EURASIP Journal on Applied Signal Processing 2001:1, 45–52

Multi-Channel Sub-Band Speech Recognition

Iain A. McCowan

Speech Research Laboratory, RCSAVT, School of EESE, Queensland University of Technology,
GPO Box 2434, Brisbane QLD 4001, Australia
Email: iain@ieee.org

Sridha Sridharan

Speech Research Laboratory, RCSAVT, School of EESE, Queensland University of Technology,
GPO Box 2434, Brisbane QLD 4001, Australia
Email: s.sridharan@qut.edu.au

Received 22 December 2000 and in revised form 16 February 2001

Two distinct fields of research into robust speech recognition are the use of microphone arrays for signal enhancement and the use of independent frequency sub-band models for robust recognition. In this article, we propose and investigate the integration of these two techniques on two different levels. First, a broad-band beamforming microphone array allows for natural integration with sub-band speech recognition as the beamformer is implemented as a combination of band-limited sub-arrays. Rather than recombining the sub-array outputs to give a single enhanced output, we fuse the output of separate hidden Markov models trained on each sub-array frequency band. Second, a dynamic sub-band weighting algorithm is proposed in which the cross- and auto-spectral densities of the microphone inputs are used to estimate the reliability of each frequency band. The proposed multi-channel sub-band system is evaluated on an isolated digit recognition task and compared to both a standard full-band microphone array system and a single channel sub-band system.

Keywords and phrases: microphone array, sub-band, beamforming, speech recognition.

1. INTRODUCTION

An emerging area of research is the use of microphone arrays for the purpose of speech enhancement. In particular, microphone arrays have shown much promise in improving the performance of hands-free speech recognition systems in adverse environments [1, 2]. While such microphone array systems have shown good performance, potential for further improvement exists in closer integration of the multi-channel input with the speech recognition system. Brandstein [3] observes that while single channel speech enhancement and robust recognition techniques have sought to exploit various features of the speech signal, multi-channel techniques to date have primarily focused on improving the spatial filtering process. He suggests that some of the current limitations of the field could be addressed by researching multi-channel techniques based upon explicit modeling of speech characteristics.

In this article, we investigate the integration of a sub-band based speech recognition system with a microphone array. Sub-band speech recognition is a relatively new field of research which has been shown to improve robustness to noise where frequency bands are corrupted in a nonuniform manner [4, 5]. The sub-band approach is motivated by the psychoacoustic evidence that auditory processing decisions in humans are formed from the combination of independently processed frequency sub-bands [6, 7].

The proposed system integrates the microphone array with sub-band speech recognition in two ways. First, spatial filtering is done on the input channels to enhance the input to each sub-band recognizer. As the spacing of microphone array elements is dependent on the frequency of interest, a common technique of covering the broad frequency range of speech is to implement the beamformer using band-limited sub-arrays, each having elements spaced appropriately for a different frequency sub-band. Rather than recombining these sub-array outputs and performing speech recognition on the single full-band signal, we propose independent recognition of the sub-array outputs followed by likelihood combination using the sub-band recognition approach. This should show improved performance over both single channel sub-band recognition and microphone array full-band recognition by combining the advantages of both, namely the noise reduction provided by the microphone array and the noise robustness provided by the sub-band recognition system.

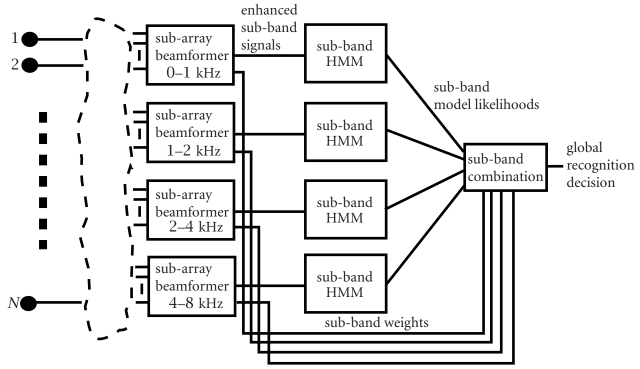

1
2
N
sub-array beamformer 0–1 kHz
sub-array beamformer 1–2 kHz
sub-array beamformer 2–4 kHz
sub-array beamformer 4–8 kHz
enhanced sub-band signals
sub-band HMM
sub-band HMM
sub-band HMM
sub-band HMM
sub-band model likelihoods
sub-band combination
global recognition decision
sub-band weights

Figure 1: System block diagram.

