletion and substitution errors, divided by the total number of words in the reference transcription. The word startup cost was tuned over all noise environments and channels jointly, but has only minor importance given the nature of the task.

The (word-independent) word insertion penalty was tuned over all noise environments and channels jointly by maximizing recognition accuracy. The word insertion penalty only marginally affects the accuracy via the pruning mechanism because in an isolated word task the same penalty is applied to all hypotheses.

4.2.2 Preprocessing

The acoustic feature vectors consisted of MEL-frequency log power spectra: $K = 22$ frequency bands with center frequencies starting at 200 Hz (the first MEL-band is not used). 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. During training, mean normalization is applied to the features. During decoding, the features are normalized by a more sophisticated technique which works by updating an initial channel estimate through maximization of the log-likelihood of the best-scoring state sequence of a recognized utterance [36]. In the MDT experiments, as described in section 2.2, the spectra and their derivatives are transformed to the PROSPECT domain [32] during decoding. Missing data masks for the derivative features were created by taking the first and second derivative of the missing data mask [34]. In the uncompensated baseline experiments, the MEL-frequency features are transformed using the Mutual Information Discriminant Analysis (MIDA) linear transformation [9]. The MIDA transformation maximizes class separation much like Linear Discriminant Analysis (LDA) does, but is based on a mutual information criterion.

4.2.3 Acoustic model training

First, for each of the two training sets an acoustic model was trained based on the MIDA feature representation. After training, the clean speech model contains 2534 tied states using 28,917 Gaussians. Because the multi-condition training data is larger in size and richer in variation (the clean data plus its noisy variants), more tied states (4476) and slightly more Gaussians (32,747) are retained by the decision tree inference algorithm. We have chosen to allow the multi-condition model to exploit the augmented data set to maximize its accuracy and hence not to constrain its size. For the MDT experiments, we then created two new acoustic models in

the PROSPECT domain by single-pass re-training. This retraining procedure consisted of replacing the means and variances of the MIDA acoustic model with their PROSPECT counterparts.

For each of the two training sets an acoustic model was trained based on the MIDA feature representation. This is achieved in several steps. First, a set of 46 context-independent phone models plus four filler models and a silence model are trained using Viterbi re-estimation. Each HMM state has a set of up to 256 unshared Gaussians. Subsequently, a phonetic decision tree (c.f [10]) defines the 2534 tied states (for the clean training data) in the cross-word context-dependent models. The final acoustic models are obtained by allowing sharing across all Gaussians and subsequently retaining only those with maximum occupancy [11], resulting in an average of 96.4 Gaussians per state for the clean training data.

Because the multi-condition training data is larger in size and richer in variation (the clean data plus its noisy variants), more tied states (4476) and slightly more Gaussians (32,747) are retained by the decision tree inference algorithm. We have chosen to allow the multi-condition model to exploit the augmented data set to maximize its accuracy and hence not to constrain its size. On average, 90.2 Gaussians per state are retained in this model. For the MDT experiments, we then created two new acoustic models in the PROSPECT domain by single-pass re-training. This retraining procedure consisted of replacing the means and variances of the MIDA acoustic model with their PROSPECT counterparts.

5 MDT and multi-condition training

In this section, we investigate the effectiveness of a classical MDT recognizer on speech recorded in real-world environments in combination with a multi-condition trained acoustic model. To that end we determined the recognition accuracy using a number of different mask estimation methods: the semi-oracle mask (cf. section 4.1.1), the VQ mask (cf. section 4.1.2) and the SVM mask (cf. section 4.1.3). Each mask estimation method was tested using two different acoustic models: a model trained on the clean speech training set and a model trained on the multi-condition training set. In order to give a baseline recognition result, we also discuss recognition experiments without using any additional noise-robust preprocessing other than what's inherent in the acoustic model and channel compensation. In section 5.1 we describe the results of our experiments and in section 5.2, we discuss the results.

5.1 Recognition results

The speech recognition results from our experiments, depicted as word error rate (WER) as a function of SNR are displayed in Fig. 1. The left pane corresponds

