# ROBUST SPEECH RECOGNITION USING MODEL-BASED FEATURE ENHANCEMENT AND HARMONIC DECOMPOSITION IN THE MISSING DATA FRAMEWORK

*Maarten Van Segbroeck*, Hugo Van hamme*

{maarten.vansegbroeck,hugo.vanhamme}@esat.kuleuven.be
Katholieke Universiteit Leuven - Dept. ESAT
Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

## ABSTRACT

*Missing Data Theory* (MDT) has shown to improve the robustness of automatic speech recognition (ASR) systems. A crucial part in a MDT-based recogniser is the computation of the reliability masks from noisy data. For each component of the feature vector extracted from the noisy observation, the *Missing Data Detector* (MDD) makes a decision about the presence of speech and noise. The components that are likely to be dominated by the noise, are labelled as unreliable and their values will be estimated from the reliable observations. In this paper, we exploit the *Model-Based Feature Enhancement* (MBFE) technique in the reliability decisions of the MDD. This technique makes use of statistical models of clean speech and noise to generate estimates of the original speech and noise feature vectors. Decision criteria obtained by combining the MBFE-estimates with the components derived from the harmonic decomposition of the noisy speech signal showed a further increase in performance of the recognition system operating on noisy signals.

## 1. INTRODUCTION

Additive noise results in a decrease in performance of ASR-systems due to the mismatch between the speech models (obtained in clean conditions) and the statistics of speech in the noisy test conditions. Missing Data Techniques have shown their effectiveness in reducing this mismatch. In the MDT approach, a MDD decides for each frame which features of the noisy speech are unreliable (missing) or reliable. The latter are treated as clean speech data in the acoustic models of the recogniser's back end. The missing features on the other hand are either removed (marginalisation) or their value is estimated from the reliable data and using the state distribution of the ASR as a prior (data imputation), see [1]. In the present paper, the data imputation method will be applied.
The missing data masks can easily be obtained in the spectral domain by considering the noise-dominated regions in the time-frequency plane. For reasons of accuracy, the ASR-system operates in a domain that is a linear transformation of log-spectra. Therefore, the data imputation technique needs to be extended to cover such linear transformations. In [2] it has been shown that a higher accuracy and robustness can be obtained by using missing data techniques in the cepstral domain. The major drawback of this approach is that it requires significantly more computation since imputing the unreliable data requires the solution of a *Non-Negative Least Square* problem. An alternative MDT formulation was presented in [3] through the introduction of the PROSPECT features, such that the computational load is reduced while maintaining the accuracy. The missing data approach in this domain will be briefly restated in section 2.

In order to make decisions about the reliability of the data, information about the clean speech and noise must be derived from the noisy signal. In section 3, we introduce a MDD that computes the noisy masks by exploiting the MBFE-technique of [4]. In this technique, we incorporate a priori knowledge of the speech and the background noise by using a Hidden Markov Model (HMM) with Gaussian observation probabilities for the clean speech cepstral feature vectors and another Gaussian HMM for the noise cepstral feature vectors. Based on these statistical models, the parameters of a combined HMM of the noisy speech can be obtained. This combined model is then used to calculate the global MMSE-estimate of the clean speech and the noise, given the noisy speech. If we integrate those estimates in a local *signal-to-noise ratio* (SNR) criterion, reliability decisions on the data can be made.

Section 4 describes the MDD presented in [5] which employs the decomposition of the speech signal into a harmonic and a random part. In section 5, this harmonic decomposition is combined with MBFE to build missing data masks. Firstly, we integrate the MBFE-estimates in the harmonicity MDD of [5]. In a second approach, missing data decision are made by the linear combination of the MBFE-estimates and the harmonic decomposed components. The experimental results are presented in section 6 and section 7 concludes.

*This work was supported by "Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen)".

