COMBINING MISSING-DATA RECONSTRUCTION AND UNCERTAINTY DECODING FOR ROBUST SPEECH RECOGNITION

*José A. González, Antonio M. Peinado, Angel M. Gómez**

Dpt. de Teoría de la Señal, Telemática y Comunicaciones
University of Granada, Spain
{joseangl,amp,amgg}@ugr.es

Ning Ma, Jon Barker

Dpt. of Computer Science
University of Sheffield, UK
{n.ma,j.barker}@dcs.shef.ac.uk

ABSTRACT

This paper proposes a novel approach for noise-robust speech recognition which combines a missing-data (MD) derived spectral reconstruction technique and uncertainty decoding based on the weighted Viterbi algorithm (WVA). First, the noisy feature vectors are compensated by using a novel MD imputation technique based on the integration of truncated Gaussian pdfs. Although the proposed MD estimator has both the advantages of MD techniques and the use of cepstral features, it may still be affected by a number of uncertainty sources. In order to deal with these uncertainties, WVA-based uncertainty decoding is proposed. Our experiments on the Aurora-2 and Aurora-4 tasks show that the proposed MD estimator outperforms other MD imputation techniques. Also, we show that the combination of MD imputation with WVA provides better results than the combination with other uncertainty processing techniques such as the use of evidence pdfs for the estimated features.

Index Terms— Missing data imputation, uncertainty decoding, MMSE estimation, speech recognition

1. INTRODUCTION

Feature compensation (or feature enhancement) techniques for robust speech recognition deal with speech distorted by different kinds of noise sources (e.g. additive noise, channel distortion). The aim of these techniques is to remove as much noise as possible while keeping speech intelligible. Over the last few years, one approach that has proved to be very effective in this task is the missing data framework [1, 2]. This framework considers that, when speech is distorted by noise, some parts of speech spectra will be more affected than others. In this sense, the parts where the energy of speech dominates can be considered *reliable*, whereas those regions dominated by the energy of the noise are *unreliable*.

Two different approaches have been considered in the missing data framework to perform speech recognition with incomplete data: *marginalization* and *imputation* [1]. In the *marginalization* approach, speech decoding relies on the reliable parts of spectra, while the unreliable parts are discarded or marginalized up to the observed values. The *imputation* approach makes use of the redundancy in speech signals to estimate the missing data and, then, speech recognition is performed as usual. As can be noted, the strength of both approaches is that few or no assumptions about the corrupting noise are made (these assumptions are embedded in the reliable/unreliable classification process).

*This work has been supported by an FPU grant from the Spanish Ministry of Education and by projects MICINN TEC2010-18009, CEI BioTIC GENIL (CEB09-0010) and UK EPSRC grant EP/G039046/1 .

While marginalization is known to perform optimal classification with missing data [2], it suffers from several drawbacks. First, recognition has to be carried out with spectral features. It has been shown that recognition using cepstral features outperforms that using spectral features [2]. Second, no usual feature compensation techniques such as cepstral mean normalization (CMN) can be applied. Third, the standard decoding algorithm must be modified to account for the missing values in the marginalization approach. Finally, since spectral features are correlated with each other, the acoustic model (hidden Markov models) needs to employ Gaussian mixtures with with full covariance matrices or an increased number of Gaussians with diagonal covariance. This could be computationally prohibitive in some cases, e.g. in large vocabulary continuous speech recognition systems [1].

This paper proposes a novel imputation technique that produces full-band reconstructed spectra, which can be later transformed to the cepstral domain. In order to apply this spectral reconstruction, a Gaussian mixture model (GMM) is trained on clean speech. In this way, the correlation between features is explicitly modeled and used for reconstruction. Moreover, the use of truncated Gaussian distributions allows the exploitation of the boundary information provided by the masking noise.

It has been shown [1, 2] that imputation techniques tend to be very sensitive to errors in the identification of missing data. These errors result in poor spectral reconstruction that degrades recognition performance. Thus, we also propose here a joint scheme where the reliability of the reconstruction is first estimated and, then, this information is propagated to the decoder as a weighting factor, so that more reliably reconstructed frames are weighted more.