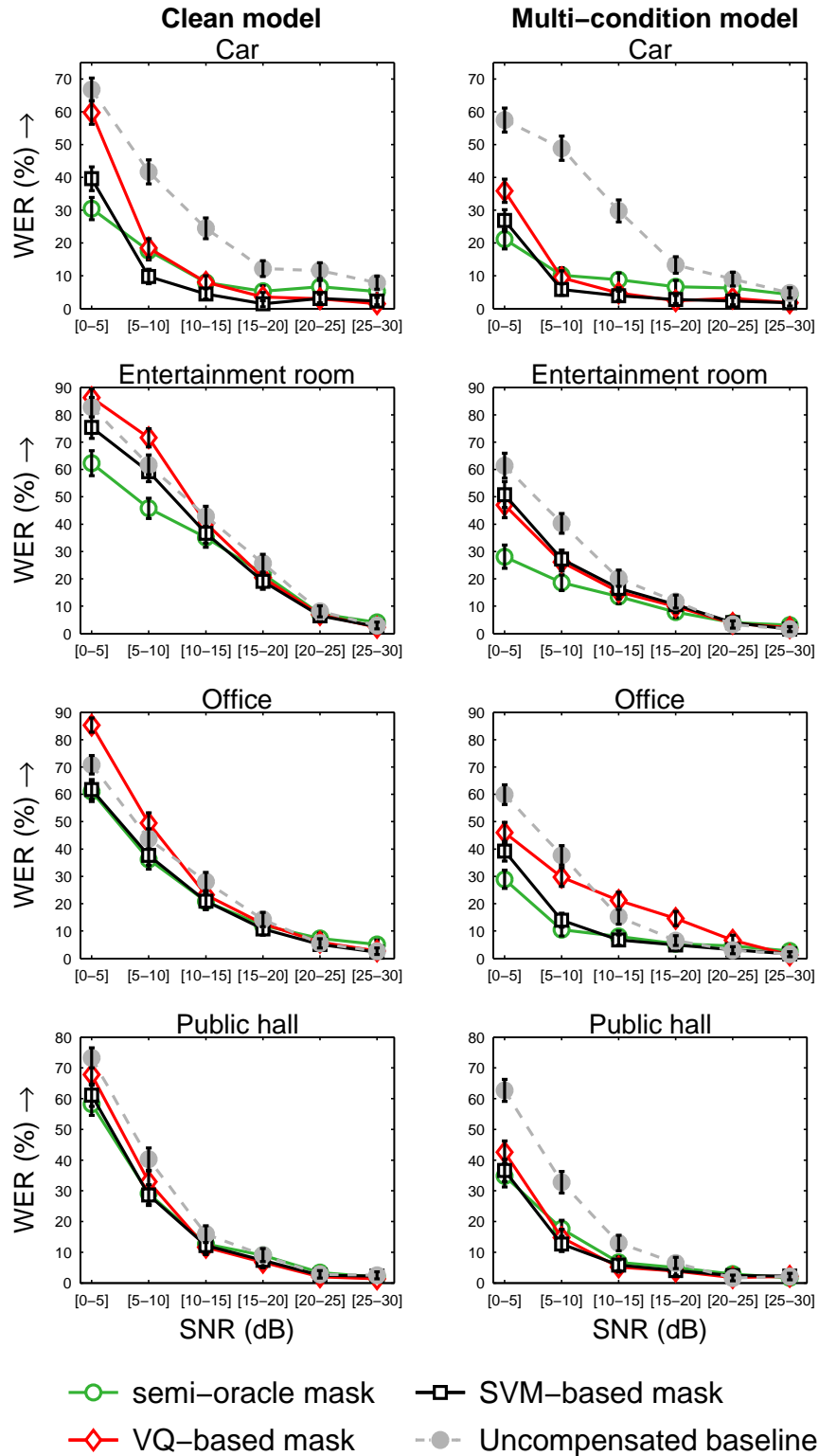


Fig. 1 Word error rate (WER) as a function of SNR displayed for clean speech (left) and multi-condition trained models (right). From top to bottom the rows represent different noisy environments, viz. car, entertainment room, office, and public hall. In each panel the results are shown for the uncompensated baseline, semi-oracle mask, VQ mask and SVM mask. Vertical bars around the data points indicate 95% confidence intervals.

to recognition using the clean acoustic model while the right pane corresponds to recognition with the multi-condition acoustic model. From top to bottom, the respective rows represent different noise environments, viz. car, entertainment room, office and public hall, respectively.

Within each plot we display the results of the following methods:

- The uncompensated baseline, with no further noise robustness processing beyond what’s inherent in the channel compensation or acoustic model training.
- MDT with the ‘cheating’ semi-oracle mask described in section 4.1.1
- MDT with the VQ-based missing data mask described in section 4.1.2
- MDT with the SVM-based missing data masks described in section 4.1.3

In the top left plot of Fig. 1, corresponding to recognition in the car environment using a clean speech acoustic model, we can observe large differences between the approaches. It is apparent that all MDT approaches improve substantially over the baseline. The ‘cheating’ semi-oracle mask (SO) performs better than the other missing data masks in the 0 – 5 dB range; at higher SNRs it is outperformed by the estimated masks. In the highest SNR bins the VQ mask and the SVM mask perform comparably, while at lower SNRs the SVM mask performs better than the VQ mask.

When using a multi-condition acoustic model (right pane, top row), the WERs at SNRs below 10 dB are much lower. Especially the VQ mask benefits, achieving up to 24% lower WERs (absolute) in the 0 – 5 dB range. The baseline, on the other hand, has performance gains of $\approx 10\%$ at lower SNRs. The ranking of the missing data methods is roughly the same as when using the clean acoustic model, although the differences between the methods are much smaller. Importantly, there is no significant difference in speech performance at high SNRs when using a multi-condition acoustic model: 1.5% WER (multi-condition model) vs 1.8% (clean model) for the VQ mask.

Compared to the car environment, the entertainment room (second row of Fig. 1), is a more challenging environment: even the semi-oracle, which does best in this environment, has a 62.3% WER in the 0 – 5 dB range when using a clean acoustic model. While SVM performs up to 8% better (absolute) than the baseline, the VQ mask performs worse than the baseline in the 0 – 10 dB range. As before, there is no significant difference between the methods at high SNRs. When using a multi-condition acoustic model, there is again a substantial overall drop in WER. As in the car environment, the VQ mask benefits especially and now performs comparably to the SVM mask.

In the office and public hall environments (third and fourth rows in Fig. 1), we can observe many of the same trends described for the entertainment room environment. The SVM and semi-oracle masks perform comparably, and both methods perform up to $\approx 12\%$ better (absolute WER) than the baseline. The VQ mask does worse in the office environment but comparably in the public hall environment. When using a multi-condition acoustic model, overall WERs are much lower and the gap between MDT and the baseline is much bigger.

5.2 Discussion

5.2.1 Effect of multi-condition training

Comparing the uncompensated baseline scores in Fig. 1 with those obtained with MDT, it is clear that MDT manages to substantially improve upon the clean model baseline, reaching comparable recognition accuracies as the multi-condition model baseline. Moreover, comparing the left and right-hand panes of Fig. 1 we can see that the use of a multi-condition acoustic model improves MDT recognition accuracy substantially, especially at lower SNRs. In fact, the performance increase when using a multi-condition model with MDT is much larger than for the uncompensated baseline. From these results we conclude that our hypothesis — that ‘reliable’ features do not match the statistics of the clean acoustic model — is correct.

The mismatch of the reliable features with the acoustic model has two causes. First, mask estimation techniques make errors and sometimes unjustly label features dominated by noise as ‘reliable’, *false reliables*. On real-world data, conventional mask estimation techniques do not take into account that the speech signal can be corrupted by channel and reverberation effects as well as by environmental noise. As a result, speech features that should have been masked because they are dominated by any of these effects are also unjustly labeled ‘reliable’. The multi-condition model (partly) corrects false reliables because the acoustic model matches a much larger variance of the speech features. Second, even if all features which are not too heavily affected by additive noise or reverberation would correctly be labeled as ‘reliable’, the resulting features can, in contrast to artificially noisified data, still mismatch the speech distributions trained on close-talk channel data due to remaining microphone characteristics and reverberated speech energy. The multi-condition acoustic models will also compensate for this effect.

5.2.2 Mask estimation accuracy

The semi-oracle mask, a ‘cheating’ mask created with knowledge of all channels, in general hardly performs better than the estimated masks except in the lowest SNR bins. And even there, the differences are small, unlike the performance differences between estimated masks and ‘true’ oracle masks in experiments with artificially noisified data [14, 22]. While it cannot be established what the performance of a ‘true’ oracle mask would be, especially given the test/training mismatch issues discussed above, we can point out two shortcomings of the semi-oracle mask. First, the semi-oracle is derived under the assumption that the close-talk signal can be considered as ‘clean’ speech, resulting in all-reliable masks for close-talk channel speech. Even the features of close-talk channel speech, however, may be occasionally be corrupted and should have been labeled ‘unreliable’. Second, the AEC captures not only the transmission path between the close-talk microphone and a far-talk microphone, but also reverberation. The semi-oracle thus does not label speech dominated

by reverberation of channel effects as unreliable. We will explore this effect in more detail in section 6.