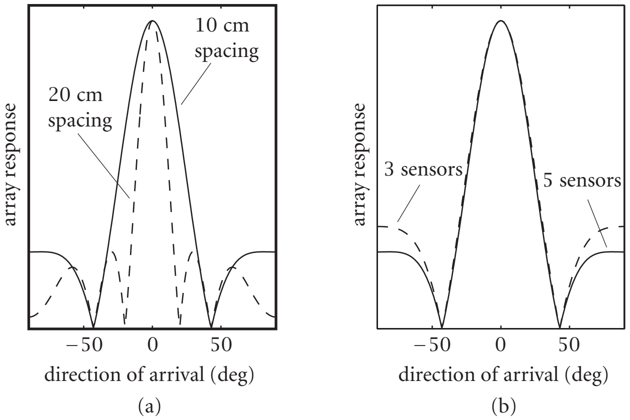


Figure 2: Array response comparison. (a) Same number of sensors, different array length. (b) Different number of sensors, same array length.

The second proposed level of integration is a multichannel algorithm to determine the weights to apply to each sub-band recognition result in forming the global decision. The best method of performing this recombination is currently an open issue with sub-band recognition. The reliability of each sub-band result depends to some extent upon the proportion of speech and noise energy present in that frequency band. With the multi-channel input from the microphone array, an effective estimate of the sub-band noise levels can be made by examining the cross- and auto-spectral densities of the different channels. The proposed algorithm uses such a multi-channel noise estimation technique to determine the reliability of each sub-band on a word by word basis.

A block diagram of the proposed system is shown in Figure 1. The system can be broken down into three main components: the sub-array beamformer, the sub-band speech recognition system and the calculation of the sub-band weights. Each of these components is discussed in detail in the following sections.

The proposed multi-channel sub-band recognition system is compared to a standard full-band microphone array recognition system, and a single channel sub-band recognition system in isolated digit speech recognition experiments. The results of the proposed dynamic weighting scheme are compared to those obtained using both fixed equal sub-band weights, as well as optimal sub-band weights calculated from a priori knowledge of the correct results.

2. SUB-ARRAY BEAMFORMING

The response of an array of sensors approximates that of the continuous aperture which it samples. A linear array of N sensors with uniform inter-element spacing, d, has a horizontal directivity pattern given by

$$D(\phi) = \sum_{n=1}^{N} a_n e^{-(j2\pi f(n-1)d\sin\phi)/c}, \tag{1}$$

where a_n is the gain associated with the nth sensor, ϕ is the angle measured normal to the array axis, and c is the speed of propagation. From this equation we see that the characteristics of the array response depend on the frequency of interest, the inter-element spacing, and the number of elements in the array. For a given number of elements, the dependency on element spacing is effectively a dependency on the length of the continuous aperture that is being sampled ($L = Nd$). Figure 2 demonstrates how, for a given frequency, the array response depends upon the length of the array and the number of sensors. As is seen from Figure 2(a), for the same number of elements, the array length determines the main lobe width of the response—the longer the array, the narrower the main lobe. Specifically, the beam-width is inversely proportional to the product fL, where L is the array length. Conversely, as shown in Figure 2(b), varying the number of elements for a given array length has the effect of changing the sidelobe level—the more sensors, the lower the sidelobes.

