MISSING DATA SPEECH RECOGNITION IN REVERBERANT CONDITIONS

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

[1]Department of Computer Science, University of Sheffield,
211 Portobello Street, Sheffield S1 4DP, United Kingdom.

[2]Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing
P.O. Box 3000, FIN-02015 HUT, Finland

Email: kalle.palomaki@hut.fi, g.brown@dcs.shef.ac.uk, j.barker@dcs.shef.ac.uk

ABSTRACT

In this study we describe an auditory processing front-end for missing data speech recognition, which is robust in the presence of reverberation. The model attempts to identify time-frequency regions that are not badly contaminated by reverberation and have strong speech energy. This is achieved by applying reverberation masking. Subsequently, reliable time-frequency regions are passed to a 'missing data' speech recogniser for classification. We demonstrate that the model improves recognition performance in three different virtual rooms where reverberation time T60 varies from 0.7 sec to 2.7 sec. We also discuss the advantages of our approach over RASTA and modulation filtered spectrograms.

1. INTRODUCTION

Human listeners have little difficulty in recognising speech in moderately reverberant conditions, whereas reverberation substantially degrades the performance of current automatic speech recognition (ASR) systems. It is reasonable to argue, therefore, that ASR performance in the presence of reverberation could be improved by adopting an approach that models auditory processing more closely.

Room reverberation introduces convolutional interference that can be characterised as both spectral distortion and additive noise. Spectral shaping of the speech signal arises from room modes that emphasize some frequencies more than the others. Room reflections can be divided into early reverberation, which is highly correlated with the speech signal, and late reverberation which is less correlated. Therefore the interference caused by late reverberation can be characterised as additive noise.

Broadly, four strategies have been proposed to handle reverberation (for an overview see [9]): training speech models in the presence of reverberation, dereverberation, source separation via microphone array processing and the search for more robust feature vectors for the recogniser. The dereverberation approach attempts to estimate a model of the room impulse response and tries to remove it by deconvolution. However, blind deconvolution from single microphone input has remained a difficult problem. Probably the best results so far have been achieved with microphone array processing, but the problem with this technique is that at least two microphone signals are needed when one speech signal is separated. An alternative approach is to seek noise robust feature vectors. An interesting feature of this approach is that better performance in reverberant or noisy conditions is usually achieved when the system mimics functions of the human auditory system [5],[6],[7]. A good example of this is RASTA-PLP (RelAtive SpecTrAl Perceptual Linear Predictive analysis) [5], which mimics several aspects of auditory processing rather closely. Another approach, especially optimised for reverberation, is the modulation filtered spectrogram (MSG) representation [7].

However, knowledge about human auditory processing can inform ASR beyond the feature extraction stage. Human listeners are able to perceive speech robustly even when parts of the signal are masked by noise or deleted by band-limiting. According to Cooke and his co-workers [4], this implies that the auditory system has a mechanism for dealing with 'missing data'. They have exploited this notion in ASR by adapting a hidden Markov model (HMM) classifier to deal with missing or unreliable features. The missing data paradigm is complementary to a 'computational auditory scene analysis' (CASA) approach; an auditory model can be used to decide which acoustic components belong to a target speech source, and only these 'reliable' features are passed to the recogniser. Indeed, auditory front-ends have been combined with missing data speech recognition systems in several previous studies [2],[3],[10].

In this study we propose a new method using an auditory front-end, which enhances recognition performance in the precence of reverberation. Our model is based on a reverberation masking algorithm which attempts to find spectro-temporal regions which are not severely contaminated by reverberation and discards those which are. The model is evaluated in three different virtual rooms, in which the reverbation time T60 varies from 0.7 sec to 2.7 sec. The results obtained with the new method are compared against a baseline recogniser which uses a mean normalised mel-cepstral coefficient (MFCC) front-end.

2. MODEL

2.1. Monaural pathway

To produce the feature vectors for the recogniser a simple monaural model of the auditory pathway is used (see Fig 1). Cochlear frequency analysis is simulated by a bank of 32 bandpass gammatone filters with centre frequencies spaced on the equivalent rectangular bandwidth (ERB) scale between 50 Hz and 8 kHz. The output of each filter is half-wave rectified and compressed to give a representation of auditory nerve activity. Then the instantaneous Hilbert envelope is computed at the output of each filter. This is smoothed by a first-order low-pass filter with an 8 ms time constant, sampled at 10 ms intervals, and finally cube

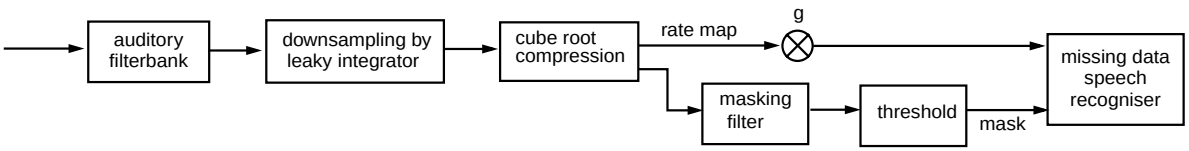

Figure 1: Schematic diagram of the model.

root compressed to give a representation of auditory nerve firing rate ('rate map'; see Figure 3 for an example).

