IDIAP RESEARCH REPORT

Improving Speech Recognition Performance of Small Microphone Arrays using Missing Data Techniques

Iain A. McCowan [1] Andrew Morris [1]
Hervé Bourlard [1, 2]

IDIAP–RR 02-09

April 2002

SUBMITTED FOR PUBLICATION

Dalle Molle Institute for Perceptual Artificial Intelligence • P.O.Box 592 • Martigny • Valais • Switzerland

phone +41 – 27 – 721 77 11
fax +41 – 27 – 721 77 12
e-mail secretariat@idiap.ch
internet http://www.idiap.ch

1 IDIAP—Dalle Molle Institute of Perceptual Artificial Intelligence, P. O. Box 592, CH-1920 Martigny, Switzerland, {mccowan, morris, bourlard}@idiap.ch
2 EPFL, Lausanne

IDIAP Research Report 02-09

Improving Speech Recognition Performance of Small Microphone Arrays using Missing Data Techniques

Iain A. McCowan   Andrew Morris   Hervé Bourlard

April 2002

submitted for publication

Abstract. Traditional microphone array speech recognition systems simply recognise the enhanced output of the array. As the level of signal enhancement depends on the number of microphones, such systems do not achieve acceptable speech recognition performance for arrays having only a few microphones. For small microphone arrays, we instead propose using the enhanced output to estimate a reliability mask, which is then used in missing data speech recognition. In missing data speech recognition, the decoded sequence depends on the reliability of each input feature. This reliability is usually based on the signal to noise ratio in each frequency band. In this paper, we use the energy difference between the noisy input and the enhanced output of a small microphone array to determine the frequency band reliability. Recognition experiments with a small array demonstrate the effectiveness of the technique, compared to both traditional microphone array enhancement and a baseline missing data system.

1 Introduction

Microphone arrays have been shown to give significant improvements over single distant microphones in speech recognition applications, particularly in high levels of noise [1]. While traditional array enhancement techniques can lead to good performance, the hardware and processing requirements of large numbers of microphones can inhibit their application in practice. For this reason, it is desirable to investigate ways of reducing the number of microphones without significantly affecting system performance.

A promising avenue of research towards this goal is the closer integration of traditional microphone array processing techniques with other robust speech recognition techniques. The recent *missing data* approach to robust speech recognition is based upon the concept that only reliable information should be taken into account when recognising noise-corrupted speech [2]. When the feature space consists of energies from frequency bands, one means of determining the reliability of the features is the estimation of the signal to noise ratio (SNR) in each band. The set of reliability estimates (for each feature, for each frame) for missing data recognition is termed a *reliability mask*.

Missing data experiments using *a priori* knowledge of the noise spectrum have demonstrated the potential of using an SNR-based reliability mask, with performance in high levels of noise approaching that of clean speech. In practice however, estimation of the local SNR is problematic with a single channel signal. This is particularly true when the noise is non-stationary, as single channel techniques typically rely upon estimating the noise during non-speech periods.

In the proposed technique, we compare the enhanced output of a small microphone array with the noisy input in order to estimate the reliability of each frequency band. The technique allows the generation of robust reliability masks using only small microphone arrays. Moreover, this technique is equally applicable in stationary or non-stationary noise conditions.

This paper is organised as follows. Section 2 gives a brief overview of the missing data approach to robust speech recognition, and details the standard procedure for calculating an SNR-based missing data mask. Section 3 then details the proposed technique for generating a missing data mask using a microphone array. In Section 4, the performance of the proposed robust speech recognition system is investigated using a small microphone array in varying levels of office noise. In these experiments, the new technique is shown to give improvement over both a standard microphone array enhancement system, as well as a baseline single channel missing data system.

2 Missing Data Speech Recognition

2.1 Overview

In automatic speech recognition (ASR), some of the observation data for the target signal can be masked by noise signals. One way to compensate for noise in ASR is to train the recogniser on data which has been corrupted by multiple different noise conditions. A limitation of this approach is that it is not as effective when the test noise conditions differ significantly from those used in training.