Compared to the other mask estimation methods, the VQ mask has a lower performance and in various conditions does even worse than the baseline. When using a mask derived with multi-condition acoustic models, however, the VQ mask performs much better. As with the semi-oracle mask, this is probably because the codebook is created using channel #1 speech, under the assumption it contains clean speech. Because in the other channels the observed speech results in harmonic decomposition components that will often be poorly described by the codebook, many reliable features are likely to get mis-labeled as unreliable or vice versa.

The SVM mask generally performs very well, often performing comparably to the semi-oracle mask. Its performance is a testament to the generalizability of SVM-based machine learning; after all, the mask estimation was trained using the AURORA-4 corpus. The use of this corpus, which contains noisified Wall Street Journal (WSJ) speech, means there is a mismatch in noise types, language and content. Still, it generally performs better than the VQ mask, even though they share many of the features such as the harmonic decomposition. Moreover, the SVM mask does not require the tuning of a threshold parameter. A downside of the SVM method, not discussed in this chapter, is its high computational cost.

6 MDT and de-reverberation

In order to investigate whether MDT can be combined with de-reverberation and to what extent reverberation can affect the recognition performance of our MDT recognizer, we performed two experiments that will be described in more detail in this section.

In the first experiment, we determined to what extent MDT can be used to improve recognition accuracy of artificially reverberated speech. In [7,21] it was shown that treating features which are dominated by reverberation as unreliable, can be quite effective when using clean speech models. Here, we investigate to what extent this approach can be combined with the use of multi-condition models to provide robustness for the artificial reverberation. Using artificially reverberated speech constructed by filtering clean speech with a known room impulse response filter, we created an oracle mask by considering the difference between the reverberated and the non-reverberated versions of the speech as ‘noise’. Using the conventional mask definition of (1) and the previously trained clean speech models and multi-condition models (cf. section 4.2.3), we investigated to what extent the recognition accuracy of our MDT recognizer can be improved using this oracle mask that labels features unreliable when they are dominated by reverberation.

In the second experiment we applied the insights gained from the first experiment on real-world data. Under the assumption that the channel #1 data has negligible reverberation and that the reverberation effects become more pronounced in microphone channels #2 – #4, we use the estimated room impulse response filter

of the AEC to estimate the underlying non-reverberated speech signal. Thus, we can construct, similarly as with the artificially reverberated speech, an improved (semi-)oracle mask that not only labels features unreliable when they are affected too severely by environmental noise, but also when they are dominated by reverberation. By comparing improvements in recognition accuracy, both for acoustic models that were trained on clean speech and for multi-condition models, we try to estimate an upper and lower bound on the impact reverberation can have in our MDT recognition experiments on real data.

6.1 Experimental setup

6.1.1 MDT for de-reverberation of artificially reverberated data

We created artificially reverberated speech as follows. First, we measured two room impulse responses (RIR) using a microphone at 261 cm from the speaker. The room of 36 m³ has curtains on all walls. In the first RIR, the curtains were closed ($T_{60} = 140$ ms), while in the second RIR they were open ($T_{60} = 250$ ms). The RIRs were measured with Gaussian white noise excitation using a least-squares estimation approach. The resulting FIR filter had a length of 125 ms (2001 coefficients at 16 kHz). Next, these two FIR filters were applied to the channel #1 utterances from all four environments in the 20 – 25 dB bin of our SPEECON and SpeechDat-Car data. This results in two new artificially reverberated test sets, each containing 1236 utterances.

Subsequently, we created a delayed version of the original signal by filtering it with the same FIR, but with the tail of the filter coefficients (representing the echo’s from the non-direct path) zeroed out. This was done by manually setting all FIR filter coefficients of the AEC filter beyond 3 ms after the first peak to zero. Then, we calculated the residual between the delayed and the reverberated channel #1 data. Using the residual as the ‘noise’ and with the delayed channel #1 data taking the place of the clean speech, we applied (1) to obtain our oracle mask. This oracle mask was then used in the MDT recognizer to decode the artificially reverberated signal.

In order to obtain the optimal oracle mask, we performed experiments with a large number of SNR thresholds. The results that will be presented pertain to the oracle masks obtained with the threshold that resulted in the best accuracy. Tuning the threshold on the test data is justified by the fact that with these artificial test data we are only interested in an estimate of the upper bound of the improvement that can be achieved.

As before, all experiments are done both with the acoustic models trained on clean speech and multi-condition speech. No new models were trained for the experiments described here.

6.1.2 MDT using a reverb-masking semi-oracle mask on real-world data

In this experiment, we try to estimate which spectro-temporal features in the SPEECON and SpeechDat-Car data are associated with speech energy that reaches the microphone via a direct path rather than being dominated by first and higher order reflections or additive noise. To that end, we estimate a de-reverberated version of the clean speech signal by modifying the room impulse response filter in the AEC described in section 4.1.1. Analogously with the experiments on artificially reverberated data, the de-reverberated clean speech estimate is obtained by manually setting all FIR filter coefficients of the AEC filter to zero beyond 3 ms after the first peak. Thus, the resulting FIR filter ideally should only represent the transfer function of the direct path between the two microphones, discarding any reflections caused by the room acoustics.

Next, we consider the residual of the features of the observed signal and the non-reverberated clean speech estimate as noise. We use (1) to label only those speech features reliable which can be assumed not to be excessively affected by additive noise or reverberation. This improved semi-oracle mask, which we will denote by reverb-masking semi oracle (RMSO) mask, is then used to decode the original noisy speech features. This allows us to test whether or not in our framework, masking reverberated speech can improve recognition accuracy as it did in [7, 21].

As before, only the thresholds with the best accuracy are shown and all experiments are done both with the acoustic models trained on clean speech and with multi-condition models.