## 2. MISSING DATA TECHNIQUES

The speech recogniser is assumed to have a mainstream HMM-based architecture with Gaussian mixture models (GMM). In the front-end, a low resolution MEL-spectral representation is computed by a filter bank with $D$ channels through windowing, framing, FFT and filter bank integration. At frame $t$, the output of the filter bank with centre frequency $f$ will be denoted by $|Y_t(f)|$, $|S_t(f)|$ and $|N_t(f)|$ for the noisy speech, clean speech and noise respectively. The log-MEL spectral noisy features $\mathbf{y}_t$ are then obtained by stacking $log(|Y_t(f)|)$ for all filter banks in a vector, and likewise for $\mathbf{s}_t$ and $\mathbf{n}_t$.
In missing data theory, the following assumption is made for the noisy speech:

$$\mathbf{y}_t \approx max(\mathbf{s}_t, \mathbf{n}_t), \tag{1}$$

where the max-operator works element-wise over the MEL-spectral components. The ideal MDD decision criterion labels the components of the noisy speech vector $\mathbf{y}_t$ as missing if:

$$|S_t(f)| < |N_t(f)|. \tag{2}$$

If we further remark that regions with a low local SNR will be dominated by the noise, we can rewrite (2) as the local SNR-criterion:

$$10log_{10}\frac{|S_t(f)|^2}{|N_t(f)|^2} < \delta. \tag{3}$$

with some threshold $\delta$. To illustrate, Figure 1 shows the ideal mask of a mixture of speech and babble noise.
In this way, the noisy data vector $\boldsymbol{y}_t$ is partitioned into a reliable and an unreliable part, $(\boldsymbol{y}_{t,r}, \boldsymbol{y}_{t,u})$. The reliable part of $\boldsymbol{s}_t$ is estimated as $\boldsymbol{y}_{t,r}$. In the maximum likelihood per Gaussian-based imputation, the missing part of $\boldsymbol{s}_t$ is estimated by minimizing the (negative) log-likelihood for each Gaussian mixture component over $\mathbf{s}_t$:

$$\frac{1}{2}(\boldsymbol{s}_t - \boldsymbol{\mu})'\boldsymbol{P}(\boldsymbol{s}_t - \boldsymbol{\mu}) \tag{4}$$

subject to the equality and inequality constraints:

$$\boldsymbol{s}_{t,r} = \boldsymbol{y}_{t,r} \text{ and } \boldsymbol{s}_{t,u} \leq \boldsymbol{y}_{t,u} \tag{5}$$

where $\boldsymbol{\mu}$ is the Gaussian mean and $\boldsymbol{P}$ is an inverse covariance or precision matrix. $\boldsymbol{\mu}$ and $\boldsymbol{P}$ are both estimated on clean training data.
In most MDT systems, GMMs with diagonal covariance in the log-spectral domain are used, resulting in a diagonal structure for $\boldsymbol{P}$ and a tractable MLE for $\boldsymbol{s}$. Higher accuracies are obtained with GMMs with a diagonal covariance in the cepstral domain, in which case $\boldsymbol{P}$ becomes non-diagonal. Imputation then becomes computationally more complex since the imputation of the unreliable part now requires the solution of a *Non-Negative Least Square* problem. In this paper, the PROSPECT features defined in [3] will be used. Just like cepstra, they are computed by a linear transform of the logarithm of the filter bank energies. While these features can be applied in any speech recognition system, they show especially a clear benefit in MDT-based recognition since they reduce the computational requirements over cepstral MDT while the accuracy is maintained.

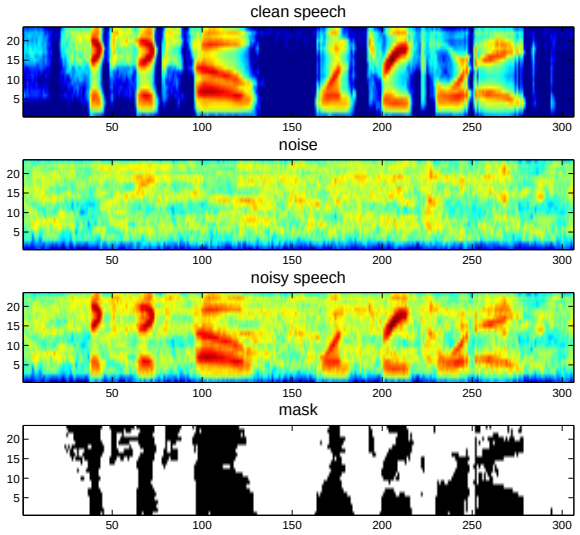


***Figure 1:*** *Row 1-3 depict log-MEL spectrograms of the clean speech, noise and noisy mixture of the TIDigit sequence 'six six oh one three four eight' in 10dB babble noise. The bottom row shows the missing data mask. The regions in black have been assigned to the speech by the ideal local SNR criterium.*