Another approach which avoids this restriction is to use models trained on clean speech together with a noise estimation technique. Noise estimation can be used to improve recognition performance in two ways. One is to enhance the speech data and then process it as clean speech. Another is the *missing data* (MD) approach, in which spectral data components which are estimated to be dominated by noise are ignored [2]. This approach was partly motivated by studies on human auditory scene analysis [3]. Perceptual experiments have shown that humans are able to maintain a high level of recognition when a high proportion of spectral data is masked by noise. Two notable improvements to the basic missing data technique have been proposed which significantly improve performance at low SNR's.

The first improvement was to make use of the *bounds constraint* whereby the unknown clean speech energy (at each point in frequency and time) is known to be between zero and the observed noisy energy [2]. In terms of computational expense, the only difference in a standard HMM system using Gaussian mixture PDFs is that while for clean components it is necessary to evaluate the Gaussian density, for missing components it is necessary to evaluate the Gaussian integral over the range of possible clean data values (which is possible using the standard error function).

The second improvement was the *soft missing data* technique, in which the decision as to whether each component is clean or noisy is modelled by a soft (probabilistic) rather than hard (deterministic) decision [4]. In this case the missing data mask specifies a probability that each component of the feature vector is clean or dominated by noise. Here both the density and the integral must be evaluated for every component. These are then combined in a weighted sum, where the weights for the density/integral are the probabilities that the feature is speech dominated and noise dominated, respectively.

The missing data approach to robust speech recognition essentially poses two separate problems : missing data mask estimation, and how to perform recognition using this mask. The above improvements both address the latter problem, while in this paper we focus on the problem of effective estimation of the missing data mask.

2.2 Missing Data Mask Estimation

The key element of the missing data approach is the missing data mask, which indicates the reliability $r_i(k)$ of each input feature $x_i(k)$ - for filter-bank index i, at each frame k. This reliability value must lie between 0 and 1, and represents the 'probability' that the feature is clean enough to be used for reliable recognition.

Provided (log) filter-bank speech features are used, an estimate of the frequency band noise energy can be used to determine reliable features by comparing to an SNR threshold value β (in the natural-log energy domain). If the input energy in a frequency band exceeds the estimated noise energy by more than this threshold, it can be assumed to be reliable for speech recognition. Assuming we have an estimate of the noise log energy feature $n_i(k)$, then we can estimate the clean speech log energy feature (transforming to and from the linear amplitude spectral domain) as

$$s_i(k) = \log\left(\left(\sqrt{e^{x_i(k)}} - \sqrt{e^{n_i(k)}}\right)^2\right) \tag{1}$$

and then the ratio of signal to noise energy in the feature is

$$snr_i(k) = s_i(k) - n_i(k) \tag{2}$$

where $snr_i(k)$ is the estimated SNR for the feature $x_i(k)$. The reliability of the feature is then determined as

$$r_i(k) = \begin{cases} 1 & : \quad snr_i(k) \geq \beta \\ 0 & : \quad snr_i(k) < \beta \end{cases} \tag{3}$$

In the soft missing data approach, this value is softened to take a value in the range 0 to 1 according to

$$r_i(k) = \frac{1}{1 + e^{-\alpha(snr_i(k)-\beta)}} \tag{4}$$

where α is the gradient of the smoothing function [4].

To date, most missing data experiments have used a simple noise estimation technique which averages the spectrum over the first K input frames (assumed to be noise only). While this gives reasonable performance in stationary noise for the purpose of experimentation, it is desirable to investigate more robust and practical noise estimation techniques.