6.2 Results and discussion

6.2.1 MDT for de-reverberation of artificially reverberated data

In Fig. 2 we can observe a clear trend that on artificially reverberated speech, masking the reverberation improves performance. This holds both for the clean speech model (left panel) and for the multi-condition model (right panel). Moreover, comparing the overall performance of the multi-condition model and the clean model in the various reverb conditions, clearly shows that the multi-condition model is the most robust against the artificial reverberation. Surprisingly the performance for the no-reverb condition shows a similar trend: with the multi-condition model performance is better than with the clean model. Although we do not have a solid explanation why the apparent mismatch between the current test set and the clean training data would be greater than with the multi-condition training data, we take this observation as yet another illustration that recognition with a clean speech model is sensitive to even the slightest training/test mismatch and that such a mismatch can often be compensated by using a multi-condition model.

In the case of the reverberation caused by a RIR of a room with closed curtains, the oracle missing data mask is able to completely compensate for the performance

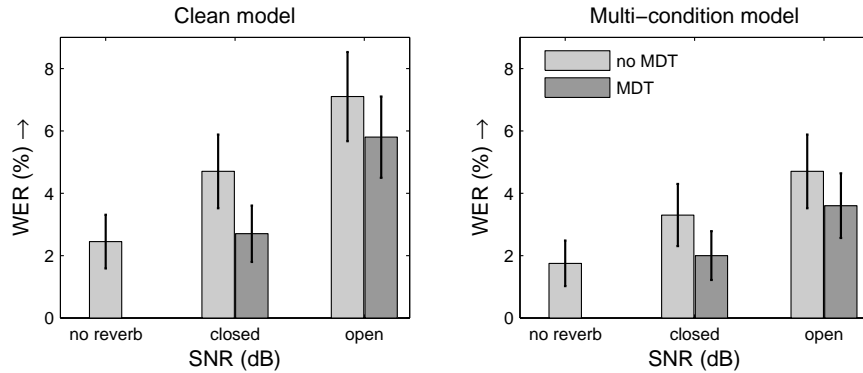


Fig. 2 Word error rate (WER) for recognition with clean (left) and multi-condition models (right). Vertical bars around the maxima indicate 95% confidence intervals. In each bar-graph results are shown for the non-reverberated ‘clean’ test set, the reverberated test set with closed curtains and the reverberated test set with open curtains.

loss due to reverberation, yielding a performance which is indistinguishable from the no-reverb condition. This holds true both for the clean speech model and the multi-condition model.

In summary, these results suggest that loss in recognition accuracy due to reverberation can be compensated by using a multi-condition acoustic model as well as by a suitable missing data mask which labels the features affected by reverberation unreliable. In fact, it seems that these two approaches are to some extent complementary and can be combined to combat reverberation.

6.2.2 MDT using a reverb-masking semi-oracle mask on real-world data

In Fig. 3 we observe that when using the clean acoustic model, masking the reverberation generally increases performance substantially. Especially in the entertainment room and public hall environments, performance differences can be up to 23% (absolute WER). From this we conclude that we were successful in estimating which features were excessively affected by reverberation and therefore should be labeled unreliable, and thus better approximate the ‘true’ oracle mask (cf. section 5.2.2). Moreover, these results show that MDT can be used to compensate both noise and reverberation at once by using a suitably chosen missing data mask.

In none of the environments, however, does the reverb-masking semi-oracle mask (RMSO) in combination with a clean acoustic model perform better than the original semi-oracle mask (SO) in combination with a multi-condition acoustic model. This means that masking the reverberation effects (using the modified RIR filter approach) does not account for all mismatch between the noisy speech features and the clean acoustic model. When using the RMSO mask in combination with a multi-condition model, the results vary between environments. In the entertain-

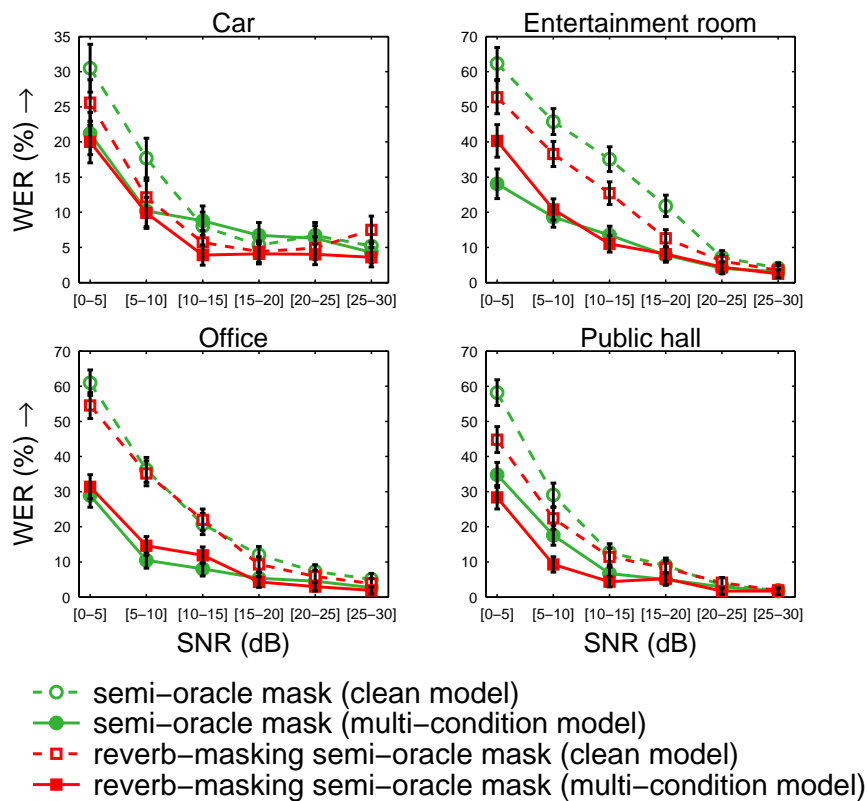


Fig. 3 Word error rate (WER) for four different noise environments. In each panel the results are shown for recognition with the semi-oracle masks in combination with the clean acoustic model, semi-oracle mask with the multi-condition acoustic model, and reverb-masking semi-oracle mask with the clean or multi-condition acoustic model. Vertical bars around the data-points indicate 95% confidence intervals.

ment room and office environments, the performance is worse than when using the SO mask. This implies that in these environments, the reverberation is already fully compensated by the multi-condition model, and masking more features only results in masking features that are useful for imputation or recognition.

In the car environment, the RMSO and SO masks perform comparably at high and low SNRs, with the RMSO mask performing better in the 5 – 25 dB SNR range. In the public hall environment, which is the most reverberant environment, masking the reverberation lowers the WER by $\approx 7\%$ (absolute) at SNRs < 15 dB. It seems that in the public hall environment, the impact of reverberation is substantial in the channels that contribute most to the lower SNR bins (i.e., channels #3 and #4).