This paper is organized as follows. In Section 2, the mathematical formulation of the missing data reconstruction is derived. Section 3 describes the proposed joint scheme for uncertainty computation and its exploitation in the decoder. The experimental framework and the results are presented in Section 4. Finally, Section 5 concludes this paper and discusses future work.

2. SPECTRAL RECONSTRUCTION USING TRUNCATED GAUSSIAN DISTRIBUTIONS

Let $\boldsymbol{x}$, $\boldsymbol{n}$, and $\boldsymbol{y}$ be the log-filterbank feature vector representations (e.g. log-Mel) corresponding to frames of clean speech, additive noise, and noisy speech, respectively. Because of the logarithmic compression applied to the filterbank outputs, the following approximation can be made,

$$\boldsymbol{y} \approx \log(e^{\boldsymbol{x}} + e^{\boldsymbol{n}}) \approx \max(\boldsymbol{x}, \boldsymbol{n}) \tag{1}$$

According to (1), $\boldsymbol{y}$ can be rearranged into $\boldsymbol{y} \equiv (\boldsymbol{y}_r, \boldsymbol{y}_u)$, where

$\boldsymbol{y}_r \approx \boldsymbol{x}_r$ are the reliable features (those where the speech energy dominates, $\boldsymbol{x}_r > \boldsymbol{n}_r$), and $\boldsymbol{y}_u$ ($-\infty \leq \boldsymbol{x}_u \leq \boldsymbol{y}_u$) are the unreliable features (the noise energy dominates in those features, $\boldsymbol{x}_u < \boldsymbol{n}_u$)[1].

In order to compensate for the effects of noise, a minimum mean square error (MMSE) estimator for $\boldsymbol{x}_u$ exploiting the known correlations with $\boldsymbol{x}_r$ can be derived. The general form of this estimator is given by,

$$\hat{\boldsymbol{x}}_u = E\left[\boldsymbol{x}_u | \boldsymbol{x}_u \leq \boldsymbol{y}_u, \boldsymbol{x}_r\right] = \int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u p\left(\boldsymbol{x}_u | \boldsymbol{x}_r, \boldsymbol{y}_u\right) d\boldsymbol{x}_u \quad (2)$$

If we now make the usual assumption that clean speech can be well modeled by a GMM, the resulting MMSE estimator is the well-known cluster-based reconstruction proposed in [2]:

$$\hat{\boldsymbol{x}}_u = \sum_k P\left(k | \boldsymbol{x}_r, \boldsymbol{y}_u\right) \hat{\boldsymbol{x}}_u^k \quad (3)$$

where $P\left(k|\boldsymbol{x}_r, \boldsymbol{y}_u\right)$ is the posterior probability for the kth Gaussian in the GMM and $\hat{\boldsymbol{x}}_u^k$ is the partial estimate for $\boldsymbol{x}_u$ given this Gaussian.

As can be observed in (3), the MMSE estimator reduces to the computation of the probabilities and the expected values in the sum. First, we analyze the problem of computing $P\left(k|\boldsymbol{x}_r, \boldsymbol{y}_u\right)$. Applying Bayes' rule, we obtain:

$$P\left(k|\boldsymbol{x}_r, \boldsymbol{y}_u\right) = \frac{p\left(\boldsymbol{x}_r, \boldsymbol{y}_u|k\right) P(k)}{\sum_{k'} p\left(\boldsymbol{x}_r, \boldsymbol{y}_u|k'\right) P(k')} \quad (4)$$

with $P(k)$ being the prior probability for the kth Gaussian. The probability $p\left(\boldsymbol{x}_r, \boldsymbol{y}_u|k\right)$ can be computed by marginalizing over the unreliable features $\boldsymbol{x}_u$ up to the observed ones $\boldsymbol{y}_u$:

$$\begin{aligned} p\left(\boldsymbol{x}_r, \boldsymbol{y}_u|k\right) &= \int_{-\infty}^{\boldsymbol{y}_u} p\left(\boldsymbol{x}_r, \boldsymbol{x}_u|k\right) d\boldsymbol{x}_u \\ &= p(\boldsymbol{x}_r|k) \int_{-\infty}^{\boldsymbol{y}_u} p(\boldsymbol{x}_u|\boldsymbol{x}_r, k) d\boldsymbol{x}_u \end{aligned} \quad (5)$$