3 Reliability Mask Estimation using Microphone Arrays

The success of the missing data recognition approach depends critically on the calculation of the reliability mask. Experiments using *a priori* knowledge of the noise spectrum show that performance

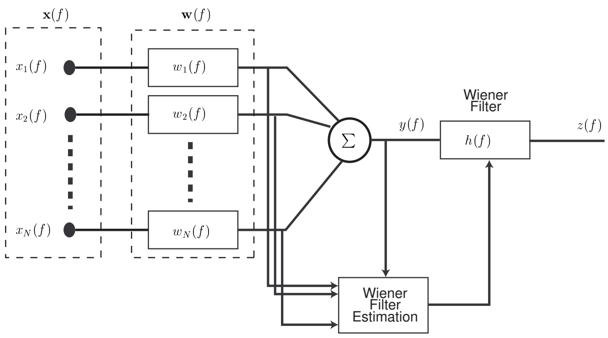

Figure 1: Filter-sum beamformer with post-filter

approaching that of clean speech can be achieved in high levels of corrupting noise. It is thus desirable to approach this performance level by investigating more robust methods for obtaining the reliability mask. In this section we propose such a technique using a small microphone array.

3.1 Microphone Array Speech Enhancement

Microphone arrays use directional discrimination to permit distant, hands-free signal acquisition, allowing for speech enhancement by reducing the level of undesired noise sources. The directional discrimination of the array is exploited by beamforming algorithms, and often the beamformer output is further enhanced by applying a post-filter. Using such techniques, significant enhancement of the speech signal can be achieved, in turn resulting in improvements in speech recognition performance [1].

Consider the microphone array system shown in Figure 1. The system consists of two stages : a filter-sum beamformer with output $y(f)$, and a Wiener post-filter with final output $z(f)$. The operation of the system is described by the equations :

$$y(f) = \mathbf{w}(f)^T \mathbf{x}(f) \tag{5}$$

$$z(f) = h(f)y(f) \tag{6}$$

where $\mathbf{w}(f)$ is the beamforming weight vector, $\mathbf{x}(f)$ is the input data vector, and $h(f)$ is the Wiener post-filter transfer function.

A variety of techniques exist for calculating the beamforming filters $\mathbf{w}(f)$. In this paper we use the simplest beamforming technique, known as delay-sum beamforming, in which the beamforming filters consist purely of delays to align for the direction of the desired signal. While more advanced beamforming techniques exist, the delay-sum technique is sufficient to demonstrate the potential of the proposed technique.

The post-filter transfer function, $h(f)$, is calculated using a combination of the auto- and cross-spectral densities of the input channels, according to a technique first proposed by Zelinski [5]. This technique was further developed and analysed in [6] and [7], and was shown to lead to improved speech recognition performance in [8].

3.2 Proposed SNR Mask Estimation Technique

In the proposed microphone array SNR mask estimation technique, we simply use the enhanced output of the microphone array, $z(f)$, as an estimate of the clean signal. In the log filter-bank energy feature space we have both the noisy input features, $x_i(k)$ and the estimated clean speech features $z_i(k)$. Rather than simply using these estimated clean speech (enhanced) features for recognition, as is the

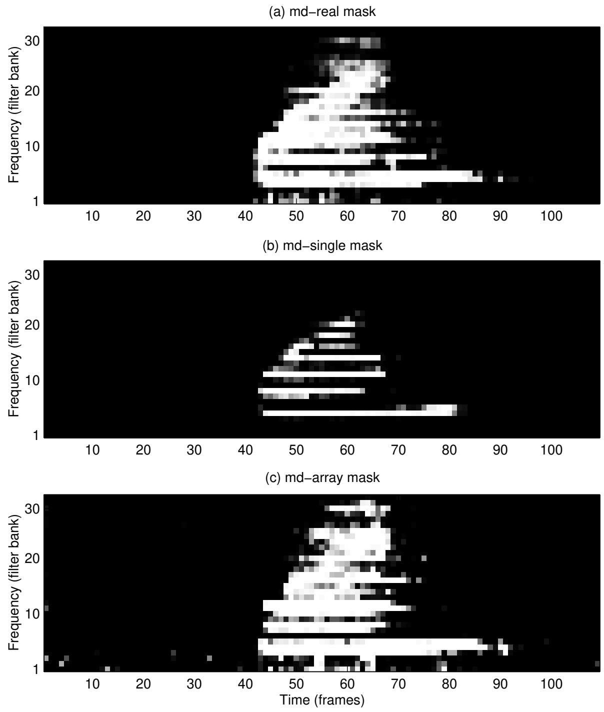

Figure 2: Sample SNR Masks (input SNR=0dB, utterance is "one")