From these results we can roughly estimate an upper bound on the impact of reverberation on our recognition results by taking the difference between the WER obtained with worst performing method (SO mask with a clean acoustic model), and

the best performing method (usually RMSO mask when using the multi-condition acoustic model). This would imply that at most 10% to 35% of the errors in the lower SNR bins, depending on the environment, can be attributed to reverberation.

For some environments, we can also try to establish a lower bound by taking the difference in performance obtained with the RMSO mask and the original SO mask in combination with the multi-condition acoustic model. The rationale behind this would be that the multi-condition acoustic model already accounts for various sources of variation, to some extent including reverberation. So if explicitly masking reverberation helps, it must be due to the reverberation that multi-condition models did *not* capture. Following this approach, we might conclude that in the public hall environment, at least at least about an 7% WER loss (absolute) at SNRs under 15dB is due to reverberation.

7 MDT and feature enhancement

In the sections above we successfully combined MDT with conventional noise robustness techniques, such as multi-condition training and de-reverberation. In section 5 it was shown that replacing a clean speech model by a multi-condition acoustic model dramatically improves results, confirming the hypothesis that the ‘reliable’ features no longer match the clean speech distributions. We argued that the improvement of using a multi-condition model is partly caused by a greater robustness against mask estimation errors. Multi-condition training material, however, is costly to acquire and the computational effort of training multi-condition acoustic models is substantial. In this section, we explore the combination of MDT with conventional feature enhancement techniques. Our aim is to reduce the mismatch between reliable features and the clean acoustic model, and to improve mask estimation on real-world recorded speech.

First, as an alternative to multi-condition training, we try to reduce the reliable feature mismatch by applying feature enhancement to the noisy features prior to missing data imputation. In doing so, we keep the estimation procedures for the missing data masks unaltered. For unreliable features, feature enhancement may also be beneficial since the unreliable features are used as an upper bound during imputation. Feature enhancement yields tighter imputation bounds [8]. Here, care must be taken since the uncertainty on the enhanced speech energy increases as the underlying speech signal becomes weaker and the imputation bound may become inaccurate. We therefore opted for spectral subtraction (SS) [18] as a feature enhancement method since the amount of noise suppression is easily controlled.

Secondly, we explore whether mask estimation on real-world data, which was argued to be more difficult than in artificially corrupted databases (cf. section 2.1), can be improved by applying feature enhancement on the speech features used in mask estimation while not preprocessing (NP) the features used for imputation and recognition. The ratio behind this is that on real-world data, the features used for mask estimation mismatch the clean speech features used to train or tune the mask

estimation technique, just like the close-talk channel acoustic models mismatch the observed features in recognition. In our experiments, we used the SS feature enhancement technique with the VQ-based mask estimation technique. The VQ mask was chosen because recreating the codebook was less computationally demanding than retraining the SVM mask estimators used for the SVM mask.

Finally, we combine the two approaches described above, leading to four MDT scenarios: whether SS is applied to the features used in mask estimation (mSS vs. mNP), and whether SS is applied to the features used in recognition (fSS vs fNP). These approaches are summarized below:

- mNPFNP** the VQ-based mask estimation nor the recognizer features are preprocessed with spectral subtraction. This is the VQ-mask result in section 5.1.
- mSSfNP** only the VQ-based mask estimation is preprocessed with spectral subtraction; the recognizer features are not preprocessed.
- mNPfSS** features for mask estimates are not preprocessed; only the recognizer features are preprocessed with spectral subtraction.
- mSSfSS** both the VQ-based mask estimation and the recognizer features are preprocessed with spectral subtraction.
- SS** SS feature enhancement is used without MDT.
- AFE** the ETSI AFE front-end is used without MDT.

The last two configurations serve as a baseline. The SS scenario allows us to evaluate the quality of the feature enhancement method. The advanced front-end feature extraction (AFE) baseline is included in order to compare our approach to what is currently regarded as a very good feature enhancement method (though it cannot be tuned to control the amount of noise suppression, needed for combination with MDT).

7.1 *Experimental setup*

7.1.1 Spectral subtraction

The basic principle of spectral subtraction (SS) is to provide an estimate of clean speech features (feature enhancement) by subtracting a direct estimate of the magnitude spectrum of noise from the noisy speech. In our approach, spectral subtraction was done using the multi-band spectral subtraction approach described in [18]. In summary, spectral subtraction is performed independently in each frequency band. The first 20 non-speech frames (as decided by the VAD in section 4.1.2) are used to provide a noise estimate, which is then assumed constant throughout the utterance. Negative values in the enhanced features are floored to the noisy spectrum using a flooring parameter β set to 0.1 (c.f. [18]). Other parameter settings were the same as in [18].

For experiments with SS, new acoustic models are generated through retraining with aligned MIDA and cepstral feature streams. This retraining procedure consisted of replacing the means and variances of the MIDA acoustic model with their cepstral counterparts.

7.1.2 AFE

The AFE algorithm proposed by ETSI [2] is based on a two-stage Wiener filtering noise reduction. Since the parameters of the two Wiener filters are updated on a frame-by-frame basis the ETSI AFE can deal with dynamically changing noise. After estimating the linear spectrum of each frame, the power spectral density is smoothed along the time axis. A voice activity detector (VAD) determines whether a frame contains speech or background noise; the estimated spectrum of both speech and noise are used to calculate the frequency domain Wiener filter coefficients. Frames labeled as non-speech by the VAD are dropped. The AFE produces cepstral features which are directly used for recognition.

As for the experiments with SS, new acoustic models are generated for AFE through retraining with aligned MIDA and AFE feature streams. When retraining, we do not drop the non-speech frames in order to properly align the MIDA and AFE features. The resulting AFE model is then updated using one pass of Viterbi training. Finally, the AFE model is updated by another pass of Viterbi retraining on AFE features with frame dropping enabled.

7.1.3 VQ mask estimation after SS

The VQ-codebook for the mSSfNP and mSSfSS experiments was learned from the same close-talk channel data used in section 4.1.2, to which SS was applied. The VAD and harmonic decomposition method were applied to the spectral subtracted noisy test utterances. The mask threshold was again optimized over the complete test set and set to $\theta = 8$ dB.

