Observation uncertainty measures for sparse imputation

Jort F. Gemmeke¹, Ulpu Remes², Kalle J. Palomäki²

¹Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands
²Adaptive Informatics Research Centre, Aalto University, Finland

j.gemmeke@let.ru.nl
{ulpu.remes, kalle.palomaki}@tkk.fi

Abstract

Missing data imputation estimates the clean speech features for automatic speech recognition in noisy environments. The estimates are usually considered equally reliable while in reality, the estimation accuracy varies from feature to feature. In this work, we propose uncertainty measures to characterise the expected accuracy of a sparse imputation (SI) based missing data method. In experiments on noisy large vocabulary speech data, using observation uncertainties derived from the proposed measures improved the speech recognition performance on features estimated with SI. Relative error reductions up to 15% compared to the baseline system using SI without uncertainties were achieved with the best measures.

Index Terms: compressive sensing, noise robustness, speech recognition, uncertainty decoding

1. Introduction

Missing data techniques (MDT) [1] improve automatic speech recognition (ASR) in noisy environments by finding reliable information in the noisy corrupted speech signal. Decoding is either based on the reliable information alone or the unreliable i.e. missing information is reconstructed using missing data imputation prior to decoding. Sparse imputation (SI) [2] is an exemplar-based reconstruction method which is based on representing segments of the noisy speech signal as linear combinations of as few as possible clean speech example segments referred to as exemplars. SI uses more time-context than traditional frame-based MDT and was shown to outperform a conventional imputation technique [3]. Especially in low-SNR conditions, where the reliable information is scarce, SI showed an impressive performance.

As with any feature enhancement or missing data imputation method, the estimated clean speech features calculated using SI are not completely accurate. Since the accuracy depends on several factors such as the amount of reliable information available in the speech segment, the uncertainty i.e. the expected squared error varies across feature components. Taking the uncertainties into account in decoding could improve the ASR performance on the estimated clean speech features as proposed in [4]. Recently, using observation uncertainties [5] with an imputation method resulted in significant improvements in speech recognition performance [6]. Using observation uncertainties allows for emphasising the reliable or less uncertain information in decoding and thus compensates for the imperfections in the imputation process.

In this work, we investigate using observation uncertainties with sparse imputation to improve speech recognition performance on the reconstructed features. Since the SI method used in this work does not directly support variance estimation, it is necessary to find heuristic measures to characterise the expected accuracy of the estimated features. The uncertainty measures proposed in this work include accuracy estimates derived from the feature enhancement or MDT principles in general as well as measures specific to SI. The proposed uncertainty measures are evaluated in a large vocabulary continuous speech recognition task with artificially noise corrupted clean speech data and the results compared to the SI baseline.

2. Sparse imputation

2.1. Missing data techniques

Missing data techniques [1] divide the log-spectral representation \( Y \) of a noisy utterance into speech and noise dominated components. Assuming the noise is additive and independent of the clean speech, the speech dominated components \( \hat{Y}_s(t,k) \) are reliable estimates for the underlying clean speech features, \( Y_s(t,k) \approx \hat{S}(t,k) \), where \( \hat{S}(t,k) \) denotes the \( k \)-th spectral channel and \( t \)-th frame of the clean speech signal in the log-spectral domain. The noise dominated components, on the other hand, are unreliable and provide only an upper bound for the corresponding clean speech components, \( \hat{Y}_n(t,k) \). The clean speech information in these components is effectively missing [1]. Reconstruction approaches like SI replace the unreliable values with clean speech estimates \( \hat{S}_u(t,k) \) and approximate clean speech as \( \hat{S} = \hat{Y}_r \cup \hat{S}_u \), where \( \hat{Y}_r \) denotes the reliable observations and \( \hat{S}_u \) the clean speech estimates derived for the unreliable features.

2.2. Missing data masks

A missing data mask \( M \) labels each spectro-temporal component in \( Y \) as reliable, \( M(t,k) = 1 \), or unreliable, \( M(t,k) = 0 \). The mask estimation approach used in this work is based on the negative energy criterion proposed in [1]. The observed features \( Y(t,k) \) are considered reliable if \( \exp(Y(t,k)/N(t,k)) > \gamma_1 \), where \( N \) denotes an estimate for the noise \( \gamma_1 \) is a threshold parameter. In this work, \( N \) is a blind estimate calculated from automatically detected speech pauses [3]. We also conduct experiments with the so-called oracle masks which assume exact knowledge of the speech and noise and label the observed features \( Y(t,k) \) reliable if \( \exp(S(t,k)/N(t,k)) > \gamma_2 \). The threshold parameters are optimised based on performance on the development data described in Section 4.2.

