# RECOGNITION OF REVERBERANT SPEECH USING FULL CEPSTRAL FEATURES AND SPECTRAL MISSING DATA

*Kalle J. Palomäki*[1,2], *Guy J. Brown*[1] *and Jon P. Barker*[1]

[1]Department of Computer Science, University of Sheffield, United Kingdom
[2]Laboratory of Computer and Information Science, Helsinki University of Technology, Finland
kalle.palomaki@hut.fi, g.brown@dcs.shef.ac.uk, j.barker@dcs.shef.ac.uk

## ABSTRACT

We describe a novel approach to feature combination within the missing data (MD) framework for automatic speech recognition, and show its application to reverberated speech. Likelihoods from a spectral MD classifier are combined with those from a full cepstral feature vector-based recogniser. Even though the performance of the cepstral recogniser is substantially below that of the MD recogniser, the combined recogniser performs better in all conditions. We also describe improvements to the generation of time-frequency masks for the MD recogniser. Our system is compared with a previous approach based on a hybrid MLP-HMM recogniser with MSG and PLP feature vectors. The proposed system has a substantial performance advantage in the most reverberated conditions.

## 1. INTRODUCTION

Room reverberation remains a significant challenge for robust automatic speech recognition (ASR) in real-world environments. Previously, we have shown that the missing data (MD) technique for dealing with additive noise in ASR [1, 2] can also be used to improve robustness to convolutional interference, such as reverberation [3]. In this approach, a conventional hidden Markov model (HMM) recogniser is modified to deal with missing or unreliable acoustic features [1, 2]. More specifically, the decoder is provided with spectral features and a time-frequency mask; each element in the mask indicates whether the corresponding feature constitutes reliable evidence of the speech signal or not.

Reverberation consists of a direct sound component followed by an exponentially decaying tail of reflections. The latter effectively smooths the temporal structure of speech, while only strongest low-frequency speech modulations remain less affected. Hence, in our previous approach a missing data mask was derived by selecting time-frequency regions in which strong speech modulations were present, as determined by modulation filtering. Previously, modulation filtering has been used to obtain noise robust feature vectors for reverberant speech recognition [4], and for dealing with transmission line distortion and additive noise [5].

A drawback of the MD approach is that it requires spectral features, which are more correlated than the cepstral features normally used with HMM-based ASR systems. Robust estimation of full covariance spectral models requires more data than is typically available. So rather than use full covariance, the data is usually modelled crudely using a Gaussian Mixture Model (GMM) with a small number of components. The problem of adequately modelling spectral data may result in the robustness obtained using MD techniques being offset by a fall in baseline recognition accuracy. This was evident in our previous study [3].

Previously, Kingsbury [4] has shown that good performance across a range of reverberation conditions can be obtained by combining the posterior probabilities from two recognisers. Specifically, he combined recognisers that used modulation filtered spectrogram (MSG) and cepstral perceptual linear prediction (PLP) coefficients. Even though the performance of the PLP system was substantially below that of the MSG system in reverberant conditions, the combined system was better than either of these alone. Posterior probability combination has also been used for noise robust ASR in the multiband approach [6]. To date, the use of feature combination in the missing data framework for ASR has remained an untouched issue. We address this issue here by showing how spectral features and cepstral features can be combined within the missing data framework, and also describe improvements to the missing data mask generation.

## 2. METHOD

### 2.1. Speech material

The Aurora 2.0 English language telephone digit recognition corpus (sampling rate 8 kHz) was used for evaluation of the system [7]. Acoustic models were trained using the clean (noiseless) speech from the clean training section of the corpus (8440 utterances). For recogniser development and testing, clean speech samples were used for the non-reverberant condition, and the same clean speech samples were convolved with room impulse responses (RIRs) to provide the reverber-

ant condition. All 1001 `Clean1` test set utterances were selected for the test set, and 300 (150 male, 150 female) different utterances from `Clean3` were randomly selected for the development set. Six RIRs were used which were characterized in terms of reverberation time T60 (i.e. the time required for the reverberant sound field to drop by 60 dB after sound offset) and source to microphone distance (S/R), as shown in Table 1. RIR1–RIR4 were those used by Kingsbury, and were recorded in a varechoic chamber [4]; page 90. A further two impulse responses denoted RIR5 and RIR6 (not used by Kingsbury) were measured in a larger room. All RIRs were used for recogniser testing and RIRs 2, 4 and 6 were used for development.

