Coupling identification and reconstruction of missing features for noise-robust automatic speech recognition

Ning Ma and Jon Barker

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{n.ma, j.barker}@dcs.shef.ac.uk

Abstract

The standard missing feature imputation approach to noise-robust automatic speech recognition requires that a single foreground/background segmentation mask is identified prior to reconstruction. This paper presents a novel imputation approach which more closely couples the identification and reconstruction of missing features by using a probabilistic framework based on the speech fragment decoding technique. Using fragment decoding, the most joint-likely state sequence and segmentation hypothesis is identified with which the missing data region is imputed. Crucially, however, imputation can exploit the speech state sequence recovered by the fragment decoding. Further, using N-best decodings allows the clean spectrogram to be estimated as a weighted combination of reconstructions which provides some allowance for uncertainty in the estimates. Experiments on the PASCAL CHiME Challenge task show that system performance is highly dependent on the complexity of the speech models used for segmentation and imputation, and by exploiting the temporal constraint of speech the system significantly outperforms those that ignore the constraint.

Index Terms: Missing feature reconstruction, noise-robust speech recognition, feature compensation, fragment decoding.

1. Introduction

Unpredictable background noise remains one of the greatest challenges for automatic speech recognition (ASR) systems. Over decades, an extremely diverse set of techniques have been proposed in an effort to combat noise corruption. Some techniques have focused on feature compensation which tries to reduce the effect of noise in the incoming features, while others have focused on model compensation which transforms speech models to better accommodate background noise.

This paper is concerned with an alternative uncertainty-based technique – the missing data approach to noise-robust ASR [1]. This approach assumes that when speech is corrupted by additive noise, in some ‘reliable’ time-frequency (T-F) regions the effect of the noise can be ignored while in ‘unreliable’ regions speech is completely masked by noise. The unreliable regions are treated by either marginalisation – considering all possible values that the clean speech may have taken given the noisy observation, or imputation – replacing noise-corrupted regions with estimates of the clean speech signal. The marginalisation technique has better theoretical justification but is difficult to apply in the cepstral domain, while imputation-based techniques can be used as a front-end which generates cepstral features for processing by a conventional ASR system.

The imputation approach requires solutions to two problems. First, reliable speech regions need to be identified. Such information is typically represented by a spectro-temporal segmentation mask, in which each T-F component is labelled as being either reliable or unreliable. Second, in the unreliable regions speech features need to be reconstructed. A widely-used method, the minimum mean square error (MMSE) estimation, exploits the distribution of unreliable features conditioned on the reliable features to estimate the masked clean speech, and the distribution is typically modelled by a Gaussian mixture model (GMM). Imputation is approximated as the expected value of such a distribution [1–4]. More recent work has focused on exploiting the temporal pattern of speech signals for imputation [4, 5]. However, all the aforementioned imputation methods require that a single foreground/background segmentation mask is identified prior to reconstruction.

This paper presents a novel imputation approach which more closely couples the identification and reconstruction of missing feature by using a probabilistic framework based on the speech fragment decoding (SFD) technique [6]. Representing clean speech spectra using hidden Markov models (HMMs) in order to exploit the temporal pattern, the system searches for the jointly-optimal segmentation masks and state sequences based on a maximum likelihood criterion. Crucially, imputation can exploit the speech state sequence recovered by using marginalisation. Further, using N-best decodings allows the clean spectrogram to be estimated as a weighted combination of spectrograms reconstructed under each segmentation/state sequence hypothesis using the MMSE estimation. An overview is shown in Fig. 1. Next section will review the MMSE imputation technique. The speech fragment decoding framework is reviewed in Section 3. Section 4 presents the novel imputation approach using this framework. Experiments and discussion are given in Section 5. Section 6 concludes this paper.

2. Missing feature imputation

Let \( x \), \( n \) and \( y \) represent feature vectors of log-spectral energy of speech, noise and their mixture, respectively. The log-spectral energy of the mixture is well approximated by the maximum of the individual log-spectral properties, commonly known as the log-max approximation [7]. Under this approximation,
the noisy spectra can be segmented into two mutually exclusive sets: (i) reliable regions, \( R \), where the observed energy is close to the speech energy, i.e. \( x_i \approx y_i \), where \( r \in R \), and (ii) unreliable regions, \( U \), where speech is masked by the noise energy, but the unobserved speech energy has a value less than the observed masking energy, i.e. \( x_i < y_i \), where \( u \in U \).