2.3. Sparse imputation

In sparse imputation [2], a moving window is applied to divide the log-spectral representation \( S \) of an utterance into \( T \times K \) dimensional spectrograms, where \( T \) is the window length and \( K \) the number of spectral channels. By concatenating subsequent time frames \( t \), each window is reshaped into \( D = T \cdot K \) dimensional vector \( s(\tau)' = [S(\tau - T/2), \ldots, S(\tau + T/2)]' \) that is represented as a linear combination of example windows i.e.
exemplars $a_n$ as
\[
s(\tau) = \sum_{n=1}^{N} x_n(\tau)a_n = Ax(\tau),
\]
where $x(\tau)$ is an activation vector corresponding to the $\tau$-th window and $A = [a_1 \ldots a_N]$ is a predetermined dictionary of size $D \times N$. In this work, the exemplars $a$ are randomly extracted from clean speech training data as described in Section 4.2. Inspired by research on compressive sensing [7] we determine the activations $x(\tau)$ by solving a minimum least squares problem with an $l^1$ penalty:
\[
x^*(\tau) = \arg\min_{x \in \mathbb{R}^N} \{ ||Ax - s(\tau)||_2 + \lambda ||x||_1 \},
\]
where $\lambda$ is a regularisation parameter. With a sparse activation vector $x^*(\tau)$, the speech token $s(\tau)$ is represented as a linear combination of as few exemplars as possible. When the speech signal is corrupted with additive noise, we calculate the sparse representation based on only the reliable components $Y_{\tau}(t, k)$. The missing data mask $M$ and the observed features $Y$ are windowed and reshaped into vectors $m(\tau)$ and $y(\tau)$ of length $D$ as described earlier. The activations $x(\tau)$ are found using the following norm minimisation:
\[
x^*(\tau) = \arg\min_{x \in \mathbb{R}^N} \{ ||WAx - Wy(\tau)||_2 + \lambda ||x||_1 \},
\]
where the matrix $W = \text{diag}(\mathbb{m}(\tau))$ is used to select the reliable components.

The activation vectors $x^*(\tau)$ could be used to estimate the clean speech features in each window as $s^*(\tau) = Ax^*(\tau)$. In practice, however, only the unreliable elements are imputed: $\hat{s}(\tau, d) = y(\tau, d)$ if the $d$-th component in $y(\tau)$ is reliable and $\hat{s}(\tau, d) = \min\{s^*(\tau, d), y(\tau, d)\}$ if the component is unreliable. The component-wise mini-operation reflects the knowledge that when noise is additive, the clean speech feature cannot exceed the observed noisy speech value. The overlapping clean speech estimates that result from using overlapping windows in the segmentation process are averaged to form the clean speech estimates $\hat{S}(t, k)$ as proposed in [2].

3. Observation uncertainties

3.1. Decoding with observation uncertainties

In traditional feature enhancement or missing data imputation, the reconstructed features $\hat{s}(t)$ in frame $t$ in the acoustic model domain are assumed exact and the likelihood of the $q$-th state of the clean speech acoustic models $M$ is calculated as
\[
L(q) = p(\hat{s}(t) | M, q).
\]
The imputed values in $\hat{s}(t)$ are not, however, perfect estimates for the underlying clean speech but rather they vary in accuracy from frame to frame depending on the available information. Therefore, as proposed in [4], in order to take into account the imperfections in the imputation process, the likelihoods should be calculated based on the complete posterior $p(\hat{s}(t) | \theta, y(t))$, where $\theta$ denotes the parameters used in missing data imputation and $y(t)$ the observed noisy speech features. Calculating the likelihoods then requires integration over all possible feature candidates,
\[
L(q) = \int p(\hat{s}|\theta, y(t)) p(\hat{s}|M, q) \, ds.
\]
Here, the feature candidates $s$ are weighted according to their observation probabilities given $y(t)$. The approach is closely related to uncertainty decoding and has been successfully tried with several feature enhancement techniques; see [8] for a review and discussion.