### 2.2. Features

Spectral features were obtained from a peripheral auditory model, which was based on a gammatone filterbank with 32 channels. Channel center frequencies were spaced uniformly on the equivalent rectangular bandwidth (ERB)-rate scale, and had a constant ERB bandwidth of 0.887. The lowest and highest center frequencies were 50 Hz and 3850 Hz, respectively. In order to produce features for the recognizer, the Hilbert envelope of each channel was extracted, smoothed with a low-pass filter (8-ms time constant), sampled at 10 ms intervals and compressed by raising to the power 0.3. These features were supplemented with their temporal derivatives, giving a total of 64 features per vector. Spectral normalisation of features was performed as described in [3].

Mel frequency cepstral (MFCC) feature vectors were computed as described in Aurora 2.0. From the two alternatives (with marginal differences) for generating MFCC features in Aurora 2.0, we used the version which is based on the HTK-implementation. The MFCC feature vectors consists of 12 mel-cepstral coefficients (the zeroth term was excluded) with cepstral liftering, logarithmic frame energy, and first and second order temporal derivatives (a total of 39 features).

### 2.3. Training

Three systems were trained using HTK 3.2, which used spectral features alone, MFCC features alone, or combined spectral and MFCC features. The combined training of the spectral features and MFCCs was required to guarantee that automatically defined temporal alignments matched. However, a concern was that training these features in combination could produce suboptimal alignments for either feature alone. To test this possibility, we trained a systems based on spectral and MFCC features alone for comparison.

Diagonal-covariance Gaussian mixture models (GMM) were used to model the spectral features and MFCCs, with seven and three mixture components (for both combined and alone systems), respectively. For the combined system, separate streams were defined for spectral and MFCC features in HTK training. Then for all systems, whole word models were trained for digits "oh", "zero" and "one" to "nine" with 16 no-skip straight through states. In addition, a silence model and short-pause model were trained, with three and one states respectively.

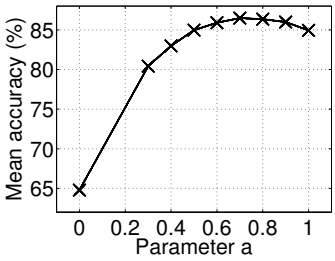


**Fig. 1**. Dependence of mean (over RIRs 2, 4 and 6) recognition accuracy on the combination factor $a$, obtained from the development set. Crosses show the measured data points.

The setting of the GMM variance floor was found to be critical in order to achieve good results. The variance floor was set based on experiments with the development set. For combined spectral and MFCC feature training the floors were set to 0.5 and 1.0 times the global variance for spectral and MFCC features, respectively. For training spectral and MFCC features alone the floors were set to 0.5 and 0.3, respectively. We note that those values are much larger than the variance floor of 0.01 used in the original Aurora 2.0 framework.

### 2.4. Recognition with combined features

Multiple representations are useful in ASR because different sets of features can contain complementary information. For example, if heavy processing to achieve noise robustness is required for one set of features, this may produce a counter effect with clean speech due to loss of fine structure. Previously, feature stream combination has been implemented by either supplementing the existing feature vector with another set of features (such as deltas), or by combining posterior probabilities (or likelihoods) of two recognisers that use different sets of features. In this paper, we introduce an approach in which the likelihoods from spectral missing data and cepstral feature streams are combined.

In the missing data approach, unreliable spectral features $x_{s,u}$ are classified differently from reliable ones $x_{s,r}$, which are passed directly to the recogniser. Here we used bounded marginalisation to deal with unreliable features. If the true value of the unreliable features is known to lie within low $x_{s,u,low}$ and high $x_{s,u,high}$ bounds, an estimate $\overline{f(x_s|C)}$ of the likelihood $f(x_s|C)$ can be obtained as follows,

$$\overline{f(x_s|C)} = \frac{\sum_{k=1}^{M} P(k|C) f(x_{s,r}|k,C) \frac{1}{x_{s,u,high} - x_{s,u,low}}}{\int_{x_{s,u,low}}^{x_{s,u,high}} f(x_{s,u}|k,C) dx_{s,u}} \quad (1)$$

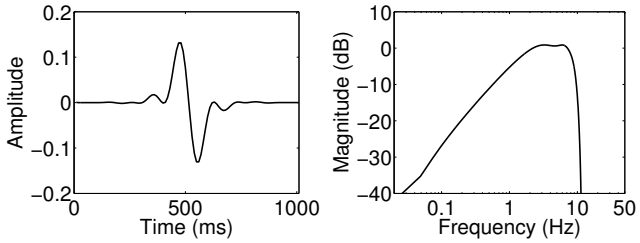


**Fig. 2**. Impulse (left) and magnitude (right) responses of the modulation filter used in this study.

where $k$ denotes the Gaussian mixture component with a diagonal covariance and $C$ is a class of speech sound.