A popular imputation method exploits the knowledge of the reliable features in conjunction with the joint statistics of all the features from the same vector \([1, 2]\). The MMSE estimation of the unreliable feature \( x_u \) is

\[
\hat{x}_u = E[x_u| x_R, x_u < y_u, q] = \int_{-\infty}^{y_u} p(x_u|x_R, y_u, q) x_u dx_u
\]

where \( q \) represents a Gaussian mixture model trained on clean speech spectra. The conditional distribution can be expanded by marginalising over the mixture component \( k \),

\[
\hat{x}_u = \sum_{k=1}^{M} P(k|x_R, y_u, q) \int_{-\infty}^{y_u} p(x_u|x_R, y_u, k, q) x_u dx_u
\]

where \( M \) is the number of mixture components. Let \( \hat{x}_u^k \) represent the integral, which is the expected value from the mixture component \( k \). Then (3) is computed as a linear combination of \( \hat{x}_u^k \). The weight \( P(k|x_R, y_u, q) \) is given by Bayes’ rule,

\[
P(k|x_R, y_u, q) = \frac{P(k|q)p(x_R, y_u|k, q)}{\sum_{k=1}^{M} P(k|q)p(x_R, y_u|k, q)}
\]

(4)

\( p(x_R, y_u|k, q) \) can be computed using the bounded marginalisation technique [1],

\[
p(x_R, y_u|k, q) = \prod_{r \in R} p(x_r|k, q) \prod_{u \in U} \int_{-\infty}^{y_u} p(x_u|k, q) dx_u
\]

(5)

If the independence between the reliable and unreliable features is assumed, the expected value \( \hat{x}_u^k \) in (3) becomes

\[
\hat{x}_u^k = \int_{-\infty}^{y_u} p(x_u|k, q) x_u dx_u
\]

(6)

When the GMM employs a diagonal covariance matrix, the expectation is effectively the mean value of the mixture component \( k \) at the index of \( u \), but bounded by the observation.

### 3. Speech fragment decoding

Speech fragment decoding, largely inspired by Bregman’s account of the auditory scene analysis [8], couples two problems that are usually addressed sequentially: source segmentation and source recognition. Let \( Y \) represent a sequence of noisy observation vectors. The decoding task is to find the most joint-likely state sequence \( Q \) and foreground/background segmentation \( S \) given these observations. The underlying speech vector sequence \( X \) is not directly observed but can be introduced by integrating over all possibilities,

\[
\hat{Q}, \hat{S} = \operatorname{argmax}_{Q,S} \int_X P(Q, X, S|Y) dX
\]

(7)

Using the product rule and Bayes’ rule, (7) can be expanded as

\[
\hat{Q}, \hat{S} = \operatorname{argmax}_{Q,S} P(S|Y) \int_X P(X|Q) \frac{P(X|Y, S)}{P(X)} dXP(Q)
\]

(8)

\( P(X|Q) \) represents the distributions learned from noise-free speech and is typically modelled by Gaussian mixtures. The term \( P(X|Y, S) \) is the masking model: the speech observation \( X \) must take the value of the noisy observation \( Y \) for features that are labelled as reliable by \( S \). The \( X \) must have a value less than \( Y \) for features that are marked as unreliable by \( S \). For values of \( X \) which violate these rules the probability is set to 0.

\( P(S|Y) \) is the segmentation model and the segmentation search is equivalent to selecting the optimal segmentation mask. An exhaustive search is clearly not practical. Fortunately, in speech fragment decoding a ‘segmentation’ front-end uses a variety of primitive grouping processes (e.g. multi-pitch analysis [9]) to locally group T-F elements into fragments of energy that are believed to originate from a common source. Elements in a fragment belong together and hence share the same foreground or background label. The fragments are imposing a form for \( P(S|Y) \) – it assigns equal probability to any foreground/background segmentation that can be constructed from the set of fragments. All other segmentations are assigned a probability of 0.

Under a foreground/background segmentation hypothesis, the state likelihood can be evaluated using bounded marginalisation as shown in (5). The maximisation over both state sequence and segmentation is achieved via a Viterbi search over a lattice of segmentation and state sequence hypotheses.

### 4. Imputation based on SFD

SFD simultaneously searches for the optimal state sequence \( Q \) and the optimal segmentation \( S \), and its output can be used in various ways to impute the missing data. First, the optimal segmentation mask can be employed solely so that any imputation technique that requires a priori defined mask can be applied (e.g. the MMSE-based estimation [2]). This method will be called ‘mask-based imputation’ and discussed in Section 4.1. Second, the overall optimal HMM state sequence, identified together with the optimal segmentation mask, can be used to restore the speech spectrogram. This will be called ‘SFD-based imputation’ and discussed in more details in Section 4.2.