The dependency on the operating frequency means that the response characteristics (beam-width, sidelobe level) will only remain constant for narrow-band signals, where the bandwidth is not a significant proportion of the centre frequency. Speech, however, is a broad-band signal, meaning that a single linear array design is inadequate if a frequency invariant beam-pattern is desired. One popular and simple method of covering broadband signals is to implement the array as a series of sub-arrays, which are themselves linear arrays with uniform spacing. These sub-arrays are designed to give desired response characteristics for a given frequency range. Due to the dependencies discussed above, as the frequency increases, a smaller array length is required to maintain constant beam-width. In addition, to ensure the sidelobe level remains the same for different frequency bands, the number of elements in each sub-array should remain the same. The sub-arrays are generally implemented in a nested fashion, such that any given sensor may be used in more than one sub-array. Each sub-array is restricted to a different frequency range by applying band-pass filters, and the overall broad-band array output is formed by recombining the outputs of the band-limited sub-arrays. To illustrate the concept, an example of such a nested sub-array structure designed to cover 3 different frequency bands and employing simple

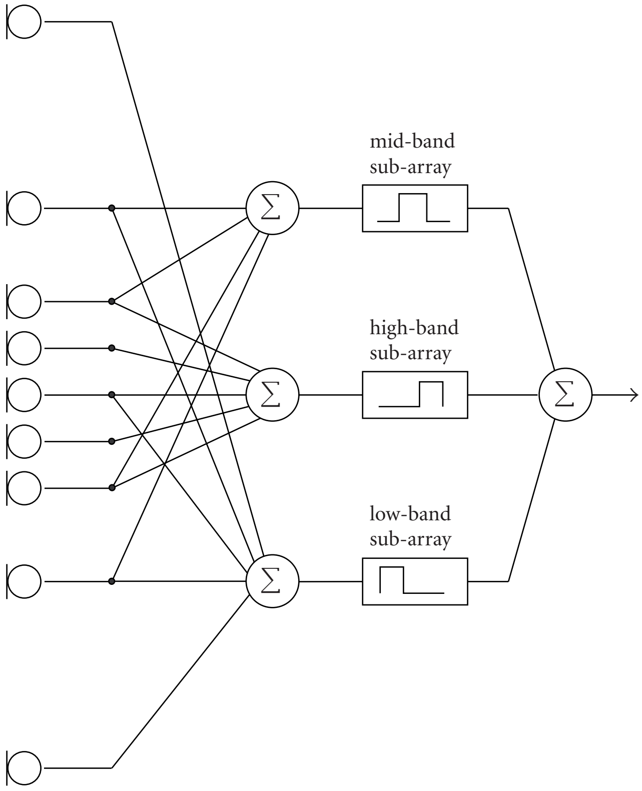


Figure 3: Sample nested sub-array structure.

delay-sum beamforming, is shown in Figure 3. In this case, each sub-array employs 5 microphones, but due to the nested structure the 3 sub-arrays can be implemented using a total of 9 microphones.

Beamforming techniques are algorithms that can be applied to the input signals of a sensor array in order to steer the main lobe of the directivity pattern to a desired direction, and also to add further enhancement to the directional characteristics of the array. A variety of beamforming techniques exist, most of which involve applying filters to each input channel prior to combination. Different beamforming algorithms calculate these channel filters for different design criteria, and so the choice of beamforming algorithm is governed by the particular application.

For a general sub-array broadband beamformer, the beamforming channel filters are band-pass filtered between the specified upper and lower frequencies for each sub-band. At the output of each channel filter we have

$$v_i^s(f) = b_i^s(f)x_i(f), \tag{2}$$

where $x_i(f)$ is the input to channel i of the array, and the superscript s represents the sub-array index. The output of sub-array s, is then given by the normalized sum across channels as

