ugr Universidad de Granada

Combining missing-data reconstruction and uncertainty decoding for robust speech recognition

J.A. González[1], A.M. Peinado[1], A.M. Gómez[1], N. Ma[2], and J. Barker[2]

Dpt. Signal Theory, Telematics and Communications, University of Granada (Spain)

Dpt. Computer Science, The University of Sheffield (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.

Introduction

- **Noise robustness** is a crucial component of the speech-enabled applications for mobile devices.
- **Methods for noise robust speech recognition:**
  - Feature-domain techniques.
  - Model-domain techniques.
  - *Uncertainty based approaches.*
- **Our proposal:** uncertainty based approach combining feature compensation and uncertainty exploitation:
  - First, the missing features in the log-spectral domain are MMSE estimated .
  - Then, the uncertainty of the estimation is exploited by the recognizer.

Missing-data theory

Model for log-spectral feature distortion due to additive noise:

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

Information about the feature reliability is provided by a *missing-data mask*.

According to this mask, the noisy feature vector can be rearranged into $\boldsymbol{y} = (\boldsymbol{y}_r, \boldsymbol{y}_u)$:

- **Reliable features** $(\boldsymbol{y}_r \approx \boldsymbol{x}_r)$, i.e. unaffected features.
- **Unreliable features** $(-\infty \le \boldsymbol{x}_u \le \boldsymbol{y}_u)$**:** speech is masked by noise.

Missing-data imputation

Aim: To estimate (impute) the unreliable features by taking advantage of the *speech redundancy* and the upper bound information provided by the observation.

The **MMSE estimate of the unreliable features** is given by,

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

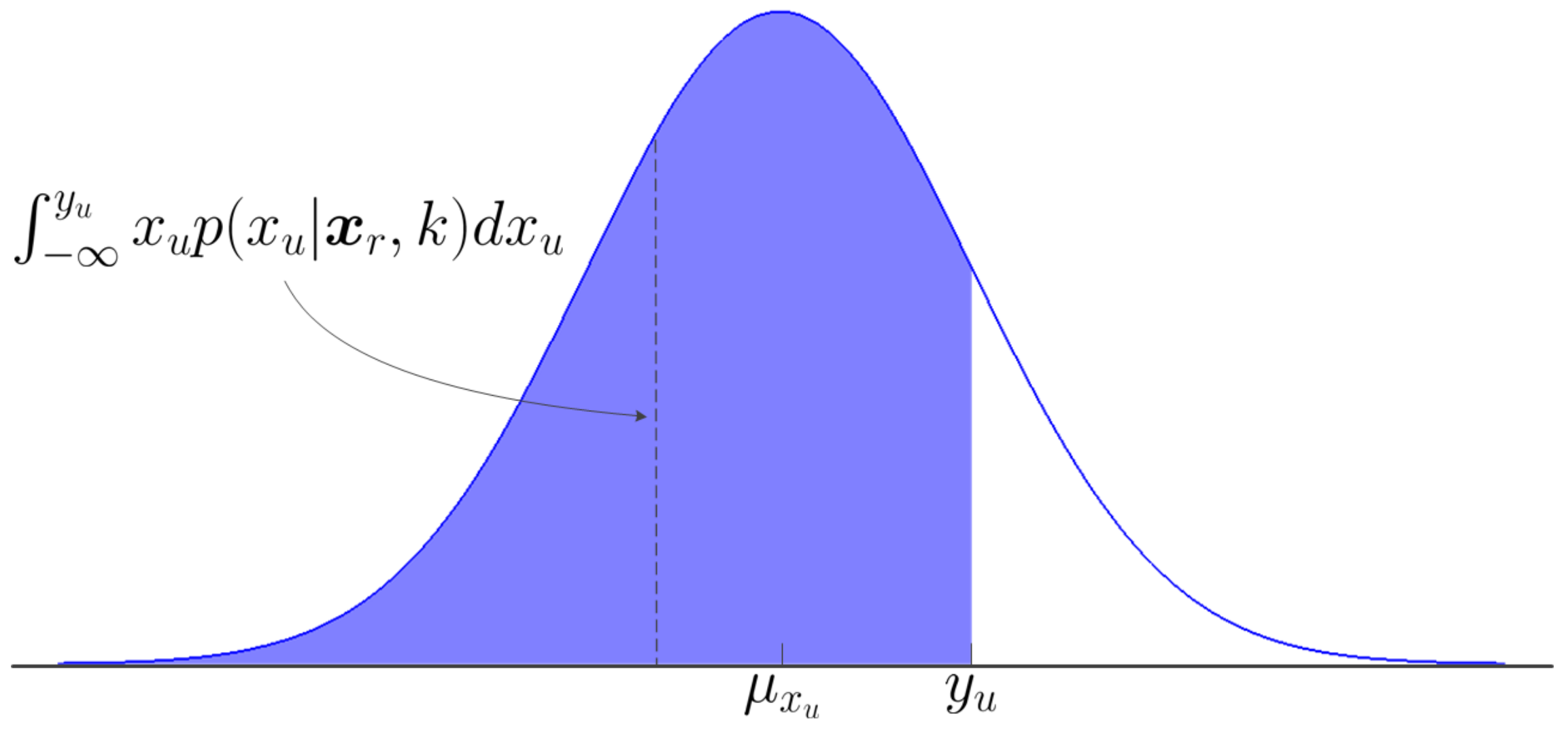
Using a *Gaussian mixture model (GMM)* to represent the clean speech features, the MMSE estimator can be rewritten as,