## 3. MISSING DATA DETECTION WITH MBFE-ESTIMATES

In practise, the values of $|S_t(f)|$ and $|N_t(f)|$ are not available in (3) and they must be estimated from $\mathbf{y}_t$. Speech and noise estimates can be obtained by embedding the Model-Based Feature Enhancement technique in the MDD. In [4] it has been shown that this technique was intrinsically suited for the estimation of the clean speech. We can easily extend the MBFE-algorithm to obtain the noise estimates as well. Since the MBFE-algorithm works in the cepstral domain (with dimension $K$), we denote the clean speech, noisy speech and noise cepstral feature vector of frame $t$ with $\mathbf{c}_t^s$, $\mathbf{c}_t^n$ and $\mathbf{c}_t^y$ respectively. As illustrated in figure 2, both $\mathbf{s}_t$ and $\mathbf{n}_t$ are modelled by a fully connected HMM. Since a mixture of Gaussians can be decomposed into an ergodic single Gaussian HMM, we will use single Gaussian state-conditional probability density functions

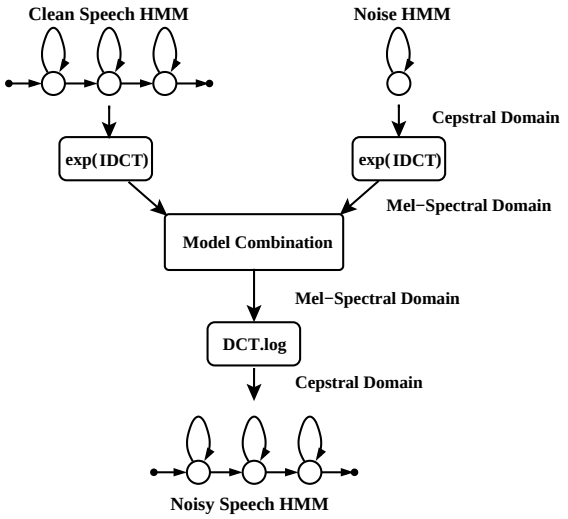


***Figure 2:*** *HMM combination principle.*

(pdfs) for simplicity of notation. Let $q_t^s \in \{1, \ldots, M^s\}$ be the speech state and $q_t^n \in \{1, \ldots, M^n\}$ the noise state at time $t$, then the state-conditional pdfs of clean speech and noise are :

$$p(\mathbf{c}_t^s | q_t^s = i) = N(\mathbf{c}_t^s; \boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{s}}, \boldsymbol{\Sigma}_i^s), \tag{6}$$

$$p(\mathbf{c}_t^n | q_t^n = j) = N(\mathbf{c}_t^n; \boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{n}}, \boldsymbol{\Sigma}_j^n), \tag{7}$$

where the covariance matrices $\boldsymbol{\Sigma}_i^s$ and $\boldsymbol{\Sigma}_i^n$ are diagonal. These statistical models are then combined to a fully connected HMM which models the noisy speech. For each combined state $(i, j)$ the emission density is given by:

$$p(\mathbf{c}_t^y | q_t^s = i, q_t^n = j) = N(\mathbf{c}_t^y; \boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{y}}, \boldsymbol{\Sigma}_{(i,j)}^y), \tag{8}$$

where $\boldsymbol{\mu}_{\boldsymbol{i}}^{\boldsymbol{y}}$ and $\boldsymbol{\Sigma}_{(i,j)}^y$ are estimated by a first order Vector Taylor Series approximation in the spectral domain as mentioned in [4]. This yields:

$$\boldsymbol{\mu}_{(i,j)}^y \approx \frac{1}{2} C \, log\big(\exp(2C^+ \boldsymbol{\mu}_i^s) + \exp(2C^+ \boldsymbol{\mu}_j^n)\big), \tag{9}$$

$$\boldsymbol{\Sigma}_{(i,j)}^y \approx F_{(i,j)} \boldsymbol{\Sigma}_i^s F_{(i,j)}' + G_{(i,j)} \boldsymbol{\Sigma}_j^n G_{(i,j)}', \tag{10}$$