$$y^s(f) = \frac{1}{\sum_{i=1}^{N} b_i^s(f)} \sum_{i=1}^{N} v_i^s(f), \tag{3}$$

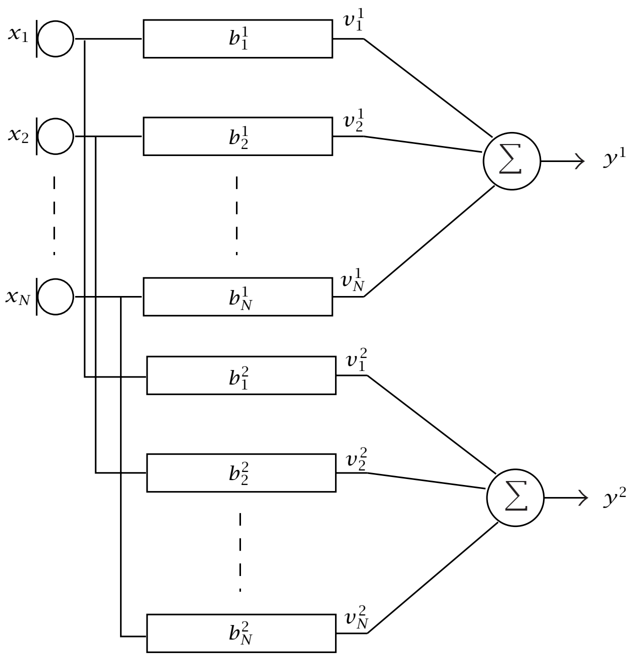


Figure 4: Sub-band beamformer structure.

where there are N microphones in the array. The summation in each sub-band is shown up to N for simplicity of notation, although in practice only the channels belonging to each sub-band are used. The beamformer structure for 2 sub-bands is shown in Figure 4.

One beamforming technique which has been shown to give good performance in speech recognition applications is *superdirectivity* [2, 8]. Superdirective techniques aim to calculate channel filters that maximize the array gain, which is defined as the improvement in signal to noise ratio between the array inputs and output. A near-field modification to the superdirective technique, termed near-field superdirectivity, was proposed by Täger [9] for the case where the desired speech source is located close to the array. Previous work has demonstrated the suitability of near-field superdirectivity for speech recognition in the context of a computer workstation in a noisy office [10].

Near-field superdirectivy is an array beamforming technique that succeeds in achieving good noise reduction across all frequencies by compensating for both the phase and amplitude differences in the desired signal across the different sensors. The technique is formulated as an optimization of the array gain in the direction of the desired signal source under the assumption of a diffuse noise field. For the experiments in this paper, near-field superdirective beamforming was performed using the geometry of Figure 5 and the following sub-array configurations:

(1) $f < 1$ kHz: microphones 1–11,

(2) 1 kHz $< f <$ 2 kHz: microphones 1, 2, 5, 8, 9,

(3) 2 kHz $< f <$ 4 kHz: microphones 2, 3, 5, 7, 8,

(4) 4 kHz < f < 8 kHz: microphones 3–7.

All microphones are used in the low frequency range as this is where the amplitude differences exploited by the near-field superdirective technique are most significant. The microphones for the remaining three sub-bands were selected to give uniform response characteristics, with each sub-array containing 5 microphones with inter-element spacings of 10 cm, 5 cm, and 2.5 cm, respectively.

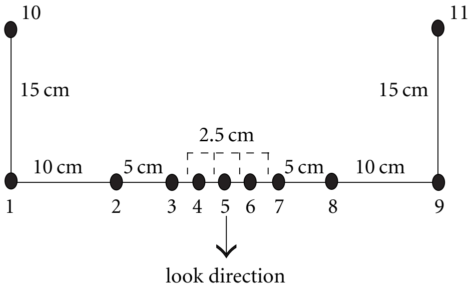


Figure 5: Array geometry.

In the experiments, the sub-array channel filters, $b_i^s(f)$, are calculated using the algorithm detailed by Täger [9], and are band-pass filtered between the specified upper and lower sub-array frequencies for each sub-band.