$$\hat{\boldsymbol{x}}_u = \sum_k P\left(k | \boldsymbol{x}_r, \boldsymbol{y}_u\right) E\left[\boldsymbol{x}_u | \boldsymbol{x}_u \le \boldsymbol{y}_u, \boldsymbol{x}_r, k\right]$$

In order to obtain $P(k|\boldsymbol{x}_r, \boldsymbol{y}_u)$, the following probability must be computed:

$$p\left(\boldsymbol{x}_r, \boldsymbol{y}_u | k\right) = p\left(\boldsymbol{x}_r | k\right) \int_{-\infty}^{\boldsymbol{y}_u} p\left(\boldsymbol{x}_u | \boldsymbol{x}_r, k\right) d\boldsymbol{x}_u$$

$E[\boldsymbol{x}_u | \boldsymbol{x}_u \le \boldsymbol{y}_u, \boldsymbol{x}_r, k]$ corresponds to the mean of a right-truncated Gaussian distribution:



Exploiting the estimation uncertainty

The proposed reconstruction cannot be considered as completely reliable. A **weighted Viterbi algorithm** performs decoding by taking into account the estimation uncertainty:

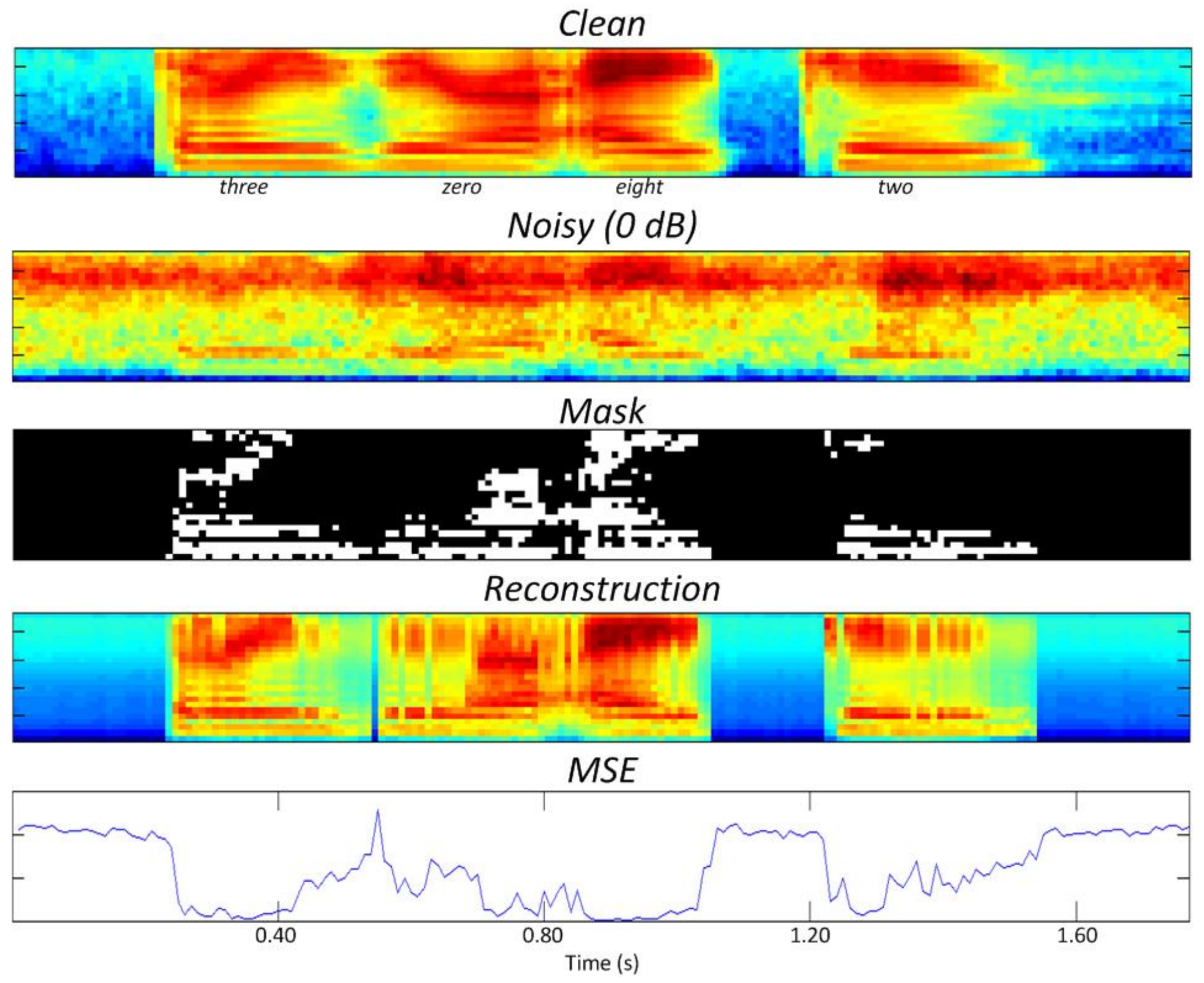
$$\phi_t(s_j) = \max_{s_i} \left\{\phi_{t-1}(s_i) a_{ij}\right\} p(\boldsymbol{x}_t | s_j)^{\gamma_t}$$

The exponential weight γ_t is computed by applying a sigmoid compression to the *mean square error (MSE)* of the estimate. The MSE is defined as the trace of:

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

with $\tilde{\boldsymbol{\Sigma}}_{u|r}^k$ being the diagonal covariance matrix of the truncated distribution.

Illustrative example



Performance Evaluation

- **Task: Aurora-2** (English connected digits) and **Aurora-4** (large vocabulary) databases.