Because reverberation introduces level changes which degrade recogniser performance, a gain adjustment g was applied to the rate maps. We use g=1 for the non-reverberant case, g=0.738 for all the reverberant cases.

2.2. Missing data speech recogniser

In this study an HMM speech recogniser is adapted to exploit the missing data technique [4]. Automatic speech recognition is a classification problem in which an acoustic observation vector x must be assigned to a class of speech sound C. However, when noise is present some components of x may be unreliable or missing. In these cases, the likelihood $f(x|C)$ cannot be computed in the usual manner. The 'missing data' technique addresses this problem by partitioning x into reliable and unreliable components, x_r and x_u. The reliable components x_r are directly available to the classifier. In practice, a binary 'mask' $m(i,j)$ is used to indicate whether the acoustic evidence in each time-frequency region is reliable.

In the simplest approach, the components of the unreliable part x_u are simply ignored so that classification is based on the marginal distribution $f(x_r|C)$. However, when x is an acoustic vector additional constraints can be exploited, since it is known that uncertain components will have bounded values (the 'bounded marginalisation' method [4]). In this study, x is an estimate of auditory nerve firing rate, so the lower bound for x_u is zero and the upper bound is the observed firing rate. We also use first order temporal derivatives and word insertion penalties, which are known to improve the performance of the missing data approach [3].

2.3. Mask generation heuristics

In this study we use two different mask estimation heuristics. Firstly, we produce a mask exploiting *a priori* information by measuring the difference d between the clean signal x and its reverberation contaminated counterpart $\hat{x}$. Then the mask values m are set as follows:

$$m(i,j) = \begin{cases} 1 & d(i,j) < \theta_{ap} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where i denotes frequency channel, j is the current time instant, θ_{ap} is a threshold, 1=reliable and 0=unreliable. The purpose of (1) is to test the limits of the missing data approach by producing 'ideal' masks, and to test how close to this limit we can reach with mask estimation based on *a posteriori* information only.

The second heuristic attempts to detect spectro-temporal regions that contain strong speech energy and are not badly contaminated by the reverberation. For this purpose we define a reverberation masking filter,

$$H(z) = \sum_{k=0}^{n} \sin(k\omega) \cdot z^{-k} \quad (2)$$

where the filter coefficients were computed from one period of a sinusoid. This corresponds to a band-pass filter that has lowpass characteristics with an additional zero at DC. Two filters were used, one optimised for the shortest reverberation time and another for the two longest: their frequency responses are plotted in Fig. 2.

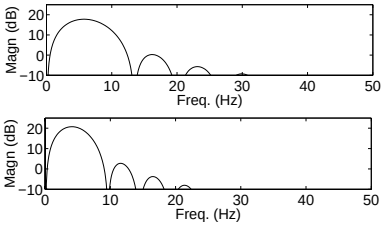

Figure 2: Frequency response of the filters used for reverberation times T60 = 0.7 sec, centre frequency 6.7 Hz (top) and for T60 = 1.7, 2.7 sec, centre frequency 4.8 Hz (bottom).

Mask values are computed by thresholding the filtered reverberation contaminated signal $y = \hat{x} * h$, where h is the impulse response corresponding to (2). Hence the mask was described by

$$m(i,j) = \begin{cases} 1 & \text{if } y(i,j) > \theta_b \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

Longer reverberation times caused the rate maps to be blurred, and therefore a longer integration period (resulting a narrower pass band) was necessary. To compensate for the filter delay, the masks were shifted backwards in time by the corresponding delay. Thresholds were experimentally tuned to give optimal result for each reverberation conditions.

During the experiments, several different types of filter design were investigated and a smooth sinusoidal impulse response was found to give the best results. The key issue in the filter design was firstly to have a long enough integration phase followed by a differentiation phase. This made it possible to detect the least reverberated areas that usually coincided with the strongest modulation frequencies of speech. Figure 3 shows that our model detects these areas rather reliably. As shown in the figure, this gives a mask estimate with wide clean areas that correspond to the *a priori* values well.