where $\boldsymbol{\Sigma}_{(i,j)}^y$ is no longer diagonal and with $C^+$ the Moore-Penrose generalised inverse of the non-square DCT-matrix $C$ and

$$F_{(i,j)} = C \, diag\left(\frac{1}{1 + exp\left[2C^+(\boldsymbol{\mu}_j^n - \boldsymbol{\mu}_i^s)\right]}\right) C^+, \tag{11}$$

$$G_{(i,j)} = I - F_{(i,j)}. \tag{12}$$

For each noisy speech state $(i, j)$ of this combined HMM, we can derive the state-conditional MMSE-estimates $\hat{\mathbf{c}}_{t,(i,j)}^s$ of the clean speech and $\hat{\mathbf{c}}_{t,(i,j)}^n$ of the noise, given the noisy speech at time $t$, as follows:

$$\hat{\mathbf{c}}_{t,(i,j)}^s = \boldsymbol{\mu}_i^s + \boldsymbol{\Sigma}_i^s F_{(i,j)}' \left(\boldsymbol{\Sigma}_{(i,j)}^y\right)^{-1} (\mathbf{c}_t^y - \boldsymbol{\mu}_{(i,j)}^y), \tag{13}$$

$$\hat{\mathbf{c}}_{t,(i,j)}^n = \boldsymbol{\mu}_j^n + \boldsymbol{\Sigma}_j^n G_{(i,j)}' \left(\boldsymbol{\Sigma}_{(i,j)}^y\right)^{-1} (\mathbf{c}_t^y - \boldsymbol{\mu}_{(i,j)}^y). \tag{14}$$

The global MMSE-estimates of the clean speech and noise are then obtained as a linear combination of the state-conditional MMSE-estimates weighted by the a posteriori probabilities $\gamma_t^{(i,j)}$ (see [4]), respectively:

$$\hat{\mathbf{c}}_t^s = \sum_i \sum_j \gamma_t^{(i,j)} \hat{\mathbf{c}}_{t,(i,j)}^s, \tag{15}$$

$$\hat{\mathbf{c}}_t^n = \sum_i \sum_j \gamma_t^{(i,j)} \hat{\mathbf{c}}_{t,(i,j)}^n. \tag{16}$$

Finally, by transforming (15) and (16) to the spectral domain, we find the components $|\hat{S}_t^{MBFE}(f)|$ and $|\hat{N}_t^{MBFE}(f)|$ of the MBFE-estimated spectral feature vector of the clean speech and noise. These can be filled in (3) to obtain the missing data criterion used by our MBFE-based MDD. This approach differs from [6] in that we use cepstral models, Vector Taylor Series, a GMM for the noise and imputation in the back end.

## 4. MISSING DATA DETECTION AND HARMONIC DECOMPOSITION

In [5] an algorithm was described to decompose a signal into a harmonic and a random spectral component. A pitch estimate is first computed by a subharmonic summation method. The signal is then subsequently framed in overlapping segments with a length of two pitch periods and a single period of frame-shift. If $j$ is the pitch period index en $\Omega_j$ the estimate of the double pitch period, then the noisy speech signal is written as:

$$y_j(n) = h_j(n) + r_j(n) \;\text{ with } 0 \leq n < \Omega_j \tag{17}$$

where $r_j(n)$ is the random component and the harmonic component is given by:

$$h_j(n) = \left(1 + \frac{e_j n}{\Omega_j}\right) \cdot \left[\sum_{k=0}^{K_j} a_{k,j} cos(2\pi f_{0,j} k n) + \sum_{k=1}^{K_j} b_{k,j} sin(2\pi f_{0,j} k n)\right] \tag{18}$$

where $f_{0,j}$ is the pitch estimate for segment $t$ and the number of harmonics $K_j$ is the largest integer such as $f_{0,j} K_j < 0.5$. The change in amplitude over the $\Omega_j$ samples, is taken into account by the linear modulation factor $(1 + e_j n / \Omega_j)$.