Assuming the states $q$ are modelled as Gaussian mixtures, the state likelihoods are calculated as a weighted sum over the likelihoods of each mixture component $l$. Furthermore, assuming a Gaussian posterior for the estimated values as proposed in [4], the likelihood of the $l$-th Gaussian component is calculated as
\[
L(l) = \int N(s; \hat{s}(t), \Sigma_1(l)) N(s; \mu(l), \Sigma(l)) \, ds
\]
where $\mu(l)$ and $\Sigma(l)$ are the mean and covariance of the $l$-th Gaussian in the uncompensated clean speech model and $\hat{s}(t)$ and $\Sigma_1(l)$ are the mean and covariance of the clean speech posterior at frame $t$. Thus, decoding with observation uncertainties reduces to adding the estimated uncertainties $\Sigma_1(l)$ to the model variances $\Sigma(l)$, for each frame $l$. 3.2. Mapping uncertainties between domains

The uncertainties $\Sigma_1(l)$ in Equation (6) characterise the observation uncertainty in the acoustic model domain whereas uncertainties derived from the reconstruction process can only reflect the uncertainty related to the imputed features in the log-spectral domain. In this work, the uncertainty estimates $\sigma$ calculated in the log-spectral domain are mapped to the acoustic model domain using a supervised learning approach as proposed in [6]. However, instead of a multilayer perceptron (MLP) proposed in [6], we use a linear transformation
\[
\sigma_1 = R \sigma,
\]
where $\sigma_1 = \text{diag}(\Sigma_1)$ denote the uncertainties in the acoustic model domain. The linear mapping $R$ is calculated from clean speech training data corrupted with pink noise (see Section 4.2). The input uncertainties $\sigma$ are estimated in the log-spectral domain using the measures introduced in Section 3.3 and the target uncertainties $\sigma_1$ are calculated as the squared error between reconstructed and clean speech features in the acoustic model domain.

Note that a linear transformation maps zero uncertainty $\sigma_1(t, k) = 0 \forall k$ in the log-spectral domain to zero uncertainty in acoustic model domain i.e. does not add uncertainty in the system.

3.3. Uncertainty measures for sparse imputation

In feature enhancement, the expected uncertainty is often based on the variance of the enhancement process [4, 6]. Since the sparse imputation method used in this work does not, however, support variance estimation, we propose using heuristic measures to characterise the uncertainty in the reconstructed features $\hat{S}(t, k)$. All the uncertainty measures $\hat{\sigma}(t, k)$ except oracle uncertainties are scaled to $0 \ldots 1$ per utterance. In order not to violate the MDT assumption (Section 2.1) that speech dominated features are reliable estimates for clean speech, the log-spectral domain uncertainties related to reliable features are always assumed zero. The following uncertainty measures are considered in this work:

OS If the clean speech features $S$ are known, an oracle uncertainty estimate can be calculated as the squared error between the reconstructed and clean speech features in the log-spectral domain, $\hat{\sigma}_S(t, k) = (S_n(t, k) - S_e(t, k))^2$.

M1 Assuming all reconstructed features are equally uncertain, the uncertainty may be characterised as $\hat{\sigma}(t, k) = 1 - \frac{1}{n}$.
M(t, k), where M is the missing data mask introduced in Section 2.2.

M2 The missing features could have any value between zero and the observed upper bound Yc(t, k). Hence, the uncertainty related to the features prior to reconstruction is \( \hat{\sigma}_u(t, k) \propto Y_c(t, k) \).

M3 Assuming the clean speech estimates which differ most from the observed values are the most inaccurate, we set the uncertainties proportional to the relative difference \( \hat{\sigma}_u(t, k) \propto (Y_c(t, k) - \hat{S}_e(t, k)) / Y_c(t, k) \).

M4 If a particular observation is difficult to represent sparsely, this can be caused by the observation not being covered by the dictionary A [10]. In a statistical framework, this would result in large predictive variance and high uncertainty related to the reconstructed features. We thus propose setting the uncertainty \( \hat{\sigma}_u(t) \) proportional to the number of exemplars used in reconstructing the segments \( y(\tau) \) that contain the t-th frame and were thus used to estimate the clean speech features \( S_e(t, k) \) in frame t. The uncertainties are defined as \( \hat{\sigma}_s(t) \propto \sum_{\tau} \sum_{f(p : t \land \tau)} f(p : x^* (\tau, d) > 0.01) \), where \( f(p) = 1 \) if the proposition \( p \) is true and zero otherwise and \( t \land \tau \) is true if the \( \tau \)-th window contains the \( t \)-th frame. The \( x^* (\tau, d) \) are the activations from Equation (3). Minor activations \( x^* (\tau, d) \leq 0.01 \) are not considered significant.

M5 Assuming that with more reliable features, it is easier to accurately estimate the unreliable values, the uncertainties \( \hat{\sigma}(t) \) are set proportional to the number of unreliable components in the segments \( y(\tau) \) that contain the \( t \)-th frame and were thus used to estimate the clean speech features \( S_e(t, k) \). The uncertainties are defined as \( \hat{\sigma}(t) \propto \sum_{\tau} \sum_{f(p : t \land \tau)} (1 - m(\tau, d)) \), where \( f(p : t \land \tau) = 1 \) if the \( \tau \)-th window contains the \( t \)-th frame and zero otherwise.