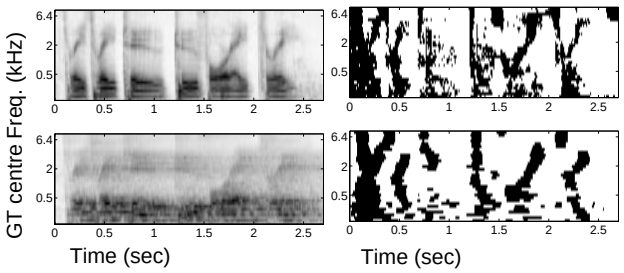

Figure 3: Left panel: Rate maps computed for anechoic (top) and reverberant (T60=1.7 D/R=-10) conditions (bottom). Right panel: *a priori* mask (top) and mask based on reverberation masking (bottom). Black areas in the mask (right panel) correspond to reliable speech regions, white areas correspond to reverberation contaminated regions.

3. EVALUATION

3.1. Corpus & HMM settings

The model was evaluated on a 240 utterance subset of male speakers from the TiDigits connected digits corpus [8]. Auditory rate maps and MFCCs were obtained for the training section of the corpus, and were used to train 12 word-level HMMs (a silence model, 'oh', 'zero' and '1' to '9') each consisting of 8 no-skip, straight-through states with observations modeled by a 10 component diagonal Gaussian mixture. All models were trained on unreverberated signals. The test utterances were then convolved with an artificially generated room impulse response. All of the utterances were presumed to start from silence. For a baseline result we used mean normalised mel-cepstral coefficients (MFCC) with 13 cepstral coefficients, and 1st and 2nd order temporal derivatives.

3.2. Testing the model under reverberation

The reverberation of an enclosure is often characterized using a simple measure called reverberation time T60, which is defined as the time required for the reverberation to drop 60 decibels below the original sound level. For example, the recommended T60 for a speech hall is 0.4 sec. A richer acoustic environment is required for music, and therefore the T60 should be longer - a typical value for a concert hall is 2.0 sec. Another useful measure considered here is the ratio of direct sound to reverberated sound (D/R). In practice D/R alters when the distance between a measurement point and the source changes, or alternatively when the direct sound pathway becomes blocked or attenuated.

In this study the room impulse responses were artificially generated by producing early reflections using the image model. The basic principle of the image model is that reflection paths from a sound source to a listener are found by reflecting the sound source against all surfaces of the room [1]. Late reverberation was taken from a real room impulse response having a high density of reflections. This was windowed to give a realistic exponential decay for different reverberation times used in our experiments. The use of the image model to generate early reflections allowed us to easily configure the room model to give different test conditions. Similarly, using a real room response for the late reverberation model made it possible to produce a reverberation tail that has a high density of reflections and a realistic distribution of energy across frequencies. Related models of room reverberation have previously been employed in speech recognition by other workers [7].

For testing the model, we generated an early reflection pattern with the image model for three different rectangular spaces (R1=15x13x6.5m^3, R2=25x20x6m^3, R3=55x35x14m^3) and fitted the exponential decay of late reverberation to these. The resulting reverberation times were T60 = 0.7, 1.7 and 2.7 sec respectively. All spaces were tested with a D/R of 0 dB and -10 dB. A -10 dB ratio was obtained with 6.5, 8 and 14 m distances between the speaker and the recogniser in the three different rooms R1, R2 and R3 respectively. For 0 dB D/R the direct sound was scaled by 10 dB without altering the pattern of reverberation**.**

3.3. Results

Table 1 shows the recognition performance in three different rooms for two D/R. The performance of the missing data speech recogniser with different mask estimation techniques is compared against a MFCC front-end. Different approaches as they appear in the table are (i) unity mask, i.e. all features are passed to the recogniser - this corresponds to a conventional HMM recogniser without missing data processing (ii) MFCC-based speech recogniser (iii) missing data mask estimation using reverberation masking (iv) using a mask based on *a priori* knowledge of the clean regions of the signal. The clean (unreverberated) utterances were tested only with the unity mask and MFCC recogniser, but not for the two missing data techniques. This is based on the

Recognition Technique	Clean	T60=0.7s D/R=0dB	T60=0.7s D/R=-10db	T60=1.7s D/R=0dB	T60=1.7s D/R=-10dB	T60=2.7s D/R=0dB	T60=2.7s D/R=-10dB
Unity mask	98.26	62.48	46.39	42.82	29.24	34.90	24.28
MFCC	99.65	60.40	47.08	47.35	34.46	40.73	28.55
Reverb. masking		90.33	85.03	82.25	71.11	66.05	45.52
a priori mask		95.82	93.73	92.42	89.12	90.60	87.99

Table 1: Speech recognition accuracy (100-WER, word error rate) in three different virtual rooms with two different D/R-ratios. Recognition performance of missing data recognition with different mask estimation methods is compared against a baseline obtained with a MFCC front-end.

assumption that for the missing data approach, optimal performance on clean speech will be obtained with a unity mask.

