A Missing Data Approach for Robust Automatic Speech Recognition in the Presence of Reverberation

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

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

Abstract

We describe a technique for robust recognition of reverberated speech using the 'missing data' paradigm. Modulation filtering is used to identify time-frequency regions of the speech signal which are relatively uncontaminated by reverberation and contain strong speech energy; only these 'reliable' acoustic features are made directly available to the recogniser. The proposed system is evaluated on a connected digit recognition task using a range of reverberation conditions. Our approach improves recognition performance when the T_{60} reverberation time is longer than 0.7 sec., relative to a baseline system which uses acoustic features derived from perceptual linear prediction and the modulation filtered spectrogram.

1. Introduction

Much progress has been made in the field of automatic speech recognition (ASR) in recent years, but significant problems still remain; in particular, the performance of ASR systems is far below that of human listeners when speech is presented in noisy or reverberant conditions (see [7] for a review).

Cooke *et al*. [1] note that human speech perception is robust even when speech is band limited or partially masked by noise. Accordingly, they propose a *missing data* approach to ASR, in which a hidden Markov model (HMM) classifier is adapted to deal with acoustic features which are known to be missing or unreliable. However, the missing data approach was conceived as a way of handling *additive* noise in ASR; as a result, little consideration has been given to its ability to handle *convolutional* interference, such as reverberation. In this paper, we propose a number of modifications to a missing data ASR system which allow it to perform robustly in the presence of reverberation.

A typical room impulse response consists of two components. Initially, sparse early reflections occur which are highly correlated with the speech signal. These may spectrally distort the speech, because the absorptive properties of room surfaces tend to vary with frequency. Following this, higher-order reflections produce dense late reverberation, which is poorly correlated with the speech signal and therefore behaves more like additive noise. The speech spectrum is also shaped by the eigenmodes of the room, which emphasize some frequencies in preference to others. Hence, the missing data approach can be applied in reverberant conditions as follows; we use conventional missing data techniques to handle late reverberation (since it resembles additive noise) and employ spectral normalisation to deal with the distortion caused by early reflections and eigenmodes of the room.

Conventional approaches to robust ASR in the presence of reverberation either perform dereverberation using multiple microphones or employ robust acoustic features. Such features include mel-frequency cepstral coefficients (MFCC) with cepstral mean subtraction [2], cepstral coefficients obtained by perceptual linear prediction (PLP) [3], and modulation spectrogram (MSG) features [5], [6]. The latter have proven to be particularly effective.

A schematic diagram of our proposed system is shown in Fig. 1. In the remainder of this paper, we review the missing data approach to ASR in Section 2 and describe a system for reverberation processing in Section 3. Our approach is evaluated in Section 4 using a number of reverberant conditions, and is compared against a system which uses MSG and PLP features [5], [6]. The paper concludes with a discussion in Section 5.

2. Speech recognition with missing features

2.1. Acoustic features

The missing data approach to ASR requires that regions of the time-frequency plane are labelled as reliable or unreliable evidence of the speech source. Accordingly, the recogniser used here employs spectral features derived from an auditory model, rather than conventional features for ASR such as cepstral coefficients.

Here, spectral features are derived from a model of cochlear frequency analysis, consisting of an array of 32 bandpass 'gammatone' filters. The centre frequencies of the filters were spaced uniformly between 50 Hz and 3850 Hz on the equivalent rectangular bandwidth (ERB) scale (see [1]). The instantaneous Hilbert envelope is

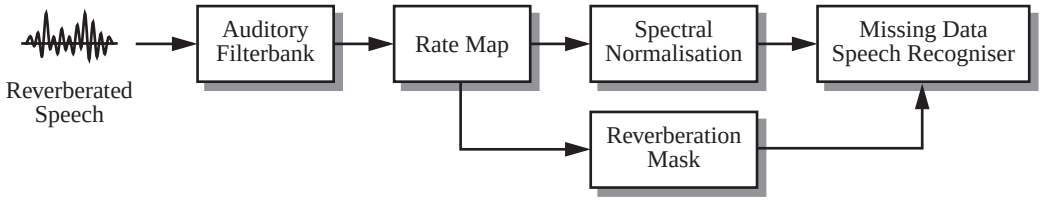

Figure 1: Schematic diagram of the speech recognition system.

computed at the output of each filter, and smoothed by a first-order lowpass filter $H(z) = 1/(1+az^{-1})$, with a chosen to give a time constant of 8 ms. The smoothed envelope is sampled at 10 ms intervals and compressed by raising to the power 0.3 to give a *rate map*, which may be regarded as a crude simulation of auditory nerve firing rate. Henceforth, we denote the rate map by $y(i,j)$ where i indexes the time frame and j is the frequency channel.