For each segment $t_p$, the parameters $a_{k,j}$ and $b_{k,j}$ together with $e_j$ and $f_{0,j}$ are estimated in the least square sense by minimizing $(\mathbf{y}_j - \mathbf{h}_j)'(\mathbf{y}_j - \mathbf{h}_j)$ and are plugged into eq. (18). More details can be found in [5]. The random part is the difference between the noisy signal and the the harmonic signal. The MEL-filter bank outputs computed on $h_j$ and $r_j$ will be denoted by $|H_t(f)|$ en $|R_t(f)|$ respectively.
The proposed MDD in [5] firstly estimates the clean speech spectrum at frame $t$ as:

$$|S_t(f)|^2 = \gamma_t(f)|H_t(f)|^2, \tag{19}$$

with $\gamma_t(f)$ a frequency dependent gain. The missing data criterion (2) can then be rewritten as

$$(2\gamma_t(f) - 1)|H_t(f)|^2 < |R_t(f)|^2. \tag{20}$$

As mentioned in [5], when an analysis window of twice the pitch period is used, the harmonic and random part of a signal have interleaved spectral lines. Hence, the harmonic part of the noise $|H_t^{(N)}(f)|$ can be obtained by MEL-integrating the spectral lines of the noise spectrum at multiples of the pitch and the MEL-integration of the spectral lines between pitch multiples results in the random part $|R_t^{(N)}(f)|$. An estimate for the gain factor $\gamma$ can then be derived from:

$$(2\gamma_t(f) - 1) = 1 - 2\frac{|H_t^{(N)}(f)|^2}{|H_t(f)|^2}. \tag{21}$$

If we assume that the noise has a smooth spectrum (its magnitude does not change much as the frequency changes with half the pitch), $|H_t^{(N)}(f)|$ can be approximated by $|R_t^{(N)}(f)|$. The latter can then be estimated from the minimum of $|R(f)|$ over a window of 2L+1 frames (here L = 10):

$$|R_t^{(N)}(f)| \approx \beta \min_{t-L \le s \le t+L} |R_s(f)| \;\; with\ \beta \ge 1. \tag{22}$$

## 5. MISSING DATA DETECTION BY COMBINING HARMONICITY WITH MBFE

### 5.1. Gain factor estimation

In this section, two alternative expressions for the gain factor are derived. Firstly, if the MBFE estimation of the speech is used in (19), then

$$\gamma_t(f) = \rho_1 \frac{|\hat{S}_t^{MBFE}(f)|^2}{|H_t(f)|^2}. \tag{23}$$

The parameter $\rho_1$ compensates for the approximation involved in (23). Secondly, consider the decomposition of the noise into a harmonic and a random part:

$$|N_t(f)|^2 = |H_t^{(N)}(f)|^2 + |R_t^{(N)}(f)|^2. \tag{24}$$

After replacing $|N_t(f)|^2$ by his MBFE-estimate and since $|H_t^{(N)}(f)|$
$\approx |R_t^{(N)}(f)|$, we have:

$$|H_t^{(N)}(f)|^2 \approx \frac{1}{2}|\hat{N}_t^{MBFE}(f)|^2. \tag{25}$$

Hence, substitution of (25) into (21), gives:

$$(2\gamma_t(f) - 1) = 1 - \rho_2 \frac{|\hat{N}_t^{MBFE}(f)|^2}{|H_t(f)|^2}. \tag{26}$$

Where $\rho_2$ compensates for the approximation in (25). The parameters $\rho_1$ and $\rho_2$ are both experimentally derived numbers, fixed for all noises and signal-to-noise ratios (here $\rho_1$ = 0.75 and $\rho_2$ = 0.85).

### 5.2. Weighted decision criterion

We now introduce a novel decision criterion, which is a linear combination of the decisions $log(|\hat{S}_t^{MBFE}(f)|) < log(|\hat{N}_t^{MBFE}(f)|)$ and $log(|H_t(f)|) < log(|R_t(f)|)$, namely

$$x_1 + a.x_2 + b < 0 \tag{27}$$

where the inequality-operator works component-wise and with

$$x_1 = log(|H_t(f)|) - log(|R_t(f)|) \tag{28}$$

and

$$x_2 = log(|\hat{S}_t^{MBFE}(f)|) - log(|\hat{N}_t^{MBFE}(f)|) \tag{29}$$

Criterion (27) can also be seen as the equation of a line through the point $(0, -b)$ with slope $-a$. Thus, a component corresponding to MEL-frequency $f$ is classified as unreliable if the point $(x_1, x_2)$ lies below this line. The parameter $a$ is found by a Fisher discriminant analysis [7]. $b$ is determined by minimisation of the classification error of the combined mask (27) with the oracle mask (3) as a reference. This gives $a = 0.81$ and $b = 0.1$ for all noises and signal-to-noise ratios.