4. Experiments

4.1. Baseline system

In the acoustic model domain, the speech signal is represented with 12 MFCC and a log energy feature along with their first and second differenitals. The features are normalised using cepstral mean subtraction (CMS) and maximum likelihood linear transformation (MLLT). The acoustic models are state clustered hidden Markov triphone models with 1564 states each associated with Gaussian mixtures with a varying number of components and state durations modelled using gamma probability functions; see [9] for details. The 30-hour dataset used for training the acoustic models contains headset recorded clean speech from 203 speakers selected from the Finnish SPEECON database. Among the utterances are words, sentences and spontaneous speech.

The decoder is a time-synchronous beam-pruned Viterbi token-pass system and the language model a variable-length, growing n-gram model that employs morpheme-like subword units called morphs. The morph lexicon was learned from 145 million words of Finnish book and newspaper data which was also used for training the language model. The decoding vocabulary is in practice unlimited since all words and word forms can be represented using the statistical morphs [9]. The speech recognition performance is measured in letter error rate (LER) which is better suited for Finnish than the commonly used word error rate (WER).

### Table 1: Speech recognition results from the experiments with the clean speech data corrupted with babble noise at SNR 10 and 5 dB.

<table>
<thead>
<tr>
<th></th>
<th>clean</th>
<th>SNR 10</th>
<th>SNR 5</th>
<th>SNR 10</th>
<th>SNR 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>BL</td>
<td>3.3</td>
<td>43.4</td>
<td>77.9</td>
<td>43.4</td>
<td>77.9</td>
</tr>
<tr>
<td>SI</td>
<td>3.4</td>
<td>21.6</td>
<td>47.4</td>
<td>6.7</td>
<td>7.5</td>
</tr>
<tr>
<td>OA</td>
<td>3.2</td>
<td>11.4</td>
<td>27.5</td>
<td>5.6</td>
<td>7.3</td>
</tr>
<tr>
<td>OS</td>
<td>3.2</td>
<td>15.6</td>
<td>38.2</td>
<td>5.9</td>
<td>7.4</td>
</tr>
<tr>
<td>M1</td>
<td>3.4</td>
<td>19.2</td>
<td>44.3</td>
<td>6.4</td>
<td>8.9</td>
</tr>
<tr>
<td>M2</td>
<td>3.4</td>
<td>20.8</td>
<td>46.4</td>
<td>6.3</td>
<td>9.1</td>
</tr>
<tr>
<td>M3</td>
<td>4.0</td>
<td>19.3</td>
<td>44.9</td>
<td>6.1</td>
<td>8.0</td>
</tr>
<tr>
<td>M4</td>
<td>3.3</td>
<td>18.6</td>
<td>43.6</td>
<td>5.9</td>
<td>7.8</td>
</tr>
<tr>
<td>M5</td>
<td>3.3</td>
<td>18.4</td>
<td>42.9</td>
<td>6.1</td>
<td>8.1</td>
</tr>
</tbody>
</table>

4.2. Experimental setup

In this work, sparse imputation is applied in 21-channel log-compressed mel-spectral domain. We use the MATLAB implementation described in [3] with window width \( T = 15 \) that was found optimal in our previous experiments [3] on the Finnish SPEECON data. The exemplar dictionary A with \( N = 8000 \) exemplars was randomly extracted from the 8139 read sentences (14 hours) in the SPEECON training data. The mapping \( R \) in Equation (7) was trained for each measure OS and M1–M5 independently using set of 500 sentences (52 minutes) randomly selected from the SPEECON training data. The training data for \( R \) was corrupted with pink noise from the NOISEX-92 database at SNR 10 dB.

The proposed system is evaluated in a noisy speech recognition task with artificially corrupted clean speech data. The development set used for optimising the missing data mask threshold has 1093 read sentences (115 minutes) from 40 speakers in the Finnish SPEECON database. The utterances were corrupted with babble noise and pink noise from NOISEX-92 at SNR 10 dB. The threshold parameters were set at \( \gamma_1 = 3 \text{ dB} \) for babble noise and \( \gamma_2 = 4 \text{ dB} \) for pink noise and \( \gamma_2 = -2 \text{ dB} \) for both noise types. The evaluation sets were constructed of 1118 SPEECON sentences (113 minutes) from 40 speakers and corrupted with babble noise from NOISEX-92 at SNR 10 dB and 5 dB.