case for standard microphone array speech recognition systems, we instead propose using them to determine the missing data reliability mask. In this case, the signal to noise ratio can be estimated as

$$n_i(k) = \log\left(\left(\sqrt{e^{x_i(k)}} - \sqrt{e^{z_i(k)}}\right)^2\right) \tag{7}$$

$$snr_i(k) = z_i(k) - n_i(k) \tag{8}$$

and then the soft reliability mask can be calculated according to Equation 4. As the speech and noise energy estimates are both time-dependent, such a technique is applicable in both stationary and non-stationary noise conditions.

The above approach is based upon the concept that, while the level of enhancement in $z(f)$ for small microphone arrays is often not sufficient for good speech recognition performance in itself, examination of the level of energy reduction between each feature of $x_i(k)$ and $z_i(k)$ may still give a good indication of the feature reliability. By using the array enhancement only to estimate the noisy feature reliability, rather than the feature value itself, the level of enhancement required of the array is considerably relaxed, and hence acceptable recognition performance should still be achieved with few array elements.

4 Experiments and Results

4.1 Configuration

The microphone array used in experiments consisted of 4 linearly-spaced microphones, with inter-element spacing of 5cm. The desired speaker was situated directly in front of the broadside array, at a

	WER (%)		
signal	20 dB	10 dB	0 dB
noisy	61.4	85.8	91.3
array	9.5	22.6	64.3
md-single	13.3	20.2	54.1
md-array	9.7	12.1	31.5
md-true	8.0	9.1	16.6

Table 1: Speech recognition results

distance of 70cm. For the microphone array enhancement technique, standard delay-sum beamforming was used, and the output was further enhanced using the post-filtering technique in [6].

The experiments were conducted in an office room containing a number of computer workstations. The room has a measured reverberation time of approximately RT60 $\approx 420ms$. Multi-channel recordings of the room noise were made, consisting mainly of computer noise, air-conditioning noise, and a low level of background speech. We note that this noise recording is essentially stationary in nature. In addition, the impulse responses of the acoustic path between the desired speaker and each microphone were calculated from real recordings. These impulse responses were used to generate the multi-channel desired speech input for the test database.

Speaker and gender independent speech recognition experiments were conducted on 1001 utterances from the TIDIGITS database (Aurora test set 1a). HMM's were trained on the clean TIDIGITS training data set, with no adaptation performed for the noise conditions or for the effect of the distant microphone. Standard log mel-scaled filter-bank energy features (order 32) were used along with their first order derivatives, giving a feature vector of dimension 64. The system gives an error rate of 5.6% for the clean input to one of the microphones in the array.

Five different techniques were tested :

1. standard recognition using noisy input features from a single (distant) microphone ('noisy')
2. standard recognition using the enhanced array output features ('array')
3. missing data recognition on the noisy features, with SNR mask generated from a single channel by estimating the noise spectrum over the first 10 input frames ('md-single')
4. missing data recognition on the noisy features, with SNR mask generated from the microphone array output ('md-array')
5. missing data recognition on the noisy features, with SNR mask generated from *a priori* knowledge of the clean and noisy signals ('md-true')

For all missing data techniques, soft masks were used according to Equation 4, with an SNR threshold of $\beta = 0$ and a softening gradient of $\alpha = 5$. An example of the three different masks generated at an input SNR of 0dB is shown in Figure 2 (single digit utterance "one"). In the figure, black represents noise (reliability of zero), and white represents clean speech. These masks show the success of the microphone array technique in generating a mask that closely resembles the real SNR mask.

The various techniques were tested at different input SNR's of 20, 10 and 0 dB. The speech recognition results are given in Table 1, and plotted in Figure 3 in terms of the percentage word error rate.