Here, we also have a stream of cepstral features with likelihood $f(x_c|C)$. The combined likelihood $f(x_{s,c}|C)$ is obtained as the weighted average of missing data spectral $\overline{f(x_s|C)}$ and full cepstral $f(x_c|C)$ likelihoods in the logarithmic domain

$$\log(f(x_{s,c}|C)) = a\,\log(\overline{f(x_s|C)}) + (1-a)\,\log(f(x_c|C)) \tag{2}$$

where the weight $a = 0.7$ was chosen to yield the best average performance in recognition experiments using the development set (see Fig. 1).

### 2.5. Mask estimation

Previously we have shown that modulation filtering can be used to generate a mask for missing data ASR, since it detects strong speech onsets that have not been contaminated by reverberation [3]. The spectral features $x_s(i,j)$ are filtered along their time trajectories using a band-pass modulation filter $h(k)$, with 3dB cutoffs at 1.5 Hz and 8.2 Hz,

$$x_{s,bp}(i,j) = \sum_{k=-\infty}^{+\infty} h(k)x_s(i-k,j) \tag{3}$$

where $i$ indexes discrete time (in frames) and $j$ is the frequency channel. The modulation band-pass filter was designed by convolving a linear FIR low-pass filter $h_{lp}$ with a differentiator $h_{diff}$ (i.e., $h = h_{diff} \otimes h_{lp}$). The impulse and frequency responses of the filter $h$ are shown in Fig. 2; for detailed filter parameters see [3]. Filtered features $x_{s,bp}$ are then passed through an automatic gain control (AGC)

$$x^{agc}_{s,bp}(i,j) = \frac{x_{s,bp}(i,j)}{\sum_{k=-\infty}^{+\infty} w(k)|x_{s,bp}(i-k,j)|} \tag{4}$$

where the denominator describes a convolution of $|x_{s,bp}|$ with a triangular window $w(k)$ of length 400 ms, which acts as a low-pass filter. Kingsbury [4] uses a similar approach for producing noise robust feature vectors.

The time-frequency mask $m(i,j)$ is produced by applying a threshold $\theta(j)$ to $x^{agc}_{s,bp}(i,j)$:

$$m(i,j) = \begin{cases} 1 & \text{if } \{x^{agc}_{s,bp}(i,j) - \min_i[x^{agc}_{s,bp}(i,j)]\} > \theta(j) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The threshold $\theta(j)$ is defined for each utterance according to the extent to which it is reverberated; this is estimated according to a metric $B$ which quantifies the 'blurredness' of the speech temporal structure:

$$B = \frac{1}{J}\sum_{k=1}^{J}\left\{\frac{\frac{1}{I}\sum_{i=1}^{I} x_s(i,j)}{\max_i[x_s(i,j)]}\right\} \tag{6}$$

Here, $J = 32$ is the number of frequency channels and $I$ is the length of the utterance. $B$ is then mapped to the threshold $\theta(j)$ as follows,

$$\theta(j) = \gamma\frac{\frac{1}{I}\sum_{i=1}^{I}\{x^{agc}_{s,bp}(i,j) - \min_i[x^{agc}_{s,bp}(i,j)]\}}{1 + exp(-\alpha(B-\beta))} \tag{7}$$

where $\alpha = 19$, $\beta = 0.43$ and $\gamma = 1.4$ were used for the system that included the AGC. In some experiments the AGC was not used, by omitting Eq. (4) and substituting $x^{agc}_{s,bp}$ in Eq. (5) and (7) with $x_{s,bp}$. Without the AGC, the mapping parameters were $\alpha = 16$, $\beta = 0.42$ and $\gamma = 1.24$. For a full account of the blurredness and mapping procedures see [3].

## 3. RESULTS

Table 1 shows the results of the study. For comparison a subset of the results from [3] are shown, namely our replication of Kingsbury's MLP-HMM, MSG+PLP system [4] and our previous missing data system (MD-04). In the table, MD-A and MD-NA denote the missing data system with and without the AGC stage, respectively. The label "comb" means that the recogniser was trained with a combination of spectral features and MFCCs. During testing, these feature streams were tested separately ("MD comb" or "MFCC comb") or together ("MD+MFCC comb"). The label "alone" means that the system was trained and tested using either spectral missing data or MFCC features alone.