7.2 Results and discussion

7.2.1 MDT versus feature enhancement

Comparing the results of the SS and AFE baseline feature enhancement scores with the MDT recognition scores, it becomes apparent there is a vast difference between the methods. SS, on the one hand, performs worst of all methods, often worse than the uncompensated baseline in Fig. 1 in section 5.1. The AFE, on the other hand, has a performance that is among the best of all methods. The low performance of the SS method is misleading, however, since it was set to be conservative in its noise

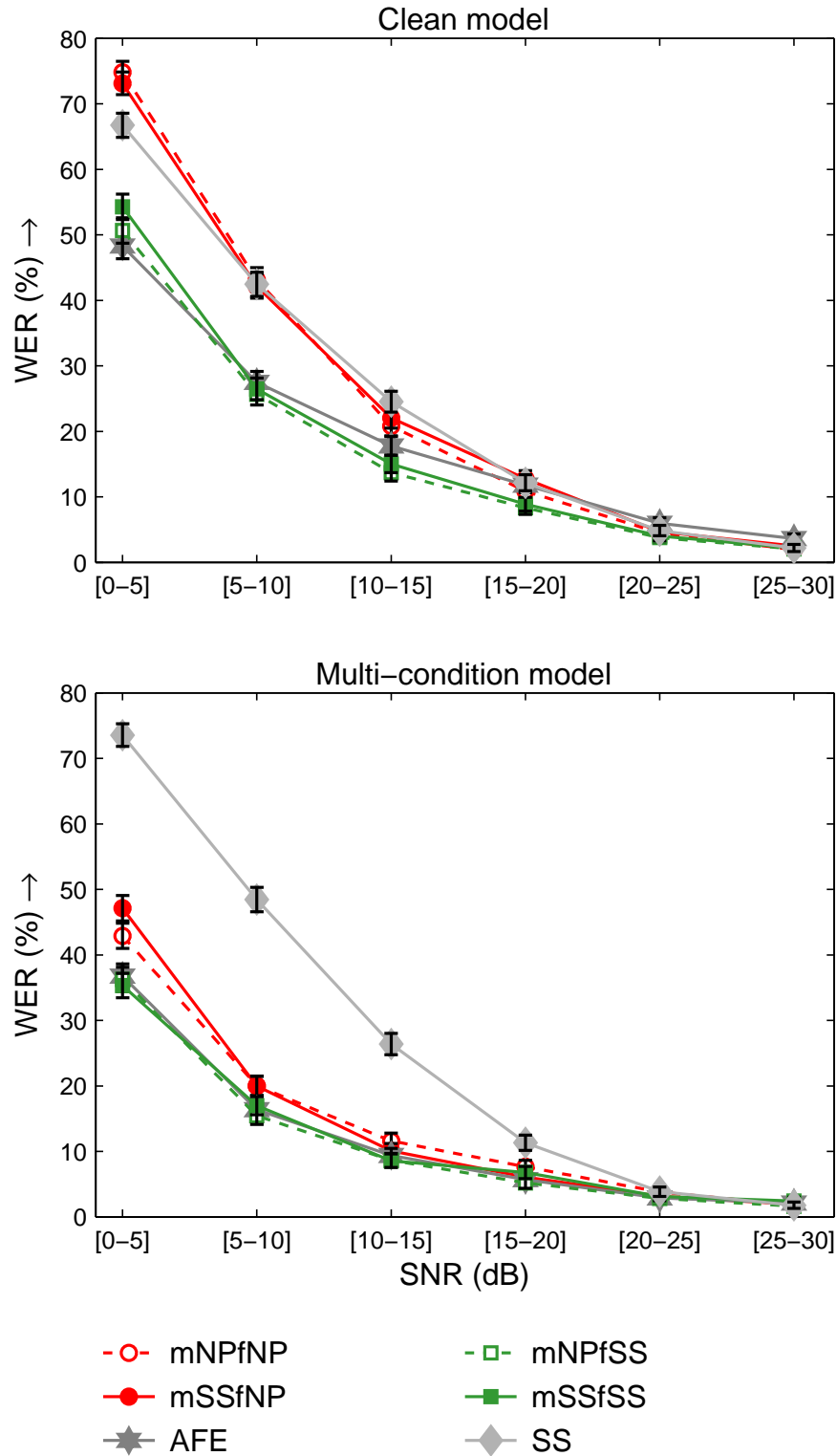


Fig. 4 Word error rate (WER), averaged over the four noise environments, is displayed as a function of SNR for the clean (top) and multi-condition (bottom) acoustic models. Vertical bars around the data points indicate 95% confidence intervals. In each figure we display the results of spectral subtraction (SS), AFE, and the four combination of applying SS to the VQ-mask features (m) or to the features used in recognition (f): mNPfNP, mSSfNP, mNPfSS, mSSfSS, with NP indicating the use of the original noisy features.

suppression. This was necessary to prevent the MDT method, in combination with which it is used, from having upper imputation bounds that are too tight.

The competitive AFE performance underlines that MDT on real-world data is difficult, since in previous work on artificially corrupted data our MDT framework was superior to AFE. While never compared directly, this can be seen for AURORA-4 by comparing the AFE recognition accuracies in [29] with the MDT results in [35], and for AURORA-2 by comparing the AFE scores in [33] with the MDT results in [35].

7.2.2 Spectral subtraction to improve mask estimation

First we compare recognition performance when doing recognition on the original noisy features in combination with the unmodified VQ mask (mNPfNP) and the mask in which SS is applied to the noisy features used in mask estimation (mSSfNP). We observe in Fig. 4 that applying spectral subtraction to improve mask estimation does not lead to significant improvements. When comparing the two MDT approaches in which spectral subtraction is also applied to the features used in recognition, i.e., mNPfSS and mSSfSS, we again observe no significant improvement when applying spectral subtraction to the features used in mask estimation. If anything, there is a trend towards higher WERs, especially at lower SNRs and when using the multi-condition acoustic model.

The reason for this failure to improve mask estimation may stem from the fact SS mostly compensates for stationary noise, something which is already covered to some degree in the VQ-mask estimation because a noise tracker is employed that uses the long-term energy minimum (cf. section 4.1.2). In other words, SS may simply not be powerful enough a technique to reduce a substantial part of the mask estimation errors which are likely due to non-stationary noises. Another possibility is that the harmonic decomposition underlying the VQ-based method fails after speech has been processed by SS because the speech now contains musical noise. Although the VQ-codebook is trained on speech processed with SS, it is conceivable that not all such errors are covered ('learned') by the codebook.