2.2. Missing data speech recognition

In ASR, classification is usually performed by finding a class of speech sound C which maximises $f(Y|C)f(C)$, where Y is an observed acoustic vector. However, the likelihood $f(Y|C)$ cannot be computed if some elements of Y are known to be missing or unreliable. In the missing data approach, this problem is addressed by partitioning Y into reliable and unreliable components, Y_R and Y_U. The reliable components are directly available to the recogniser via the marginal distribution $f(Y_R|C)$. Furthermore, the unreliable components are often known to lie within certain bounds; this additional constraint is exploited by integrating over the range of possible values. This technique is known as 'bounded marginalisation' [1]. Here, Y is a vector of simulated auditory nerve firing rates: hence the lower bound of Y_U is zero and the upper bound is the observed firing rate.

In practice, a time-frequency mask $m(i,j)$ is used to indicate whether the spectral feature j at time frame i is reliable. Here, mask values are taken to be 0 or 1, so that a binary judgment is made as to whether the acoustic evidence is reliable or unreliable.

3. Reverberation processing

3.1. Reverberation mask estimation

For reverberant conditions, we estimate a missing data mask in which reliable elements correspond to features that contain strong speech energy, and are relatively unaffected by reverberation. Regions containing strong speech energy are identified by a finite impulse response (FIR) modulation filter of the form

$$h(n) = h_{lp}(n) \otimes h_{diff}(n) \tag{1}$$

where h_{lp} is a linear phase lowpass component and h_{diff} is a differentiator. The symbol $\otimes$ denotes convolution and the time index n is measured in frames. The lowpass filter was designed using the MATLAB fir2 command [8], and is intended to detect and smooth modulations in the speech range. The differentiator emphasizes abrupt onsets, which are likely to correspond to direct sound and early reflections. Overall, $h(n)$ has a passband between 3dB cutoff frequencies of 2 Hz and 13 Hz, which is in agreement with the range of modulation frequencies known to be important for speech perception [4].

The modulation filter is applied to each channel j of the rate map $y(i,j)$, giving a filtered rate map $y_m(i,j)$:

$$y_m(i,j) = \sum_{k=-\infty}^{\infty} h(k)y(i-k,j) \tag{2}$$

Following this, a threshold is applied to the modulation-filtered rate map in order to produce a binary mask:

$$m(i,j) = \begin{cases} 1 \text{ if } y_m(i,j) > \theta(j) \\ 0 \text{ otherwise} \end{cases} \tag{3}$$

Note that filtering by $h(n)$ introduces a delay, which causes the mask $m(i,j)$ and rate map $y(i,j)$ to become misaligned. We compensate for this by shifting the mask backwards in time by an amount corresponding to the delay at which $h(n)$ reaches its peak value.

The value of the threshold $\theta(j)$ should depend upon the degree to which the speech is reverberated. In our previous study [9], $\theta(j)$ was hand-tuned to different reverberation conditions; however, we now estimate it directly from an utterance. Specifically, $\theta(j)$ is set according to a 'blurredness' metric, which exploits the fact that reverberation tends to smooth the rate map by filling the gaps between speech activity with energy originating from reflections. Blurredness B is given by

$$B = \sum_{j=1}^{N} \left\{ \frac{\frac{1}{M}\sum_{i=1}^{M} y(i,j)}{\max_i[y(i,j)]} \right\} \tag{4}$$

where $N = 32$ is the number of frequency channels and M is the number of time frames in the rate map. In practice, it is desirable for the threshold to depend not only on B, but also on the mean value over time in each channel of the modulation-filtered rate map. Hence, we compute an average firing rate $e(j)$ for each channel j according to:

$$e(j) = \frac{1}{M}\sum_{i=0}^{M} \{y_m(i,j) - \min_i[y_m(i,j)]\} \tag{5}$$

Note that the minimum value in each channel is subtracted in order to remove any negative values arising from application of the modulation filter in Eqn. (2).

Finally, the threshold $\theta(j)$ is set according to a sigmoidal function of $e(j)$ and B,

$$\theta(j) = e(j) \cdot \frac{\lambda}{1 + \exp[-\gamma(B - \delta)]} \quad (6)$$

where $\gamma = 16$ is the slope, $\delta = 0.42$ is the centre point and $\lambda = 1.3$ determines the width of the sigmoid. Note that the sigmoidal form of Eqn. (6) allows $\theta(j)$ to saturate at high blurredness values (i.e., long reverberation times).