## 6. EXPERIMENTAL RESULTS

The performance of the proposed MDT-approaches are evaluated on two test sets. The first test set is obtained by the artificially adding of 4 types of military noises (Lynx helicopter, machine gun, Destroyer tank and F16 noise) from the NOISEX-92 database to the TI-Digits speech database of Aurora-2. The second one is the Aurora-2 test set A. The recogniser is based on the HTK software package version 2.2 from Entropic. The digits are modelled as whole left-to-right word-HMMs with 16 states per digit and 20 Gaussians with diagonal covariance per state. The static features are modelled in the PROSPECT

| NOISEX-92, clean training, multiconditional testing | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MBFE | | | | | MDT MBFE | | | | | MDT harmonicity | | | | |
| | Lynx | M.Gun | Destr. | F16 | Avg. | Lynx | M.Gun | Destr. | F16 | Avg. | Lynx | M.Gun | Destr. | F16 | Avg. |
| 20 dB | 99.23 | 99.43 | 99.11 | 99.57 | 99.34 | 99.14 | 99.30 | 99.05 | 99.26 | 99.19 | 99.08 | 98.43 | 98.93 | 99.11 | 98.89 |
| 15 dB | 99.02 | 99.30 | 98.84 | 98.89 | 99.01 | 98.96 | 99.03 | 98.69 | 98.80 | 98.87 | 98.50 | 96.61 | 98.60 | 98.52 | 98.06 |
| 10 dB | 96.71 | 98.49 | 95.20 | 96.42 | 96.71 | 97.33 | 98.67 | 95.59 | 97.16 | 97.19 | 96.27 | 93.08 | 96.09 | 96.58 | 95.51 |
| 5 dB | 89.56 | 96.77 | 85.33 | 88.52 | 90.05 | 89.93 | 97.19 | 86.52 | 90.50 | 91.04 | 88.49 | 88.91 | 87.24 | 89.79 | 88.61 |
| 0 dB | 72.28 | 93.44 | 67.04 | 69.48 | 75.56 | 73.75 | 94.38 | 66.87 | 71.12 | 76.53 | 69.05 | 80.65 | 69.40 | 70.01 | 72.28 |
| Avg. | 91.36 | 97.49 | 89.10 | 90.58 | **92.13** | 91.82 | 97.71 | 89.34 | 91.37 | **92.56** | 90.28 | 91.54 | 90.05 | 90.80 | **90.67** |
| | MDT combined eq. (23) | | | | | MDT combined eq. (26) | | | | | MDT combined eq. (27) | | | | |
| | Lynx | M.Gun | Destr. | F16 | Avg. | Lynx | M.Gun | Destr. | F16 | Avg. | Lynx | M.Gun | Destr. | F16 | Avg. |
| 20 dB | 99.23 | 99.30 | 99.14 | 99.07 | 99.19 | 99.26 | 99.06 | 98.99 | 99.38 | 99.17 | 99.45 | 99.21 | 99.15 | 99.44 | 99.31 |
| 15 dB | 98.53 | 99.12 | 98.39 | 98.52 | 98.64 | 98.71 | 99.00 | 98.69 | 98.92 | 98.83 | 98.86 | 98.88 | 98.84 | 98.83 | 98.85 |
| 10 dB | 97.24 | 98.67 | 95.62 | 96.48 | 97.00 | 97.33 | 98.07 | 95.94 | 97.38 | 97.18 | 97.27 | 98.67 | 96.15 | 97.53 | 97.41 |
| 5 dB | 90.08 | 96.92 | 87.03 | 89.73 | 90.94 | 90.82 | 96.77 | 88.55 | 92.47 | 92.15 | 91.83 | 97.22 | 88.28 | 92.59 | 92.48 |
| 0 dB | 73.20 | 93.68 | 68.68 | 69.58 | 76.29 | 73.72 | 93.68 | 70.38 | 76.00 | 78.45 | 77.68 | 95.04 | 71.34 | 77.35 | 80.35 |
| Avg. | 91.66 | 97.54 | 89.77 | 90.68 | **92.41** | 91.97 | 97.32 | 90.51 | 92.83 | **93.16** | 93.02 | 97.80 | 90.75 | 93.15 | **93.68** |