4.3. Results

Table 1 presents the speech recognition performance of the baseline system (BL), baseline sparse imputation without uncertainties (SI), sparse imputation with the oracle uncertainties calculated in acoustic model domain (OA), and sparse imputation with uncertainties derived from the oracle uncertainties calculated in the log-spectral domain (OS) or from the uncertainty measures M1–M5 introduced in Section 3.3.

While all the proposed measures improved speech recognition performance over the SI baseline performance on noisy data (SNR 10 and 5 dB). Using the oracle uncertainties calculated in the log-spectral domain (OS) resulted in relative error reduction 19–28 % and using the oracle uncertainties calculated in acoustic model domain (OA) in relative error reduction 42–47 % in the same conditions.
The results in Table 1 also show that using SI resulted in speech recognition performance substantially better than the baseline system when oracle masks are used. This is in line with the results reported in our previous work [3]. Using the proposed uncertainty measures at SNR 10 dB results in a relative error reductions up to 12% compared to the baseline SI. This is equal to the relative error reduction achieved with the log-spectral domain oracle uncertainties (OS). At SNR 5 dB, however, none of the proposed measures improves the results over the SI baseline.

5. Discussion

Using the proposed uncertainty measures M1–M5 with sparse imputation improved the speech recognition performance on the reconstructed features when estimated mask were used. The best results were achieved with the measures that set the uncertainty proportional to the number of exemplars (M4) or the number of unreliable components (M5) used in sparse imputation. They differ from the other measures on two accounts: they are frame-based rather than component-based and they consider the full context used in sparse imputation. Both differences may have contributed to the result since using a single mapping R on all frames may have favoured measures that exhibit less variation, but also, using the same time-context as SI may have been an advantage.

Using the missing data mask (M1) can be considered a reference test for other uncertainty measures since each reconstructed component is assigned an equal uncertainty. M2 and M3 having a performance lower than M1 when estimated masks were used suggests that adding information specific to the component only resulted in more variation i.e. noisier uncertainties. M3 performed, however, relatively better with oracle masks, which suggests the measure may be sensitive to mask estimation errors. When oracle masks were used, M1 and M2 resulted in lowest performance. They are both a priori uncertainty measures which do not attempt to estimate the reconstruction accuracy.

When oracle masks were used at SNR 5 dB, all the proposed uncertainty measures failed to improve speech recognition performance over the baseline SI. This is not surprising considering that even the system using oracle uncertainties calculated in the acoustic model domain (OA) did not notably improve the results compared to the baseline SI system that has no knowledge of uncertainties; the relative error reduction was less than 5%.

The results with oracle uncertainties calculated in the acoustic model domain (OA) were considerably better than the results with oracle uncertainties calculated in the log-spectral domain (OS). This is partly because (i) the log-spectral domain uncertainties are restricted to compensate for estimation errors in the unreliable features while the oracle uncertainties calculated in acoustic model domain also compensate for e.g. mask estimation errors, but also because (ii) the mapping between domains is inaccurate. In theory, the dependencies between log-spectral and acoustic model domain uncertainties are nonlinear, but in a pilot experiment with OS, using a neural network for nonlinear mapping did not improve the speech recognition performance over results from experiments using the linear mapping R.

While using observation uncertainties resulted in notable error reductions especially in the noisy conditions, the results on clean speech data could be more significant for sparse imputation in general. In [3], it was shown that sparse imputation increases the number of insertion errors on clean speech data if inaccurate missing data masks are used. The type of mask estimation errors that cause the insertions are the relatively isolated noisy features that should have been labelled unreliable. Using such isolated evidence for imputation results in high uncertainty when using measures M4 and M5, and as a result, the effect of the mask estimation errors is compensated indirectly.

6. Conclusions and future work

Using observation uncertainties derived from heuristic uncertainty measures proposed in this work improved the speech recognition performance on features reconstructed using SI. Further improvements are possible if several measures were combined. In a probabilistic sparse imputation framework, the heuristic measures could either be replaced or augmented with the predictive variances. Mapping the uncertainties from log-spectral to acoustic model domain could be improved by clustering the uncertainties and training a separate transformation for each cluster. Additionally, extending the time context in the log-spectral domain should improve modelling the uncertainty related to differential features.

7. Acknowledgments

The research of JFG was carried out in the MIDAS project within the STEVIN programme funded by the Dutch and Flemish Governments. UR was supported by the Hecse graduate school and KJP by the Academy of Finland in the projects “Auditory approaches to automatic speech recognition” and AIBC.

8. References