Both probability density functions (pdfs) $p(\boldsymbol{x}_r|k)$ and $p(\boldsymbol{x}_u|\boldsymbol{x}_r, k)$ can be shown to be Gaussian distributed: the parameters of the marginal distribution $p(\boldsymbol{x}_r|k) = \mathcal{N}(\boldsymbol{x}_r; \boldsymbol{\mu}_r^k, \boldsymbol{\Sigma}_{rr}^k)$ are obtained from the original pdf $\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}^k, \boldsymbol{\Sigma}^k)$ by partitioning $\boldsymbol{\mu}^k$ and $\boldsymbol{\Sigma}^k$ into their reliable and unreliable features as,

$$\boldsymbol{\mu}^k = \begin{pmatrix} \boldsymbol{\mu}_r^k \\ \boldsymbol{\mu}_u^k \end{pmatrix} \quad (6)$$

$$\boldsymbol{\Sigma}^k = \begin{pmatrix} \boldsymbol{\Sigma}_{rr}^k & \boldsymbol{\Sigma}_{ru}^k \\ \boldsymbol{\Sigma}_{ur}^k & \boldsymbol{\Sigma}_{uu}^k \end{pmatrix} \quad (7)$$

The mean and covariance of the conditional pdf $p(\boldsymbol{x}_u|\boldsymbol{x}_r, k) = \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}_{u|r}^k, \boldsymbol{\Sigma}_{u|r}^k)$ are

$$\boldsymbol{\mu}_{u|r}^k = \boldsymbol{\mu}_u^k + \boldsymbol{\Sigma}_{ur}^k \left(\boldsymbol{\Sigma}_{rr}^k\right)^{-1} \left(\boldsymbol{y}_r - \boldsymbol{\mu}_r^k\right) \quad (8)$$

$$\boldsymbol{\Sigma}_{u|r}^k = \boldsymbol{\Sigma}_{uu}^k - \boldsymbol{\Sigma}_{ur}^k \left(\boldsymbol{\Sigma}_{rr}^k\right)^{-1} \boldsymbol{\Sigma}_{ru}^k \quad (9)$$

The main problem of computing the integral of (5) is that no closed-form solution exists for Gaussian pdfs with non-diagonal covariance. Thus, we must resort to some approximations to make the integral tractable. In [2], Raj et al. considered diagonal covariance matrices $\boldsymbol{\Sigma}^k$ in (7) for every Gaussian in the GMM. With this approximation, the integral is now computable, but the correlation between features is only captured via the Gaussian component variable k. Faubel et al. [3] proposed an alternative approach in which a linear transformation is applied to diagonalize the covariance matrix $\boldsymbol{\Sigma}_{u|r}^k$ in (9). After the transformation, the multivariate integral reduces to univariate integrals in the transformed domain, so that it can be computed. However, the integration limits are also modified by the transformation being non-aligned with the new axes. Nevertheless, this is not considered in Faubel's approach.

In this work, we will assume that the covariance $\boldsymbol{\Sigma}_{u|r}^k$ in (9) is diagonal. This way, the integral of (5) can be computed, whereas the correlation between reliable features is exploited in the form of $p(\boldsymbol{x}_r|k)$. Applying this approximation, $p\left(\boldsymbol{x}_r, \boldsymbol{y}_u|k\right)$ can be computed as,

$$\begin{aligned} p\left(\boldsymbol{x}_r, \boldsymbol{y}_u|k\right) &\approx p(\boldsymbol{x}_r|k) \prod_i \int_{-\infty}^{y_{u,i}} p(x_u|\boldsymbol{x}_r, k) dx_u \\ &= \mathcal{N}\left(\boldsymbol{x}_r; \boldsymbol{\mu}_r^k, \boldsymbol{\Sigma}_r^k\right) \prod_i \Phi\left(z_{u,i}^k\right) \end{aligned} \quad (10)$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function and $z_{u,i}^k$ is the standardized value for $y_{u,i}$ regarding the kth Gaussian:

$$z_{u,i}^k = \frac{y_{u,i} - \mu_{u|r,i}^k}{\sigma_{u|r,i}^k} \quad (11)$$

Let us now consider the computation of $\hat{\boldsymbol{x}}_u^k$ in (3). This term corresponds to the following expected value:

$$\begin{aligned} \hat{\boldsymbol{x}}_u^k &= E\left[\boldsymbol{x}_u | \boldsymbol{x}_u \leq \boldsymbol{y}_u, \boldsymbol{x}_r, k\right] = \int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u \cdot p\left(\boldsymbol{x}_u | \boldsymbol{x}_r, \boldsymbol{y}_u, k\right) d\boldsymbol{x}_u \\ &\approx \int_{-\infty}^{\boldsymbol{y}_u} \boldsymbol{x}_u \cdot p\left(\boldsymbol{x}_u | \boldsymbol{x}_r, k\right) d\boldsymbol{x}_u \end{aligned} \quad (12)$$

where $p\left(\boldsymbol{x}_u|\boldsymbol{x}_r, k\right)$ is the same pdf as in (5). Again, the integral of (12) has a closed-form only for Gaussian pdfs with diagonal covariance matrices. Thus, applying the diagonal approximation, we can compute this expected value, which corresponds to the mean of a truncated Gaussian distribution defined in the interval $(-\infty, y_{u,i}]$ for each unreliable feature i [4]:

$$\hat{x}_{u,i}^k = \mu_{u|r,i}^k - \sigma_{u|r,i}^k \frac{\mathcal{N}(z_{u,i}^k)}{\Phi(z_{u,i}^k)} \quad (13)$$

3. INCORPORATING RECONSTRUCTION UNCERTAINTY INTO DECODING

The reconstruction performed by the proposed imputation technique cannot be considered as completely reliable. Factors such as the SNR of input signals, the degree of noise stationarity and the accuracy of mask estimation will affect the performance of the proposed estimation in (3). For these reasons, we first derive an estimator for the expected accuracy of the reconstruction based on the variance of the estimation. A modified decoding algorithm based on the weighted Viterbi algorithm (WVA) [5, 6] is then employed to accommodate the estimation uncertainty.

A common way to estimate the accuracy of an estimator is by means of its variance. Assuming no uncertainty during mask estimation, the variance for frames with many unreliable features will be high, whereas for totally reliable frames it should be zero. The

[1] We assume that a missing data mask is already available, so that unreliable/reliable regions can be identified.

		Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Avg.	R.I.%
	Baseline	99.11	94.74	84.67	62.35	33.45	14.17	8.05	56.65	-
Oracle experiments	WVA	99.13	96.75	94.11	85.62	66.96	42.49	22.39	72.49	27.96
	Imputation	99.11	99.01	98.75	97.99	96.11	90.90	77.34	94.17	66.23
	Imputation+WVA	99.13	98.93	98.78	98.26	97.02	93.10	82.71	95.42	68.44
Real experiments	Imputation	99.10	96.95	94.19	87.56	74.33	48.69	20.40	74.46	31.44
	Imputation+WVA	99.17	97.36	95.05	89.11	77.03	51.09	20.98	75.68	33.59

Table 1. Word accuracy results (%) for Aurora-2 database at different SNRs.

covariance matrix associated to the MMSE estimation in (2) is defined as the following expected value:

$$\Sigma_{\hat{\boldsymbol{x}}_u} = E\left[(\boldsymbol{x}_u - \hat{\boldsymbol{x}}_u)(\boldsymbol{x}_u - \hat{\boldsymbol{x}}_u)^T | \boldsymbol{x}_u \leq \boldsymbol{y}_u, \boldsymbol{x}_r\right] \quad (14)$$

Assuming again a GMM for clean speech, we obtain the covariance for the estimation in (3) as,