- **Speech feature extractor**: ETSI Front-end.
- Oracle and real (estimated) masks are employed.
- GMMs with full covariance and 256 components are used.

Word accuracy results

Aurora-2 results

		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	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
	Imp.+WVA	99.13	98.93	98.78	98.26	97.02	93.10	82.71	95.42	68.44
Real	Imputation	99.10	96.95	94.19	87.56	74.33	48.69	20.40	74.46	31.44
	Imp.+WVA	99.17	97.36	95.05	89.11	77.03	51.09	20.98	**75.68**	**33.59**

Aurora-4 results

		T-01	T-02	T-03	T-04	T-05	T-06	T-07	Avg.	R.I. %
	Baseline	87.26	38.11	34.17	54.96	39.34	34.15	31.31	**45.61**	-
Oracle	WVA	87.26	56.19	50.76	73.30	52.23	47.49	46.29	59.07	29.51
	Imputation	87.26	85.48	84.53	86.31	84.10	83.71	83.00	84.91	86.15
	Imp.+WVA	87.26	85.78	84.46	86.33	84.46	84.16	83.41	85.12	86.61
Real	Imputation	87.00	55.86	58.47	80.93	52.98	59.07	61.70	65.14	42.82
	Imp.+WVA	87.41	59.99	62.53	82.01	55.97	61.22	64.28	**67.63**	**48.26**

- **Oracle:** oracle masks and/or oracle uncertainties.
- **Real:** estimated masks and/or estimated uncertainties
- A similar technique (Srinivasan and Wang, 2007) yields a performance of 79.42 % for sets T-02 to T-07 using oracle masks.

Conclusions

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.

Future work:

- Binary masks vs. soft masks .
- Uncertainty per frame vs. uncertainty per feature.