***Table 1:*** *Recognition accuracy for the NOISEX-92 noise types artificially added to the TIDigits speech of Aurora-2 with 3 Gaussians in the MBFE noise model.*

domain and the velocity and acceleration features in the cepstral domain (with $D = 23$ and $K = 13$). Therefore, MDT will only be applied to the static features while the dynamic cepstra are derived from the MBFE-estimate of the clean speech. The MBFE-algorithm uses a speech model with $M^s = 128$ fully connected Gaussians trained on the clean speech and a noise model trained on the noise present in the NOISEX-92 and Aurora-2 set A database. Because of the presence of highly non-stationary noise types in NOISEX-92 (machine gun), we used $M^n = 3$ here and $M^n = 1$ elsewhere.

Tables 1 and 2 list the word error rate (substitutions, deletion, insertions) of the test sets using clean speech models in the recogniser's back end. The tables present the recognition accuracy for the MBFE-estimates of the clean speech (MBFE), for the MDD of [5] (MDT harmonicity) and for the MBFE-based MDD of section 3 (MDT MBFE). The tables show also how the performance and noise robustness improve when MBFE is combined with harmonicity in the MDT-based recogniser.

An interesting remark is the bad performance of the harmonicity based MDD on noise types consisting of short impulses, such as the machine gun noise. The pulses of the shootings are too short compared to the window length of 2L+1 frames in eq. (22). Hence, the pulses do not much affect the values of $\gamma$ and therefore they will be classified as speech resulting in many insertion errors. As can be seen in table 1, this problem is solved by the combination of MBFE and harmonicity.

## 7. DISCUSSION, CONCLUSIONS AND FUTURE WORK.

Model-Based Feature Enhancement has proven to be a powerful technique for speech enhancement that exploits model knowledge of both speech and noise. Providing the enhanced speech directly to the recogniser ("MBFE" in the tables) provides only slightly inferior performance as using MBFE-derived masks ("MDT MBFE"). Although all information about the noise is summarised in a binary mask in the latter approach, the MDT-based back end succeeds equally well as the MBFE in reconstructing the corrupted speech regions using its speech model. On the other hand, this result is somehow discouraging for the missing data approach since it cannot exploit the stronger back end speech model to produce a significant advantage. This is in contrast to what has been observed with spectral subtraction [8], a speech and noise estimation method with inferior performance. We have shown, however, that the harmonicity of voiced speech can be integrated to produce better masks and hence additional characteristics of speech can be exploited successfully.

In this research, we have used a HMM model for noise. While scenarios can be envisaged where such a model can be estimated off-line or on-line, one of the main advantages of missing data techniques is that less assumptions could be made about the noise. While we have succeeded in strengthening the speech model through integration of harmonicity and MBFE masks, our next step is to weaken the assumptions about the noise.