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{x}}_u} = \sum_k P\left(k|\boldsymbol{x}_r, \boldsymbol{y}_u\right)\left(\tilde{\boldsymbol{\Sigma}}^k_{u|r} + (\hat{\boldsymbol{x}}^k_u - \hat{\boldsymbol{x}}_u)(\hat{\boldsymbol{x}}^k_u - \hat{\boldsymbol{x}}_u)^T\right) \quad (15)$$

with $\tilde{\boldsymbol{\Sigma}}^k_{u|r}$ being the covariance associated with the partial estimate $\hat{\boldsymbol{x}}^k_u$ computed in (13). In order to compute this matrix, we consider the case of a truncated Gaussian distribution defined by the conditional pdf $p(\boldsymbol{x}_u|\boldsymbol{x}_r, k) = \mathcal{N}(\boldsymbol{x}_u; \boldsymbol{\mu}^k_{u|r}, \boldsymbol{\Sigma}^k_{u|r})$ with upper bounds $\boldsymbol{y}_u$. For this case, the variance is given by [4],

$$\tilde{\sigma}^{k,2}_{u|r,i} = \sigma^{k,2}_{u|r,i}\left(1 - \frac{\mathcal{N}(z^k_{u,i})}{\Phi(z^k_{u,i})}\left(z^k_{u,i} + \frac{\mathcal{N}(z^k_{u,i})}{\Phi(z^k_{u,i})}\right)\right) \quad (16)$$

where we have again assumed independence between the unreliable features given the reliable ones.

Once estimated, the uncertainty of the missing-data reconstruction can be employed by the speech recognizer. There have been several attempts to exploit the uncertainty during speech recognition, most of them being based on the uncertainty decoding approach [7, 8, 5]. In this approach, a Gaussian evidence pdf is considered for every estimate. The uncertainty of the estimation is taken into account during recognition by adding the variance of the estimate to the model variances. The work such as [9, 10] has explored this approach under the missing-data framework. The problem in this case lies in how the estimated variances are transformed from the log-spectral domain, where the imputation is performed, to the cepstral domain. In [9] regressions trees are trained to learn the nonlinear transformation between both domains. Alternatively, a linear transformation matrix is trained using stereo-data in [10].

This paper considers an uncertainty decoding scheme based on WVA. Instead of propagating the variance of the estimation to the decoder, a time-varying weighting factor $\gamma_t \in [0, 1]$ (one per frame) is used, so that the contribution of unreliable feature vectors is diminished. Hence, the state metrics updating equation used in the decoding stage is,

$$\phi_t(s_j) = \max_{s_i}\left\{\phi_{t-1}(s_i)a_{ij}\right\} p(\boldsymbol{x}_t|s_j)^{\gamma_t} \quad (17)$$

where s_i and s_j are states of the acoustic model, a_{ij} and $p(\boldsymbol{x}_t|s_j)$ correspond to the transition and observation probabilities, and $\phi_t(s_j)$ is the likelihood of the best decoding path for the state s_j at time t.

As can be observed, WVA reduces to a simple multiplication in the logarithm domain, thus being very efficient in comparison with other uncertainty decoding approaches, e.g. variance propagation. Furthermore, we have shown in previous work [5, 11] that WVA can outperform these approaches.

In order to compute the weight γ_t in (17), we first define a uncertainty function for the estimation carried out by (3). In this work, we propose a uncertainty function based on the mean square error (MSE) for the estimate. In this way, the uncertainty of a frame will be proportional to the MSE computed for the unreliable features of this frame. The MSE for $\hat{\boldsymbol{x}}_u$ can be computed as the trace of the covariance matrix for the estimate:

$$\epsilon = \mathrm{tr}(\Sigma_{\hat{\boldsymbol{x}}_u}) \quad (18)$$

Finally, the weighting factor γ used by WVA is generated by applying a sigmoid compression to ϵ:

$$\gamma = 1 - \frac{1}{1 + e^{-\alpha(\epsilon - \beta)}} \quad (19)$$

where α, β are the sigmoid slope and centre, respectively. These parameters are empirically derived.

4. EXPERIMENTS AND RESULTS

The proposed techniques have been evaluated on Aurora-2 [12] and Aurora-4 [13] databases using acoustic models trained on clean speech. For the connected digit Aurora-2 task, left to right continuous density HMMs with 16 states and 3 Gaussians per state are used to model each digit. In the case of the large vocabulary Aurora-4 task, continuous cross-word triphone models with 3 tied states and a mixture of 6 Gaussians per state are used. The language model is the standard bigram for the WSJ0 task.