3.2. Spectral normalisation

In order to compensate for convolutional distortion, spectral features are usually normalised by the mean and variance in each frequency band (for example, see [6]). A potential problem with this approach is that clean regions of an utterance may be normalised by a mean and variance that are computed when both speech and noise are present. This is very problematic for missing data ASR, because reliable features presented to the recogniser must be scaled in the same way as the clean speech features used for training.

Here, we derive a normalisation factor from the L largest reliable features in each frequency channel. Scaling based on these regions should minimise the mismatch between (clean) training and (reverberated) testing conditions, because the corresponding acoustic features are likely to be relatively uncorrupted by reverberation. Specifically, we compute a scaling factor $\eta(j)$ as follows,

$$\eta(j) = \frac{1}{L} \sum_{i \in \Gamma(j)} y_c(i,j) \quad (7a)$$

$$y_c(i,j) = m(i,j) \cdot y(i,j) \quad (7b)$$

where $\Gamma(j)$ is a set containing the indices of the L largest values of $y_c(i,j)$ in channel j. During training and recognition, rate maps are normalised by dividing each channel j by $\eta(j)$. Note that in the training case, $m(i,j) = 1$ for all i and j, i.e. $y_c(i,j) = y(i,j)$.

Generally, we set L to M/D, where M is the number of time frames in the rate map and D is a constant whose value is tuned empirically (we use D=5). However, if channel j does not contain any speech-dominated features, i.e. when $\Gamma(j) = \varnothing$, the scaling factor $\eta(j)$ is interpolated from adjacent channels (or extrapolated in the case of the lowest and highest frequency channels).

4. Evaluation

4.1. Missing data recogniser

We evaluated the proposed missing data ASR system on a subset of the Aurora 2 connected digit corpus [10]. Rate maps and their first-order deltas were computed for the clean training section of the Aurora corpus, and were used to train 12 word-level HMMs (a silence model, 'oh', 'zero' and '1' to '9'), each consisting of 16 no-skip, straight-through states with observations modelled by a seven component diagonal Gaussian mixture. The test set consisted of 1001 utterances drawn from the clean1 test set of the Aurora corpus. All speech data was sampled at a rate of 8 kHz.

4.2. Missing data recogniser with *a priori* masks

An upper bound on the performance of the missing data approach can be obtained by considering its performance when given *a priori* information about the reliability of acoustic features. Specifically, we define an *a priori* mask as follows:

$$m_{apriori} = \begin{cases} 1 \text{ if } 20\log_{10}\left[\frac{y_r(i,j)}{y_d(i,j)}\right] < \Theta \\ 0 \text{ otherwise} \end{cases} \quad (8)$$

Here, $y_r(i,j)$ and $y_d(i,j)$ represent the rate maps for the reverberated and dry (unreverberated) utterances respectively: note that these are computed without the compression described in Section 2.1. The threshold Θ was tuned to give optimal performance for each reverberation condition.

4.3. Baseline HMM-MLP system

For comparison, we also re-implemented Kingsbury's hybrid HMM-MLP (hidden Markov model multi-layer perceptron) recogniser and evaluated it on the Aurora corpus (see [5], pages 148-152). His system combines likelihood estimates from two kinds of reverberation-robust features; cepstral coefficients (plus their deltas and double deltas) obtained by PLP, together with MSG features.

Kingsbury's system was implemented using the STRUT speech recognition toolkit [11]. Acoustic models for 23 phonemes, silence and unknown (required by STRUT) were obtained from the training part of the Aurora corpus. Durational information was included in the HMM model for each phone by setting the number of states to be proportional to the average duration of the phone, computed from the training set (see page 45 of [5] for details).