#### 4.1. Mask-based imputation

Mask-based imputation makes use of only the segmentation mask identified by speech fragment decoding and the segmentation might not be consistent with the models used for imputation. Essentially, the optimal speech model state sequence identified by SFD is discarded and imputation is performed by using a separately trained models. For example, the MMSE estimation technique discussed in Section 2 employs a GMM with diagonal covariances. In this paper, the MMSE imputation technique is extended to use a probabilistic segmentation mask.

The optimal segmentation \( \hat{S} \) represents a sequence of vectors \( \{\hat{s}^1, \ldots, \hat{s}^T\} \), where each \( \hat{s}^i \) is a vector of indicator variables stating whether the corresponding time-frequency element is reliable or unreliable. For the sake of simplicity the index \( i \) is omitted for \( \hat{s} \) and other variables, so that \( \hat{s} \) still represents a vector and \( \hat{s}_i \) is a scalar. The segmentation mask can also be probabilistic. In this case, imputation has to be performed for every feature dimension \( i \),

\[
\hat{x}_{i} = \sum_{k=1}^{M} P(k|x, \hat{s}) \hat{x}_u^k
\]

(9)

where \( \hat{x}_u^k \) is the MAP estimate from mixture component \( k \) and
the weight \( P(k|x, y, \hat{s}, q) \) is given by Bayes’ rule,
\[
P(k|x, y, \hat{s}, q) = \frac{P(k|q)p(x, y, \hat{s}|k, q)}{\sum_{k'=1}^{M} P(k|q)p(x, y, \hat{s}|k', q)} \tag{10}
\]
\( p(x, y, \hat{s}|k, q) \) can be computed as
\[
p(x, y, \hat{s}|k, q) = \prod_{i=1}^{D} \left( \hat{s}_i f(x_i) + (1 - \hat{s}_i) \int_{-\infty}^{\infty} f(x_i) dx_i \right)
\tag{11}
\]
where \( f(x_i) = p(x_i|k, q) \) and \( D \) is the feature dimensionality. Again the independence between the reliable and unreliable features is assumed. When \( q \) employs a diagonal covariance matrix, the expectation \( \hat{x}_i \) is effectively the bounded mean of the mixture component \( k \) at the index of \( i \). Finally, the restored feature is smoothed as a linear combination of the observation and the imputed value from (9),
\[
\hat{x}_i = \hat{s}_i y_i + (1 - \hat{s}_i) \sum_{k=1}^{M} P(k|x, y, \hat{s}, q) \hat{x}_i^k \tag{12}
\]
When \( \hat{s}_i \) is close to one, indicating the observation is more likely speech, the restored feature will be close to the observation. When \( \hat{s}_i \) is close to zero, indicating the observation is more likely to be masked by noise, the restored feature will be close to the the imputed value.

### 4.2. SFD-based imputation

The SFD-based imputation technique employs HMMs to model log-compressed spectral energy of clean speech signals and to reconstruct the underlying speech spectrogram from noisy signals. Each HMM state can be modelled by either a single Gaussian distribution or a mixture of Gaussians. Typically, a training set of clean speech is used and the training stage is similar to training acoustic models in speech recognition using the Baum-Welch re-estimation formulae.

Reconstruction of speech spectrogram involves the most likely HMM state sequence \( Q = \{q_1, ..., q_T\} \), together with the most likely segmentation \( \hat{S} \), obtained from the fragment decoding pass described in Section 3. The missing features according to the most likely segmentation are imputed from the state \( q_i \) for each frame in the optimal state sequence using (12).

The imputation method discussed above considers only a single state sequence and segmentation mask. To provide some allowance for uncertainty in the segmentation estimates, \( N \)-best decodings are used for imputation. Each of the \( N \)-best hypotheses provides a speech state sequence as well as a jointly optimal segmentation mask, which can reconstruct a spectrogram. The final clean spectrogram is estimated as a linear combination of reconstructions from each hypothesis weighted by their posterior probabilities at each frame.