3. SUB-BAND SPEECH RECOGNITION

Sub-band speech recognition is based upon the work of Fletcher [6] (reviewed by Allen in [7]) which investigated the way in which humans recognize speech. His research found evidence suggesting that humans process speech units in independent articulation bands (or frequency channels), and that the estimates from each of these bands are merged in some optimal fashion to determine the globally recognized speech unit. In humans, the fusion of articulation bands reduces the overall error rate according to the *product of errors rule*, which states that the full-band error rate is equal to the product of the sub-band error rates [6, 7]. This principle has inspired much recent work in so called *sub-band recognition* in an effort to improve the robustness of automatic speech recognition systems [4, 5, 11].

Sub-band speech recognition is effectively a problem in combining classifiers, where each classifier is a HMM trained on speech from a particular frequency sub-band. Classifier combination is used across many diverse fields as a means of improving the accuracy of decision making processes [12]. Rather than relying on a single *expert* to make a decision, a set of experts is employed, where each expert is trained on a different set of features. A consensus decision is reached by combining the opinions of each individual expert according to some combination rule.

We consider the general case when we wish to classify a measurement $\mathbf{x}$ given by

$$\mathbf{x} = \{x^1, \dots, x^S\}, \tag{4}$$

where there are S classifiers and x^s denotes the measurement features for the sth classifier. We wish to assign the measurement to the class λ_m which gives the maximum a posteriori probability. Expressing this decision framework formally

$$\hat{\lambda} = \lambda_m, \quad P(\lambda_m \mid \mathbf{x}) = \max_i \left(P(\lambda_i \mid \mathbf{x})\right). \tag{5}$$

Using the Bayes theorem, the a posteriori probability can be written as

$$P(\lambda_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid \lambda_i) P(\lambda_i)}{p(\mathbf{x})}, \tag{6}$$

where $p(\cdot)$ denotes a probability density function and $P(\lambda_i)$ is the a priori occurrence probability of class i. Since the denominator is class independent and assuming equally probable classes, maximizing the a posteriori probability consists of maximizing the probability density $p(\mathbf{x} \mid \lambda_i)$. Assuming conditional independence between the features for different classifiers, we have

$$p(\mathbf{x} \mid \lambda_i) = \prod_{s=1}^{S} p(x^s \mid \lambda_i). \tag{7}$$

In the framework of a hidden Markov model classifier, the output scores are generally logarithms of the average frame probability densities. In such a context, if $p(x^s \mid \lambda_i)$ represents the average frame probability densities, then (7) can be written as

$$p(\mathbf{x} \mid \lambda_i) = \sum_{s=1}^{S} \log\left(p(x^s \mid \lambda_i)\right). \tag{8}$$

And thus reformulating the decision rule from (5) in terms of the HMM log-likelihood outputs gives us

$$\begin{gathered} \hat{\lambda} = \lambda_m, \\ \sum_{s=1}^{S} \log\left(p(x^s \mid \lambda_m)\right) = \max_i \left(\sum_{s=1}^{S} \log\left(p(x^s \mid \lambda_i)\right) \right). \end{gathered} \tag{9}$$

In the case where the classifier accuracy is disturbed by noise in the measurements, the decision rule can be made more robust by weighting the output of each classifier in the combination by a term α^s as

$$\begin{gathered} \hat{\lambda} = \lambda_m, \\ \sum_{s=1}^{S} \alpha^s \log\left(p(x^s \mid \lambda_m)\right) = \max_i \left(\sum_{s=1}^{S} \alpha^s \log\left(p(x^s \mid \lambda_i)\right) \right), \end{gathered} \tag{10}$$

where the values α^s are positive and are normalized to sum to unity. The classifier weights effectively represent a confidence measure of the relative reliability of that classifier making a correct decision.

In the context of sub-band speech recognition, the above framework can be used to emulate the decision making process observed in humans by Fletcher [6]. By training a classifier for each frequency sub-band, and recombining the classifier outputs according to (10), the recognition system should exhibit greater robustness to errors caused by frequency dependent noise.