7.2.3 Spectral subtraction to improve MDT recognition

Comparing the results of the MDT approaches in which spectral subtraction is applied to the features used in recognition (mNPfSS and mSSfSS) with those in which the recognizer uses the original noisy speech features (mNPfNP and mSSfNP) in Fig. 4, we observe a substantial decrease in WER when using SS. With a clean acoustic model, the use of SS improves the results significantly at SNRs < 15 dB, with differences as large as 18% (absolute WER) in the 0 – 5 dB SNR range. When using a multi-condition acoustic model, SS improves results at SNRs < 10 dB, with differences up to 9% (absolute WER).

The results show that combining SS and MDT can be beneficial if SS is used to modify the features used in recognition. Moreover, our results show that multi-condition training and SS are complementary, making it advantageous to combine the two approaches. With the multi-condition model the impact of applying SS on recognition performance is smaller (both in absolute and relative terms).

As discussed in section 5.2, a mismatch exists between the reliable features (or false reliables) and the clean acoustic model. It is likely that the improvement in recognition performance is at least partly due to SS reducing this mismatch. The smaller impact of SS when using a multi-condition model could be explained by assuming the multi-condition model already compensates for part of the test/training mismatch. As mentioned above, however, the application of SS also results in tighter imputation bounds when applied to the unreliable features, which may also improve recognition accuracy. Our experiments do not allow us to investigate the relative contribution of these two factors when using SS; further research is needed for that. One possibility to do that would be to only apply SS to either reliable or unreliable features.

The improvements found when applying SS raise the question to what extent recognition with other types of estimated masks, such as the SVM mask described in section 4.1.3, can benefit from feature enhancement. The features used for mask estimation in the SVM mask largely overlap with those used in VQ-based mask estimation. Therefore, given the results in section 7.2.2, it is doubtful whether SVM mask estimation improves after applying SS. Given the success of applying SS to the features used in recognition, however, it seems likely that SVM-based mask estimation will also benefit, and without the additional cost of retraining the SVM and re-estimating the missing data masks.

The AFE front-end, used as a baseline in this chapter, was not used for combination with MDT due to its inflexibility. Its competitive performance, however, merits the question whether a feature enhancement technique based on, or similar to the Wiener filtering used in AFE can be used for combination with MDT. A very similar combination of techniques is described in [12].

8 General discussion and conclusions

In this chapter, we have investigated the performance of a missing data recognizer on speech recorded in real-world environments. We hypothesized that on real-world speech, which is corrupted not only by noise, but also by speaker, reverberation and channel effects, the ‘reliable’ features no longer match an acoustic model trained on clean speech. We investigated the validity of this hypothesis and explored to what extent performance can be improved by combining MDT with three conventional techniques, viz. multi-condition training, de-reverberation and feature enhancement. Using a multi-condition trained acoustic model in combination with MDT, we confirmed the hypothesis and showed that recognition accuracy improves substantially in all noise environments and at all SNR-levels. When comparing the performance

of conventional mask estimation techniques, we found that even a ‘cheating’ semi-oracle missing data mask did not perform better than VQ- or SVM-based estimated masks. We argued this was at least partly due to the semi-oracle missing data mask not being designed to label speech features dominated by reverberation as unreliable.

In a second experiment (cf. section 6), we combined MDT with de-reverberation by doing recognition with the reverberated part of speech labeled ‘unreliable’, both on real-world recordings and on artificially reverberated speech. The experiment with artificially reverberated speech confirmed previous findings, that masking reverberation improves recognition accuracy, but also revealed that the multi-condition trained acoustic model is intrinsically more robust against reverberation. To some degree, these two methods can work together for an even better performance. The experiment on real-world recordings showed that the semi-oracle mask also improved when the reverberant part of speech was labeled unreliable, and thus that with a suitable missing data mask, MDT can compensate for noise and reverberation at once. Finally, the experiments showed that reverberation has a major impact on the recognition performance in far-talk channels.

Third, we did an experiment (cf. section 7) in which we combined MDT with feature enhancement techniques. We investigated whether spectral subtraction (SS) could reduce the mismatch of reliable features to such an extent that it might serve as an alternative to multi-condition training. We also investigated whether spectral subtraction could improve the performance of VQ-based missing data mask estimation, which was found to be unexpectedly poor when using clean acoustic models. The application of spectral subtraction to the features used in VQ mask estimation did not improve results, but the application of SS to the features used in recognition proved to be quite successful: WERs decreased when using a clean as well as a multi-condition model. We argued that this is either due to a reduction of the test/training mismatch in the reliable features, or due to tighter imputation bounds on the unreliable features. Finally, even though in previous work MDT was shown to be superior to the ETSI advanced front-end (AFE) on artificially corrupted speech, we could show only a small advantage of MDT in case multi-condition training is not an option, while MDT performs comparably with the AFE under a multi-condition scenario.

From our findings we conclude that two issues make applying MDT on real-world speech difficult. The first issue is that one of the assumptions underlying MDT, viz. that reliable features remain uncorrupted, can be violated. The second issue is that conventional mask estimation techniques are not able to deal with the fact that real-world speech can be affected not only by environmental noise, but also by effects such as reverberation. In this chapter we showed that the first issue can be dealt with to some degree with conventional noise reduction techniques such as multi-condition training and feature enhancement. With even a ‘cheating’ missing data performing only marginally better than estimated missing data masks, it is clear that in order to deal with the second issue, (much) more effort is needed to improve mask estimation techniques.

Based on our results, however, it is not at all obvious whether MDT can beat a well-designed feature enhancement technique such as the ETSI advanced front-end, which operates at a fraction of the computational cost. Yet, the fact that our work shows that MDT can be combined with conventional noise robustness techniques and since mask estimation allows to integrate additional knowledge sources (e.g. harmonicity), or to use classifiers that do not integrate easily in HMMs (e.g. SVM-based classifiers), means there is still potential for improving the results, for example by combination with the aforementioned feature enhancement technique.

9 Acknowledgements

This research is financed by the MIDAS project of the Nederlandse Taalunie under the STEVIN programme. The research of Maarten Van Segbroeck was financed by the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen). We acknowledge Lou Boves for his help with the manuscript and Toon van Waterschoot for providing the PEM-AFROW code and the room impulse response estimates.