The speech models employed in SFD-based imputation can be of different complexity – a single-state HMM which ignores the temporal pattern of speech, effectively a GMM, or cross-word triphone HMMs with a well-constructed grammar. Theoretically, the better the HMMs represent the speech, the better the missing features can be reconstructed. First, with more detailed speech models the most likely state sequence identified by SFD will better represent the temporal pattern of the underlying speech signals. Second, more detailed models of the target sound source will provide more top-down constraint when selecting bottom-up fragments of the target source to form a better foreground/background segmentation. Section 5 will discuss the effect of model complexity on imputation performance based on speech recognition experiments.

### 5. Experiments and Results

Evaluation was based on the PASCAL CHiME Challenge task [10]. The task entails the recognition of short command utterances that have been mixed into binaural recordings made in a noisy domestic environment after convolution with room impulse responses. The average recognition accuracy for the letter-digit keywords is used to report the keyword accuracy.

#### 5.1. Experimental setup

The log-spectral features used by imputation were produced via a 32-channel gammatone filterbank distributed in frequency between 50 Hz and 8000 Hz on the equivalent rectangular bandwidth scale. They were then log-compressed and supplemented with deltas to form 64-dimensional feature vectors.

Four SFD-based imputation systems employing HMMs of different complexity were tested. The first system, termed S-GMM, employed a single-state HMM, i.e. a GMM with a diagonal covariance matrix to represent the distribution of the clean speech log-spectra. The GMM was trained using the training set via the standard EM algorithm and the number of Gaussian components was gradually increased to 512. The second one, S-Mono, employed a set of 3-state left-to-right monophone models (34 monophones in total), and each state is modelled by a mixture of 32 Gaussians with diagonal covariances. They were employed by SFD with an ergodic grammar so that all the monophone models are placed in parallel. The third system, S-MonoG, used the same monophone models as in S-Mono, but they were employed during decoding with a lexicon and a grammar. The last system, S-Word, employed a set of word-level HMMs in which the number of states is based on 2 states per phoneme and each state is modelled by 7 Gaussians. The same grammar as in S-MonoG was used during decoding. In all the SFD-based imputation systems, the top 5 state sequences/segmentation hypotheses were employed to reconstruct the speech spectrogram, as discussed in Section 4.2.

The mask-based imputation system, MMSE, employed the same GMM as in the S-GMM system for imputation, and the most likely segmentation mask from the S-Word system. As discussed in Section 4.1, this imputation system is similar to the commonly used cluster-based imputation [2] except that a probabilistic segmentation mask is used.

In the final speech recognition stage, the reconstructed spectral features were transformed to 13-dimensional cepstral domain via the DCT transformation. The cepstral features were then applied cepstral mean normalisation and appended with deltas and delta-deltas. Speaker-dependent word HMMs were used in all ASR experiments, following the standard model setup of the PASCAL CHiME Challenge. Each HMM state employed 7 Gaussians with diagonal covariance and the models were trained in a multicondition style using data mixed from -6 dB to 18 dB SNRs plus the noise-free set.

#### 5.2. Results and analysis

The ASR results using various imputation techniques are reported in Tables 1 and 2 for the development and test sets, respectively. The results of the multicondition (MC) training baseline are also included. All the imputation systems produced substantial improvement over the MC training baseline.
Table 1: Keyword recognition accuracies (%) of the MC baseline and various imputation systems for the development set.

<table>
<thead>
<tr>
<th></th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>57.92</td>
<td>65.67</td>
<td>74.17</td>
<td>81.50</td>
<td>87.67</td>
<td>91.08</td>
</tr>
<tr>
<td>S-GMM</td>
<td>64.25</td>
<td>68.08</td>
<td>77.17</td>
<td>83.75</td>
<td>88.83</td>
<td>90.75</td>
</tr>
<tr>
<td>S-Mono</td>
<td>69.92</td>
<td>74.50</td>
<td>81.25</td>
<td>85.83</td>
<td>89.17</td>
<td>91.50</td>
</tr>
<tr>
<td>S-MonoG</td>
<td>71.67</td>
<td>76.33</td>
<td>83.33</td>
<td>86.17</td>
<td>89.92</td>
<td>92.33</td>
</tr>
<tr>
<td>S-Word</td>
<td>75.42</td>
<td>79.58</td>
<td>84.75</td>
<td>87.67</td>
<td>91.00</td>
<td>92.50</td>
</tr>
<tr>
<td>MMSE</td>
<td>70.50</td>
<td>71.92</td>
<td>80.58</td>
<td>85.75</td>
<td>90.17</td>
<td>91.00</td>
</tr>
<tr>
<td>Oracle</td>
<td>91.75</td>
<td>92.08</td>
<td>92.58</td>
<td>93.33</td>
<td>93.92</td>
<td>94.50</td>
</tr>
</tbody>
</table>

Table 2: Keyword recognition accuracies (%) of the MC baseline and various imputation systems for the evaluation test set.