For the proposed technique, the sub-band recognition models are implemented as hidden Markov models that are trained and tested using band-pass filtered speech input. A major issue in the training of the sub-band models is the choice of parameterization. Of the parameterization methods examined, sub-band mel frequency cepstral coefficients (MFCC's) were found to give the best results in our experiments. Sub-band MFCC's differ from standard MFCC's in that the frequency banks are only distributed between the specified lower and upper frequency bounds of each band.

4. CALCULATION OF SUB-BAND WEIGHTS

Clearly the success of the sub-band recognition approach is critically reliant on the sub-band weighting factors, α^s. Several techniques to determine these weights have been proposed, with varying degrees of success, including normalized sub-band phoneme-level recognition rates, normalized sub-band signal to noise ratios, and multi-layer perceptrons [4]. In the proposed technique we propose an algorithm that makes use of the multi-channel input to give a continuous estimate of the signal to signal-plus-noise ratio.

4.1. *Dynamic sub-band weighting algorithm*

The reliability of each sub-band recognition result depends upon the amount of speech and noise energy in the given frequency band. Multi-channel techniques provide us with a convenient means of estimating the input signal to noise ratio. If we denote the speech and noise power spectral densities as Φ_{ss} and Φ_{nn}, respectively, under the assumptions that

(1) the noise and speech are uncorrelated,

(2) the noise has low correlation between sensors,

(3) the noise power spectral density is the same across sensors,

we have the following relations for the cross- and auto-spectral densities between input channels

$$\begin{aligned}\Phi_{v_i^s v_i^s}(f) &= \Phi_{ss}(f) + \Phi_{nn}(f),\\ \Phi_{v_i^s v_j^s}(f) &= \Phi_{ss}(f),\end{aligned} \tag{11}$$

where $\Phi_{v_i^s v_j^s}(f)$ and $\Phi_{v_i^s v_i^s}(f)$ are the cross- and auto-spectral densities of the channel-filtered signals v_i^s.

Of course the above assumptions are only true in an ideal scenario, and so in practice an improved estimate of the speech and noise spectral densities can be made by averaging the cross- and auto-spectral densities over all channel combinations. Using this technique, and normalizing for the effect of the channel filters, Marro et al. [13] estimate the ratio

$$W(f) = \frac{\Phi_{ss}(f)}{\Phi_{ss}(f) + \Phi_{nn}(f)}, \tag{12}$$

as

$$\begin{aligned}\hat{W}^s(f) = {}& \frac{\sum_{i=1}^{N} \left| b_i^s(f) \right|^2}{\Re\left\{ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} b_i^s(f) b_j^{s*}(f) \right\}}\\ &\times \frac{\Re\left\{ \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{\Phi}_{v_i^s v_j^s}(f) \right\}}{\sum_{i=1}^{N} \hat{\Phi}_{v_i^s v_i^s}(f)}.\end{aligned} \tag{13}$$

The values $\hat{\Phi}$ are the estimated spectral densities, which are calculated using a simple time recursive formula as

$$\hat{\Phi}^k_{v_i^s v_j^s}(f) = \gamma v_i^s(f) v_j^{s*}(f) + (1-\gamma)\hat{\Phi}^{k-1}_{v_i^s v_j^s}(f), \tag{14}$$

where k is the frame number, $(\cdot)^*$ is the complex conjugate operator and γ is typically in the range $0.7 \leq \gamma \leq 0.95$.