References

1. Spraak: Speech processing, recognition and automatic annotation kit. Website (1996). <http://www.spraak.org/>
2. ETSI standard doc.: Speech processing, transmission and quality aspects (STQ); distributed speech recognition; advanced front-end feature extraction algorithm; ES 202 050 V1.1.5 (2007)
3. C. Cortes, C., Vapnik, V.: Support-vector networks. *Machine Learning* pp. 273 – 297 (1995)
4. Chang, C., Lin, C.: Libsvm: a library for support vector machines (2001)
5. Cooke, M., Green, P., Crawford, M.: Handling missing data in speech recognition. In: In ICSLP-1994, pp. 1555–1558 (1994)
6. Cooke, M., Green, P., Josifovski, L., Vizinho, A.: Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Communication* **34**, 267–285 (2001)
7. Delcroix, M., Nakatani, T., Watanabe, S.: Static and dynamic variance compensation for recognition of reverberant speech with dereverberation preprocessing. *IEEE Transactions on Audio, Speech, and Language Processing* **17**(2), 324–334 (2009)
8. Demange, S., Cerisara, C., Haton, J.P.: Accurate marginalization range for missing data recognition. In: Proc. of Interspeech, pp. 27–31 (2007)
9. Demuynck, K., Duchateau, J., Compernelle, D.V.: Optimal feature sub-space selection based on discriminant analysis. In: Proc. of European Conference on Speech Communication and Technology, vol. 3, pp. 1311–1314 (1999)
10. Duchateau, J., Demuynck, K., Compernelle, D.V.: Fast and accurate acoustic modelling with semicontinuous hmms. *Speech Communication* **24**(1), 5–17 (1998)
11. Duchateau, J., Demuynck, K., Wambacq, D.V.C.P.: Improved parameter tying for efficient acoustic model evaluation in large vocabulary continuous speech recognition. In: Proc. ICSLP, vol. V, pp. 2215–2218. Sydney, Australia (1998)

12. Fernandez Astudillo, R., Kolossa, D., Mandelartz, P., Orglmeister, R.: An uncertainty propagation approach to robust asr using the etsi advanced front-end. accepted for publication in *IEEE Journal of Selected Topics in Signal Processing* (2010)
13. Gemmeke, J.F., Wang, Y., Van Segbroeck, M., Cranen, B., Van hamme, H.: Application of noise robust MDT speech recognition on the SPEECON and SpeechDat-Car databases. In: *Proc. of Interspeech*. Brighton, UK (2009)
14. Hirsch, H., Pearce, D.: The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: *Proc. of ISCA ASR2000 Workshop*, pp. 181–188 (2000)
15. Huang, X., Acero, A., Hon, H.W.: *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall (2001)
16. 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: *Proc. of LREC*, pp. 329–333 (2002)
17. Josifovski, L., Cooke, M., Green, P., Vizinho, A.: State based imputation of missing data for robust speech recognition and speech enhancement. In: *Proc. of Eurospeech* (1999)
18. Kamath, S., Loizou, P.: A multi-band spectral subtraction method for enhancing speech. In: *Proc. of ICASSP*
19. Kim, W., Hansen, J.H.L.: Time-frequency correlation-based missing-feature reconstruction for robust speech recognition in band-restricted conditions. *IEEE Transactions on Audio, Speech, and Language Processing* **17**(7), 1292–1304 (2009)
20. Martin, R.: Noise power spectral density estimation based on optimal smoothing and minimum statistics. *IEEE Transactions on Speech and Audio Processing* **9**, 504–512 (2001)
21. Palomäki, K.J., Brown, G.J., Barker, J.: Techniques for handling convolutional distortion with “missing data” automatic speech recognition. *Speech Communication* **43**, 123–142 (2004)
22. Parihar, N., Picone, J.: An analysis of the aurora large vocabulary evaluation. In: *Proc. of Eurospeech*, pp. 337–340 (2003)
23. Raj, B., Seltzer, M., Stern, R.: Reconstruction of missing features for robust speech recognition. *Speech Communication* **43**, 275–296 (2004)
24. Raj, B., Singh, R., Stern, R.: Inference of missing spectrographic features for robust automatic speech recognition. In: *Proc. of International Conference on Spoken Language Processing*, pp. 1491–1494 (1998)
25. Raj, B., Stern, R.: Missing-feature approaches in speech recognition. *Signal Processing Magazine* **22**(5), 101–116 (2005)
26. Ramírez, J., Górriz, J., Segura, J., Puntonet, C., Rubio, A.: Speech/non-speech discrimination based on contextual information integrated bispectrum lrt. In: *IEEE Signal Processing Letters* (2006)
27. Remes, U., Palomäki, K.J., Kurimo, M.: Missing feature reconstruction and acoustic model adaptation combined for large vocabulary continuous speech recognition. In: *Proc. of EU-SIPCO* (2008)
28. Seltzer, M., Raj, B., Stern, R.: A bayesian classifier for spectrographic mask estimation for missing feature speech recognition. *Speech Communication* **43**, 379–393 (2004)
29. Stouten, V.: *Robust automatic speech recognition in time-varying environments*. Ph.D. thesis, K.U.Leuven (2006)
30. van den Heuvel, H., Boudy, J., Comeyne, R., Communications, M.N.: The speechdat-car multilingual speech databases for in-car applications. In: *Proc. of the European Conference on Speech Communication and Technology*, pp. 2279–2282 (1999)
31. Van hamme, H.: Robust speech recognition using missing feature theory in the cepstral or lda domain. In: *Proc. of European Conference on Speech Communication and Technology*, pp. 3089–3092 (2003)
32. Van hamme, H.: Prospect features and their application to missing data techniques for robust speech recognition. In: *Proc. of Interspeech*, pp. 101–104 (2004)
33. Van hamme, H.: Robust speech recognition using cepstral domain missing data techniques and noisy masks. In: *Proc. of ICASSP*, vol. 1, pp. 213–216 (2004)

34. Van hamme, H.: Handling time-derivative features in a missing data framework for robust automatic speech recognition. In: Proc. of ICASSP (2006)
35. Van Segbroeck, M.: Robust large vocabulary continuous speech recognition using missing data techniques. Ph.D. thesis, K.U.Leuven (2010)
36. Van Segbroeck, M., Van hamme, H.: Handling convolutional noise in missing data automatic speech recognition. In: Proc. of ICASSP (2006)
37. Van Segbroeck, M., Van hamme, H.: Vector-Quantization based mask estimation for missing data automatic speech recognition. In: Proc. of ICSLP, pp. 910–913 (2007)
38. van Waterschoot, T., Rombouts, G., Verhoeve, P., Moonen, M.: Double-talk-robust prediction error identification algorithms for acoustic echo cancellation. *IEEE Transactions on Signal Processing* **55**(3), 846–858 (2007)