Speech features are extracted according with the European Telecommunication Standards Institute front-end (ETSI FE, ES 201 108) [14]. The final feature vector employed by the recognizer consists of 12 Mel-Frequency Cepstral Coefficients (MFCCs) and 0th order cepstral coefficient along with their delta and delta-delta coefficients. For spectral reconstruction, 23-component feature vectors corresponding to the outputs of the log-Mel filterbank are used. After reconstruction, the discrete cosine transform (DCT) is applied to obtain the final cepstral parameters.

Spectral reconstruction is performed using a 256 component GMM with full covariance matrices. Training is carried out on the same clean dataset as for acoustic model training. Two types of binary missing-data masks are employed to evaluate the proposed imputation technique: oracle and real masks. Oracle masks are obtained by direct comparison between clean and noisy spectra, thus allowing us to evaluate the potential of the proposed techniques. More realistic masks are obtained by estimating the noise spectrum, which is computed by averaging a given number of frames extracted from the beginning and end of every noisy utterance. In both cases (oracle and real masks), a SNR threshold is applied to obtain binary

		T-01	T-02	T-03	T-04	T-05	T-06	T-07	T-08	T-09	T-10	T-11	T-12	T-13	T-14	Avg.	R.I.%
	Baseline	87.26	38.11	34.17	54.96	39.34	34.15	31.31	62.92	29.33	25.74	40.31	28.88	23.61	22.17	39.45	-
Oracle	WVA	87.26	56.19	50.76	73.30	52.23	47.49	46.29	68.67	42.48	40.48	55.80	38.91	31.09	33.40	51.74	31.16
	Imputation	87.26	85.48	84.53	86.31	84.10	83.71	83.00	76.42	73.68	73.64	76.26	72.65	70.60	70.99	79.19	100.74
	Imputation+WVA	87.26	85.78	84.46	86.33	84.46	84.16	83.41	77.56	75.83	76.18	77.38	73.47	72.73	73.16	80.16	103.20
Real	Imputation	87.00	55.86	58.47	80.93	52.98	59.07	61.70	73.01	46.65	50.05	67.79	45.45	48.09	53.37	60.03	52.18
	Imputation+WVA	87.41	59.99	62.53	82.01	55.97	61.22	64.28	72.56	47.75	51.19	68.17	46.33	49.67	52.27	61.53	55.97

Table 2. Aurora-4 word accuracy results (%) for the different test sets.

missing data masks. As suggested in [1], a feature is considered reliable if its local SNR is greater than 3 dB. Finally, the parameters α and β of (19) are experimentally determined for each database using a small development set extracted from the multicondition training set.

Table 1 shows the word accuracy results (WAcc) for the Aurora-2 database. The three test sets of Aurora-2 are considered for computing the average results per SNR. In addition, the overall average (Avg.) and the relative improvement (R.I.) regarding the baseline for every technique are also shown. Baseline results are obtained applying acoustic models trained with clean speech and no compensation.

The oracle experiments use oracle missing-data masks and/or oracle uncertainties. Oracle uncertainties are obtained as the squared error between the utterance to be recognized and its corresponding clean one, both expressed in the cepstral domain. After the computation of this error, the sigmoid compression of (19) is applied. On the other hand, real masks and uncertainties derived from the estimator MSE are used in the real experiments.

As can be seen in Table 1, decoding with noisy utterances and oracle uncertainties (WVA) achieves an improvement of 27.96 %. This shows the potential benefits of this technique. Significant relative improvements of 66.23 % and 31.44 % over the baseline are achieved by the proposed imputation technique when oracle and real masks, respectively, are used. For comparative purposes, the relative improvements obtained by Raj's reconstruction [2] are 63.48 % with oracle masks and 29.48 % with real masks. The proposed ensemble approach Imputation+WVA produces the best recognition results. In this case, the improvement is especially noticeable at medium and low SNRs.