4.2 Discussion

We note two interesting trends from these results. First, it is apparent that the proposed technique for generating the missing data reliability mask using the microphone array (md-array) is better than

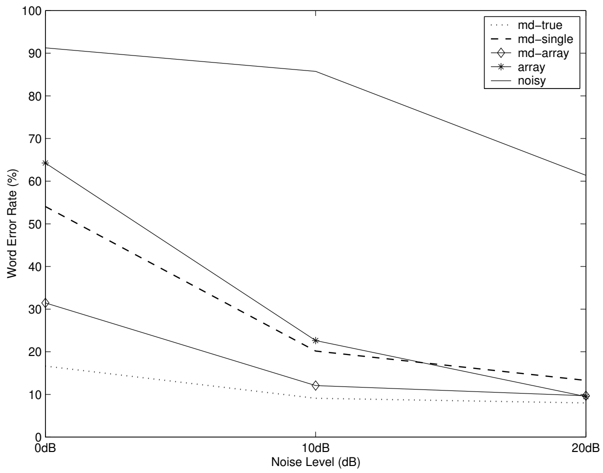

Figure 3: Speech recognition results

the single channel mask estimation technique (md-single), approaching the performance of the true SNR mask (md-true). While this is not surprising, it is pertinent to note that the noise is essentially stationary in nature, and thus the single channel technique should give a reasonable noise estimate. The difference between the two approaches should be more marked in non-stationary noise conditions, but unfortunately experiments confirming this have not yet been conducted due to the lack of appropriate multi-channel noise recordings.

The second important trend we observe, is that the performance of the proposed technique (md-array) yields significant improvements over the enhanced array output (array) for such a small microphone array. This is particularly evident in high noise, where the word error rate drops from 64% to 32% at 0dB SNR. This result is of interest as the same information (enhanced signal $z(f)$) is being used in each case, highlighting the benefit of using the information only for reliability estimation instead of purely for enhancement.

As a final point, we note that there exists room for some further improvement by tuning of the α and β parameters of the soft mask calculation.

5 Conclusions

In this paper we have investigated improving the speech recognition performance of a small microphone array system. Instead of simply recognising the enhanced array output, the difference in energy levels between the noisy input and the enhanced output is used to indicate the reliability of each feature in a missing data recognition framework. In simple experiments using an array of 4 microphones, the proposed technique is shown to give improved recognition results when compared to both the standard microphone array output and also a baseline single channel missing data technique. While the results are promising, we note that the experiments in this paper have been limited to a single array configuration, a single noise type (stationary office noise), and a single speaker location. Future work will investigate the effect of different noise types, array configurations and speaker locations on the system performance.

References

[1] M. Omologo, M. Matassoni, and P. Svaizer. Speech recognition with microphone arrays. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 15, pages 331–353. Springer,

2001.

[2] A. Morris, M. Cooke, and P. Green. Some solutions to the missing feature problem in data classification, with application to noise robust ASR. In *Proceedings of ICASSP '98*, pages 737–740, 1998.

[3] P. Green, M. Cooke, and M. Crawford. Auditory scene analysis and HMM recognition of speech in noise. In *Proceedings of ICASSP '95*, pages 401–404, 1995.

[4] J. Barker, L. Josifovski, M. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In *Proceedings of ICSLP 2000*, pages 373–376, 2000.

[5] R. Zelinski. A microphone array with adaptive post-filtering for noise reduction in reverberant rooms. In *Proceedings of ICASSP-88*, volume 5, pages 2578 –2581, 1988.

[6] Claude Marro, Yannick Mahieux, and K. Uwe Simmer. Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering. *IEEE Transactions on Speech and Audio Processing*, 6(3):240–259, May 1998.

[7] I. McCowan and H. Bourlard. Microphone array post-filter for diffuse noise field. In *Proceedings of ICASSP 2002*, 2002.

[8] I. McCowan, C. Marro, and L. Mauuary. Robust speech recognition using near-field superdirective beamforming with post-filtering. In *Proceedings of ICASSP 2000*, volume 3, pages 1723–1726, 2000.