4.4. Results

Test utterances were reverberated by convolving them with six different room impulse responses. Four of these were used by Kingsbury [5], and were recorded in a varechoic chamber with two different configurations of the wall panels. In one configuration the T_{60} reverberation time was 0.7 sec. and the distances between the source and microphone were 2.35 m and 3.05 m. In the second configuration, the T_{60} was 1.2 sec. and the

T_{60} and Source-Receiver Distance	HMM-MLP	MD-AP	MD-RM
1.5 sec., 18.3 m	59.8	88.5	63.2
1.5 sec., 6.1 m	64.0	92.4	67.6
1.2 sec., 3.05 m	69.5	88.5	75.6
1.2 sec., 2.0 m	71.5	89.9	77.6
0.7 sec., 3.05 m	93.5	94.2	91.9
0.7 sec., 2.35 m	95.1	95.0	93.0
Unreverberated	98.5	97.5	97.0

Table 1: Recognition accuracy (percent) for seven reverberation conditions. Results are shown for the baseline system (HMM-MLP), missing data recogniser with *a priori* masks (MD-AP) and missing data with the proposed reverberation masking scheme (MD-RM).

source-microphone distances were 2.0 m and 3.05 m. A further two impulse responses (not used by Kingsbury) were recorded in a larger room, with a T_{60} of 1.5 sec. and source-microphone distances of 6.1 m and 18.3 m.

The results shown in Table 1 indicate that the proposed missing data system outperforms Kingsbury's hybrid recogniser in the most reverberant test cases. However, the performance of the hybrid recogniser using MSG and PLP features was better than that of the missing data system for the shortest T_{60} condition, and when no reverberation was present. We also note that the missing data approach is very robust when reliable regions are known *a priori*. Further improvements to the mask estimation process could yield recognition performance that approaches this theoretical upper limit.

5. Discussion and conclusions

The reverberation masking technique proposed here has some parallels with the modulation spectrogram (MSG) [5], since both exploit modulation frequencies in the speech range. However, we believe that our approach has some advantages. Robust acoustic features such as MSG represent a compromise; they are intended to work in a wide variety of acoustic environments, but in any particular environment their performance may not be optimal. In contrast, our algorithm can be adapted quickly to different acoustic environments by changing the mask estimation rule. It may therefore offer advantages for ASR in mobile devices.

In practice, the baseline HMM-MLP system outperformed our missing data system in the least reverberated conditions. This may be because our method of estimating the amount of reverberation present in a speech sample is not sufficiently sensitive to distinguish between anechoic and mildly reverberant conditions. In the most reverberant cases, however, the missing data approach has a clear advantage over the baseline system.

Furthermore, our experiments with *a priori* masks suggest that there is considerable potential for further development of the missing data technique. Our future work will focus on improving the mask estimation process, with the expectation that this will yield a performance closer to that obtained with *a priori* masks.

6. Acknowledgements

GJB and JB were funded by EPSRC grant GR/R47400/01. KJP was funded by the EC TMR SPHEAR project, the Academy of Finland (project 1277811) and a Finnish Nokia säätiö grant. Thanks to Dan Ellis, Brian Kingsbury and Heidi Christensen for their help with the baseline system, and to Dan, Brian, Jim West, Michael Gatlin and Carlos Avendano for providing the room responses.

7. References

[1] Cooke, M. P., Green, P. D., Josifovski, L., Vizinho, A. "Robust automatic speech recognition with missing and unreliable acoustic data", *Speech Comm.*, 34:267-285, 2001.

[2] Davis, S. P., Mermelstein, P. "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Trans. Acoust. Speech Signal Process.*, ASSP-28:357-366, 1980.

[3] Hermansky, H. "Perceptual linear predictive (PLP) analysis of speech", *J. Acoust. Soc. Am.*, 87:1738-1752, 1990.

[4] Houtgast, T., Steeneken, H. J. M. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria", *J. Acoust. Soc. Am.*, 77:1069-1077, 1985.

[5] Kingsbury, B. E. D. "Perceptually inspired signal-processing strategies for robust speech recognition in reverberant environments", PhD thesis, Univ. California, Berkeley, 1998.

[6] Kingsbury, B. E. D., Morgan, N., Greenberg, S. "Robust speech recognition using the modulation spectrogram", *Speech Comm.*, 25:117-132, 1998.

[7] Lippmann, R. P. "Speech recognition by machines and humans", *Speech Comm.*, 22:1-15, 1997.

[8] Mathworks, Inc. MATLAB release 13 reference manual. Natick, MA, 2003.

[9] Palomäki, K. J., Brown, G. J., Barker, J. "Missing data speech recognition in reverberant conditions", *Proc. ICASSP-2002*, I:65-68, 2002.

[10] Pearce, D., Hirsch, H. G. "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions", *Proc. ICSLP-2000*, 4:29-32, 2000.

[11] STRUT Version 2.4, 1997. Step by step guide to using the speech training and recognition unified tool, http://www.tcts.fpms.ac.be/asr/project/strut/.