<table>
<thead>
<tr>
<th></th>
<th>-6 dB</th>
<th>-3 dB</th>
<th>0 dB</th>
<th>3 dB</th>
<th>6 dB</th>
<th>9 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MC</td>
<td>58.25</td>
<td>66.33</td>
<td>76.58</td>
<td>84.25</td>
<td>88.50</td>
<td>92.92</td>
</tr>
<tr>
<td>S-GMM</td>
<td>66.33</td>
<td>70.25</td>
<td>78.75</td>
<td>86.00</td>
<td>89.50</td>
<td>92.67</td>
</tr>
<tr>
<td>S-Mono</td>
<td>71.42</td>
<td>76.08</td>
<td>82.42</td>
<td>87.50</td>
<td>89.08</td>
<td>92.83</td>
</tr>
<tr>
<td>S-MonoG</td>
<td>75.58</td>
<td>77.83</td>
<td>84.67</td>
<td>87.33</td>
<td>90.83</td>
<td>92.25</td>
</tr>
<tr>
<td>S-Word</td>
<td>78.50</td>
<td>81.25</td>
<td>85.58</td>
<td>88.25</td>
<td>91.08</td>
<td>93.33</td>
</tr>
<tr>
<td>MMSE</td>
<td>71.75</td>
<td>76.50</td>
<td>81.67</td>
<td>87.92</td>
<td>90.08</td>
<td>93.00</td>
</tr>
<tr>
<td>Oracle</td>
<td>92.25</td>
<td>92.42</td>
<td>93.67</td>
<td>94.58</td>
<td>95.08</td>
<td>95.58</td>
</tr>
</tbody>
</table>

For SFD-based imputation, the level of HMM complexity is clearly a key factor for the imputation and ASR performance. The S-GMM system employed a GMM to model speech spectra and therefore exploited no temporal information. The S-Word system employed word-level HMMs and was able to apply more temporal constraints during imputation than the other systems. It is clear that the better the set of HMMs employed represent the dynamics and the context of speech, the higher keyword recognition accuracies the ASR systems are able to achieve. Not only do the temporal constraints affect decoding of the state sequences decoded, but also the quality of the segmentation masks constructed by labelling fragments as either foreground and background. This is particularly the case at the low SNR end, where there is more noise corruption and therefore better acoustic models can bring more improvement.

The mask-based MMSE system and the S-GMM system employed the same GMM for imputation, but the difference lay in the segmentation masks used. The MMSE results were produced by using the most likely segmentation from the best SFD-based imputation system, S-Word, while the mask used by S-GMM was produced via fragment decoding using the GMM. It is not difficult to see that the quality of masks affects the imputation and ASR accuracy. This is further confirmed by results achieved when the oracle masks, obtained by comparing pre-mixed signals, were used in the MMSE system.

The MMSE and S-Word systems employed the same segmentation mask, but while the MMSE system ignored the temporal pattern of speech, the S-Word system modelled the speech statistics with a set of HMMs. The advantage of exploiting the temporal constraints is clearly demonstrated in the improvement of ASR accuracies across all SNRs, even though the overall number of Gaussian states in the two systems is roughly the same.

Fig. 2 shows examples of the reconstructed spectrograms. The fragments used are shown in (c) in different colours. The best segmentation mask selected by the S-Word system is shown in (d), with black representing reliable speech regions. Although the same mask was used, it is clear that the spectrogram restored by the S-Word system (f) resembles the clean speech spectrogram (b) much better than the MMSE system (e), e.g. the MMSE system failed to impute speech in low frequency around 0.1 second and in high frequency around 1.4 seconds.

6. Conclusions

This paper has presented a novel approach that couples the problems of missing feature identification and imputation in a unified probabilistic framework based on the speech fragment decoding technique. The system models the temporal pattern of speech using HMMs, and jointly searches for the N most likely state sequences and segmentations. For each hypothesis, the state sequence is used to reconstruct the missing features identified according to the corresponding segmentation mask. The noise-free spectrogram is estimated as a weighted combination of these N-best reconstructions and is then transformed into the cepstral domain in the speech recognition stage. Experiments on the PASCAL CHiME Challenge task have shown that system performance is highly dependent on the complexity of the speech models used for segmentation and imputation, but even while using simple models performance substantially outperforms multicondition training.

7. References