| Aurora2, clean training, multiconditional testing | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MBFE | | | | | MDT MBFE | | | | | MDT harmonicity | | | | |
| | Subw. | Babble | Car | Exh. | Avg. | Subw. | Babble | Car | Exh. | Avg. | Subw. | Babble | Car | Exh. | Avg. |
| 20 dB | 98.83 | 98.94 | 99.25 | 98.89 | 98.98 | 98.74 | 98.28 | 98.99 | 98.49 | 98.63 | 99.11 | 98.82 | 99.19 | 98.61 | 98.93 |
| 15 dB | 97.67 | 97.85 | 98.48 | 97.16 | 97.79 | 97.70 | 96.80 | 98.33 | 97.04 | 97.47 | 98.22 | 97.13 | 98.30 | 97.93 | 97.90 |
| 10 dB | 94.38 | 93.62 | 94.51 | 93.00 | 93.88 | 95.39 | 93.14 | 94.75 | 93.71 | 94.25 | 95.64 | 94.38 | 95.56 | 95.31 | 95.22 |
| 5 dB | 85.82 | 78.75 | 80.85 | 81.33 | 81.69 | 88.30 | 79.08 | 82.88 | 82.75 | 83.25 | 87.44 | 83.40 | 86.16 | 84.82 | 85.46 |
| 0 dB | 63.89 | 45.19 | 50.73 | 59.67 | 54.87 | 66.60 | 45.28 | 51.92 | 60.20 | 56.00 | 63.03 | 53.90 | 55.35 | 54.46 | 56.69 |
| Avg. | 88.12 | 82.87 | 84.76 | 86.01 | **85.44** | 89.35 | 82.52 | 85.37 | 86.44 | **85.92** | 88.69 | 85.53 | 86.91 | 86.23 | **86.84** |
| | MDT combined eq. (23) | | | | | MDT combined eq. (26) | | | | | MDT combined eq. (27) | | | | |
| | Subw. | Babble | Car | Exh. | Avg. | Subw. | Babble | Car | Exh. | Avg. | Subw. | Babble | Car | Exh. | Avg. |
| 20 dB | 98.96 | 98.79 | 98.99 | 98.67 | 98.85 | 98.89 | 97.94 | 98.96 | 98.70 | 98.62 | 98.89 | 98.00 | 98.90 | 98.77 | 98.64 |
| 15 dB | 98.04 | 97.25 | 98.48 | 96.48 | 97.56 | 97.61 | 96.31 | 98.21 | 97.19 | 97.33 | 98.07 | 96.49 | 98.18 | 97.28 | 97.51 |
| 10 dB | 95.73 | 93.86 | 94.48 | 93.37 | 94.36 | 95.33 | 92.65 | 95.76 | 94.11 | 94.46 | 96.50 | 93.14 | 95.53 | 94.51 | 94.92 |
| 5 dB | 88.24 | 78.87 | 83.24 | 82.63 | 83.25 | 86.92 | 79.66 | 87.89 | 83.83 | 84.58 | 90.02 | 80.56 | 87.41 | 85.56 | 85.89 |
| 0 dB | 64.94 | 44.01 | 50.76 | 56.31 | 54.01 | 67.64 | 52.78 | 62.21 | 60.14 | 60.69 | 72.34 | 51.69 | 63.79 | 64.73 | 63.14 |
| Avg. | 89.18 | 82.56 | 85.19 | 85.49 | **85.61** | 89.28 | 83.87 | 88.61 | 86.79 | **87.14** | 91.16 | 83.98 | 88.76 | 88.17 | **88.02** |

***Table 2:*** *Recognition accuracy for test set A of Aurora-2 with a single Gaussian MBFE noise model.*

## 8. REFERENCES


[1] M. Cooke, Ph. Green, L. Josifovski, and A. Vicinho, "Robust automatic speech recognition with missing and unreliable acoustic data," in *Speech Communication*, 2001, vol. 34, pp. 267–285.

[2] H. Van hamme, "Robust Speech Recognition Using Missing Feature Theory in the Cepstral or LDA Domain," in *Proc. European Conference on Speech Communication and Technology*, Geneva, Switzerland, 2003, pp. 3089–3092.

[3] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. International Conference on Spoken Language Processing*, Jeju Island, Korea, October 2004, pp. 101–104.

[4] V. Stouten, H. Van hamme, K. Demuynck, and P. Wambacq, "Robust speech recognition using model-based feature enhancement," in *Proc. European Conference on Speech Communication and Technology*, Geneve, Switzerland, September 2003, vol. I, pp. 17–20.

[5] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. International Conference on Acoustic, Speech and Signal Processing*, Montreal, Canada, May 2004, pp. 213–216.

[6] P. Renevey and A. Drygajlo, "Estimation of unreliable data for robust speech recognition," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 2000.

[7] CR. Rao, "Linear statistical inference and its applications," 1973, John Wiley & Sons.

[8] A. Vizinho, P. Green, M. Cooke, and L. Josifovsk, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: An integrated study," in *Proc. Eurospeech99*, Budapest, September 1999, vol. 5, pp. 2407–2410.