Equation (13) was thoroughly analyzed by Marro et al. [13] as a microphone array post-filter and shown to be effective in a variety of adverse conditions. In the proposed system we use it to estimate the average proportion of speech energy in each sub-band as

$$\beta^s = \frac{1}{f_h^s - f_l^s} \sum_{f=f_l^s}^{f_h^s} \hat{W}^s(f) \tag{15}$$

and then average this across each frame in the word utterance to give $\bar{\beta}^s$. From this we determine the normalized sub-band weights as

$$\alpha^s = \frac{\bar{\beta}^s}{\sum_{i=1}^{S} \bar{\beta}^i}. \tag{16}$$

4.2. *Optimal sub-band weights*

To measure the effectiveness of the above algorithm, it is desirable to somehow compute the maximum bound to the performance that can be obtained using a simple weighted combination of sub-bands. To determine this upper bound we use an iterative minimization algorithm which uses a simple distance measure as its objective function. The global log-likelihood of each word is first calculated as

$$L_m = \sum_{s=1}^{S} \alpha^s \log\left(p\left(\tilde{y}^s \mid \lambda_m^s \right) \right), \tag{17}$$

where $\tilde{y}^s$ represents the sub-band MFCC's for the beamformed output of sub-array s. Given a priori knowledge of the correct word, the distance measure is calculated as the difference in the global log-likelihoods of the correct word and the highest scoring competing word, that is,

$$D = \max_{m \neq c} (L_m) - L_c, \tag{18}$$

where the model c corresponds to the correct word.

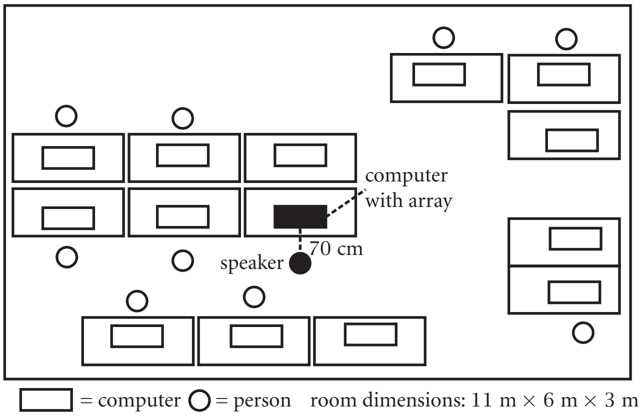


Figure 6: Experimental setup.

5. SPEECH RECOGNITION EXPERIMENTS

To assess the effectiveness of the proposed technique, hands-free speaker independent speech recognition experiments were conducted using the single digit utterances from the male adult portion of the TIDIGITS connected digits database. The recognition models were trained for each sub-band using the clean input to the centre microphone, using sub-band mel-frequency cepstral coefficients. The word recognition rates (WRR) for clean test data are shown in Table 1.

Table 1: Sub-band word recognition rates (clean speech).

sub-band	*WRR*
full-band	99.7%
1	96.5%
2	91.2%
3	85.5%
4	83.9%
combined	98.7%

The experimental context is the computer room shown in Figure 6, which has a measured reverberation time of RT60 = 250 ms. The desired speaker was situated 70 cm from the centre microphone, directly in front of the array. Impulse responses of the acoustic path between the source and each microphone were measured from recordings made in the room with the array. The multi-channel desired speech was generated by convolving the speech signal with these impulse responses.

5.1. Noise condition 1

As the advantages of the sub-band recognition technique will be most pronounced for band-limited noise, a first set of experiments was conducted using white noise that was band-pass filtered to corrupt one whole sub-band for each utterance. The corrupted sub-band was varied uniformly across the database so that all four bands were corrupted an equal number of times. The noise was added for various average segmental signal to noise ratios, calculated only across the frequency range of the corrupted sub-band. The experiments compare the performance of the proposed microphone array sub-band system with both a full-band microphone array system and a single channel sub-band system. The results for different noise levels are given in Table 2 and are plotted in Figure 7. Results are given for the following cases:

Table 2: Word recognition rates: noise condition 1.

technique	*sub-band SNR* (dB)			
	10	5	0	−5
single	63.5%	56.3%	47.3%	36.5%
BF	72.7%	63.3%	55.6%	47.6%
single-SB	89.3%	81.7%	76.4%	71.2%
BF-SB	95.5%	91.7%	86.9%	80.4%
BF-SB-DW	97.8%	96.3%	93.6%	90.0%
BF-SB-OPT	99.0%	99.0%	98.1%	97.2%