Table 2 shows the WAcc results (%) obtained for the different test sets in the Aurora-4 database: sets T-01 to T-07 are distorted by additive noise and sets T-08 to T-14 are distorted by additive and convolutive noise (T-01 and T-08 correspond to clean speech). Here the improvement achieved by the proposed approach is bigger than in Aurora-2, given the higher task complexity of Aurora-4. Among all the noise conditions, only the real mask results for set T-04, which corresponds to the quasi stationary car noise, are comparable to those obtained using oracle masks. This indicates that a better mask estimation technique is needed for this database, rather than the simple noise estimation using the first and last frames of the utterance. By comparing Imputation+WVA with the approach proposed Srinivasan et al. in [9] (i.e., imputation plus variance propagation), we see that our proposal outperforms the variance-based one. Under oracle conditions, the average WAcc (%) for sets T-02 to T-07[2] achieved by our proposal is 84.77 %, whereas Srinivasan's approach yields a performance of 79.42 %.

5. CONCLUSION AND FUTURE WORK

This paper has presented a novel noise-robust approach to automatic speech recognition by combining feature enhancement and uncertainty exploitation. A spectral reconstruction technique based on the missing data framework is proposed to estimate those spectral regions corrupted by noise. To do so, the information provided by the *reliable* regions and a joint statistical distribution modeling the correlation between features are used. As the reconstruction provided by the proposed technique cannot be considered fully reliable, a modified decoding algorithm based on the weighted Viterbi algorithm is also proposed, in which less reliable estimates are weighted less by the decoder. The experimental results show the effectiveness of this approach in both small and large vocabulary recognition tasks.

In this work, a weighting factor γ_t per frame is employed by the WVA to perform decoding. In this sense, a weighting factor per feature is expected to provide better performance. Other future work includes the extension of the proposed approach to allow the use of soft masks.

[2]Srinivasan's approach is tested only in these test sets in [9].

6. REFERENCES

[1] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable data", *Speech Comm.*, vol. 34, no. 3, pp. 267–285, June 2001.

[2] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition", *Speech Comm.*, vol. 48, no. 4, pp. 275–296, 2004.

[3] F. Faubel, J. McDonough, and D. Klakow, "Bounded conditional mean imputation with Gaussian mixture models: A reconstruction approach to partly occluded features", *Proc. ICASSP*, pp. 3869–3872, 2009.

[4] P. J. Dhrymes, "Moments of truncated (normal) distributions", 2005.

[5] J. A. Gónzalez, A. M. Peinado, A. M. Gómez, and J. L. Carmona, "Efficient MMSE estimation and uncertainty processing for multienvironment robust speech recognition", *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1206–1220, July 2011.

[6] N. B. Yoma, F. R. McInnes, and M. A. Jack, "Weighted Viterbi algorithm and state duration modelling for speech recognition in noise ", *Proc. ICASSP*, pp. 709–712, 1998.

[7] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion", *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 412–421, May 2005.

[8] H. Liao and M. J. F. Gales, "Issues with uncertainty decoding for noise robust automatic speech recognition", *Speech Comm.*, vol. 50, no. 4, pp. 265–277, 2008.

[9] S. Srinivasan and D. Wang, "Transforming binary uncertainties for robust speech recognition", *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 7, pp. 2130–2140, Sep. 2007.

[10] J. F. Gemmeke, U. Remes, and K. J. Palomäki, "Observation uncertainty measures for sparse imputation", *Proc. Interspeech*, pp. 2262–2265, 2010.

[11] J. L. Carmona, A. M. Peinado, J. L. Perez-Cordoba, and A. M. Gómez, "MMSE-based packet loss concealment for CELP-coded speech recognition", *IEEE Trans. Audio Speech Lang. Process.*, vol. 18, no. 6, pp. 1341–1353, Aug. 2010.

[12] H. G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluations of the speech recognition systems under noisy conditions", in *ISCA ITRW ASR2000*, Paris, France, 2000.

[13] H. G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends of large vocabulary task", Tech. Rep., STQ AURORA DSR Working Group, 2002.

[14] "ETSI ES 201 108 - Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms", ETSI.