Substantial improvements in performance were obtained using the combined MD and MFCC approach. Compared to the MD-04 system, the proposed combinatorial systems perform better in all test conditions, with the largest improvements in the most reverberant test cases. Our system substantially outperforms Kingsbury's approach in the most reverberant test cases, whereas performance in the least reverberant case (e.g. RIR1) still remains marginally poorer. The combined MD+MFCC recognizers always perform better than those which used only MD. The performance of the MFCC systems is substantially below that of the MD and combined

| RIR | T60 (s) | Dist S/R (m) | MLP-HMM MSG+PLP | MD-04 | MFCC alone | MD-A alone | MFCC comb | MD-A comb | MD-A+ MFCC comb | MD-NA alone | MD-NA comb | MD-NA+ MFCC comb |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| clean | | | 98.5 | 97.0 | 98.6 | 96.8 | 98.2 | 96.8 | 98.6 | 97.1 | 97.4 | 98.7 |
| RIR1 | 0.7 | 2.35 | 95.1 | 93.1 | 88.4 | 92.6 | 83.3 | 92.6 | 94.9 | 93.2 | 93.4 | 95.3 |
| RIR2 | 0.7 | 3.05 | 93.5 | 92.4 | 85.7 | 91.5 | 81.7 | 91.7 | 94.6 | 92.5 | 92.9 | 94.9 |
| RIR3 | 1.2 | 2.0 | 71.5 | 78.4 | 41.8 | 83.4 | 43.7 | 83.8 | 84.9 | 81.9 | 81.3 | 84.3 |
| RIR4 | 1.2 | 3.05 | 69.5 | 76.6 | 40.9 | 78.1 | 43.7 | 78.5 | 80.3 | 76.9 | 76.9 | 79.2 |
| RIR5 | 1.5 | 6.1 | 64.0 | 67.8 | 48.0 | 74.4 | 46.7 | 74.3 | 76.4 | 73.8 | 73.5 | 75.7 |
| RIR6 | 1.5 | 18.3 | 59.8 | 64.3 | 42.5 | 68.3 | 41.2 | 68.4 | 70.1 | 66.5 | 67.1 | 69.2 |

**Table 1**. Recognition results (accuracy %). Columns two and three show the T60 reverberation time and source/receiver distance for the reverberation conditions used. MLP-HMM/MSG+PLP data from Kingsbury [4], MD-04 data from Palomäki et al. [3]. The remaining columns show various recogniser configurations from the new study; see Sect. 3 for details.

MD+MFCC systems, except for the clean test case. The AGC technique adds a small performance gain in most of the reverberant test conditions. However, the performance without the AGC is better in the clean test case, and in the least reverberant conditions (RIR1 and RIR2).

## 4. GENERAL DISCUSSION

We have studied ASR in reverberation using an approach that combines likelihoods from a spectral missing data classifier and a conventional Gaussian mixture model classifier using full MFCC vectors. Some improvements in the missing data mask generation have also been reported. Compared to our previous system, performance gains were achieved in all test conditions. The combinatorial system performed better than Kingsbury's approach in most of the reverberant test cases, although Kingsbury's system is still better for the least reverberant case (e.g. RIR1).

Our simulations show that supplementing a spectral MD classifier with a MFCC-based classifier yields performance gains across a range of reverberation conditions, even though the performance of the MFCC system was substantially below that of the MD system. This suggests that the supplementary features contain information that is lost in missing data processing. Similar observations have been made by Kingsbury [4], although his hybrid HMM-MLP speech recognition system differs substantially from ours.

Future work will investigate the use of the combined probability approach for dealing with additive noise. Whereas binary missing data masks have been used here, a further possibility is that the approach could be applied to missing data decoders which use real-valued ('fuzzy') masks [8].

**Acknowledgements**. KJP was funded by the EC IHP HOARSE project, and was partially supported by a Finnish Jenny and Antti Wihurin rahasto grant. GJB and JPB were funded by EPSRC grant GR/R47400/01. Some of the RIRs were recorded by Jim West, Michael Gatlin and Carlos Avendano; thanks to Dan Ellis and Brian Kingsbury for making them available to us.

## 5. REFERENCES

[1] M.P. Cooke, P. Green, and M. Crawford, "Handling missing data in speech recognition," in *Int. Conf. Spoken Lang. Proc.*, 1994, pp. 1555–1558.

[2] M.P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Comm.*, vol. 34, pp. 267–285, 2001.

[3] K. J. Palomäki, G.J. Brown, and J. Barker, "Techniques for handling convolutional distortion with 'missing data' automatic speech recognition," *Speech Comm.*, vol. 43, no. 1–2, pp. 123–142, 2004, www.cis.hut.fi/kpalomak/errata04.pdf.

[4] B.E.D. Kingsbury, *Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments*, Ph.D. thesis, University of California, Berkeley, 1998.

[5] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Trans. Speech Audio Proc.*, vol. 2, pp. 578–589, 1994.

[6] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," in *Proc. Int. Conf. Spoken Lang. Proc.*, 1996, pp. 422–425.

[7] D. Pearce and H. G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems," in *Proc. Int. Conf. Spoken Lang. Proc.*, 2000, vol. 4, pp. 29–32.

[8] J. Barker, L. Josifovski, M. P. Cooke, and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in *Proc. ICSLP-2000*, 2000, vol. I, pp. 373–376.