In all the different experiments, the missing data approach with reverberation masking substantially outperforms the MFCC and unity mask approaches. The *a priori* results suggest a ceiling performance that may be achieved using the current missing data approach. When the T60 became longer the performance difference between *a priori* and reverberation masking techniques increased. This result is not suprising, since the increased reverberation causes more blur in the rate maps - thus it is more difficult to produce accurate masks based on *a posteriori* information only.

All of the utterances started from silence, and hence in the most reverberant environments the reverberation more greatly affected the last digits of the utterance. This was apparent from detailed examination of our recognition results; approaches that didn't use reverberation masking performed relatively poorly on later digits compared to the first digit.

4. DISCUSSION

In this paper we have proposed a new approach to speech recognition in reverberation, which uses a reverberation masking technique as a front-end for a missing data speech recogniser. The results in Table 1 demonstrate that our model is relatively robust in the presence of reverberation. Using our approach over 85% recognition accuracy was achieved for a 0.7 T60 reverberation time, which is already 0.3 sec longer than the 0.4 sec recommended for speech hall design. Also, with longer reverberation times rather competitive recognition results were achieved when compared to previous research. However, human performance still clearly exceeds the results presented here.

The model proposed here has some parallels to earlier work on RASTA-PLP and MSG that are used for producing noise robust feature vectors. Both of these techniques have a processing chain that firstly divides the signal into frequency bands and then (after downsampling and compression) applies a band-pass filter to emphasise the most noise tolerant speech signal regions. RASTA-PLP and MSG (also together with PLP) have both been applied to robust speech recognition in reverberation, with the latter approach perhaps being the most successful. We note, however, that MSG requires parameter adjustment for optimal performance when acoustic conditions change radically [6]. Since the same front-end settings must be used during both recognition (of reverberated speech) and training (typically on clean speech), retraining may be necessary in order to adapt a MSG-based recogniser to different reverberation conditions.

A general problem with noise robust feature vector approaches is that their performance on clean speech is sometimes compromised. Furthermore, using noise robust feature vectors sometimes gives worse performance when the type of noise is changed. In contrast, our 'missing data' recogniser can be adapted to different noise conditions simply by changing the mask estimation rule, thus avoiding the need for retraining. Hence, different types of front-end can be easily 'switched in' (for example, changing from a front-end that is optimal for additive noise to a front end that is robust for reverberation). Performing such modifications in an adaptive manner is a challenging problem that we will address in future work. In this study all thresholds were experimentally tuned for each case. Adaptive selection of these thresholds is also an issue for future research.

Kingsbury and his co-workers [6],[7] suggest that the MSG technique enhances ASR performance because the modulation frequencies remaining after their filtering approach are those which are known to be most important in human speech recognition. A similar explanation also underlies our study. Equally, though, our approach can be regarded as one whose aim is to identify the least reverberation contaminated regions of the signal that usually coincide with the strongest speech modulations. It is evident from Figure 3 that the model is able to find such regions rather reliably.

Our system showed a significant improvement compared to the baseline results obtained using a MFCC-based recogniser. Clearly, however, this baseline system does not constitute a state-of-the-art recogniser of reverberant speech. Our future intention is to compare our system against MSG and combined MSG-PLP front-ends which are known to give good performance in reverberation.

Acknowledgement. The project was funded by the EC TMR SPHEAR project and partially supported by Finnish Tekniikan edistämissäätiö grant.

5. REFERENCES

[1] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Am.,* 65, pp. 943-950, 1979.

[2] J. Barker, M. P. Cooke and D. P. W. Ellis, "Decoding speech in the presence of other sound sources," *Proc. ICSLP'00*, IV, pp. 270-273, 2000.

[3] J. Barker, M. P. Cooke, L. Josifovski and P. D. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," *Proc. ICSLP'00*, I, pp. 373-376, 2000.

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

[5] H. Hermansky and N. Morgan, "RASTA Processing of Speech," *IEEE Trans. Speech and Audio Proc*. 2(4) pp. 578-589, 1994.

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

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

[8] R. G. Leonard, "A database for speaker-independent digit recognition," *Proc. ICASSP'84*, pp. 111-114, 1984.

[9] M. Omologo, P. Svaizer and M. Matassoni, "Environmental conditions and acoustic transduction in hands-free speech recognition," *Speech Comm.*, 25, pp. 75-95, 1998.

[10]K. J. Palomäki, G. J. Brown and D. L. Wang, "A binaural auditory model for missing data speech recognition in noisy and reverberant conditions," *proc. CRAC-Eurospeech'01 satellite workshop*, 2001.