- single channel unenhanced (*single*)
- full-band beamformed (BF)
- single channel sub-band (fixed equal weights) (*single-SB*)
- beamformed sub-band (fixed equal weights) (BF-SB)
- beamformed sub-band (dynamic sub-band weighting algorithm) (BF-SB-DW)
- beamformed sub-band (optimal weights) (BF-SB-OPT)

5.2. Noise condition 2

Given that the noise has been band-limited to a single sub-band, the above experimental results represent an ideal scenario for sub-band recognition. In addition, we note that the use of random noise for the different sensors is an ideal case that fulfills the assumptions made for the dynamic weighting algorithm in Section 4.1. Thus, while the above results serve to illustrate the theoretical merit of the proposed technique, it is desirable to verify the system in more realistic noise conditions.

To this end, a second set of experiments was performed using a real multi-channel recording of background noise in the office room. This recording was made simultaneously on all microphone elements in the array. The noise recording consisted of noise from computers and air-conditioning, as

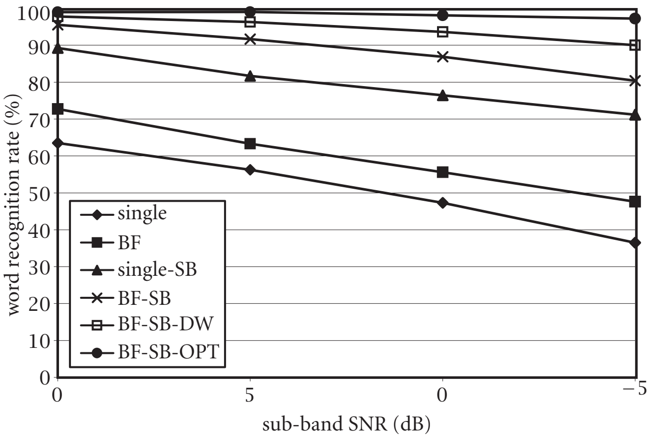


Figure 7: Speech recognition results: noise condition 1.

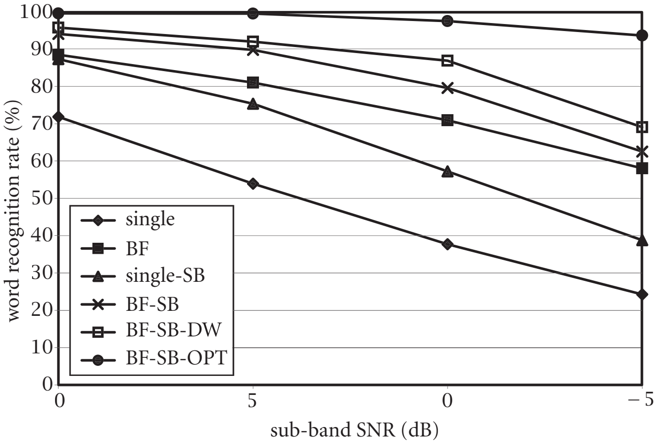


Figure 8: Speech recognition results: noise condition 2.

well as speech-like noise (taken from NOISEX database) emitted from a number of loudspeakers throughout the room. Random segments of the noise recording were added to the multi-channel speech signals at varying segmental signal to noise ratios. Due to the presence of the NOISEX speech-like noise, much of the noise energy was located in the low frequency band below 1 kHz. The results for different noise levels are given in Table 3 and are plotted in Figure 8.

Table 3: Word recognition rates: noise condition 2.

technique	*sub-band SNR (dB)*			
	10	5	0	−5
single	71.8%	53.9%	37.7%	24.3%
BF	88.5%	81.0%	70.9%	58.0%
single-SB	87.3%	75.4%	57.2%	38.8%
BF-SB	94.1%	89.7%	79.6%	62.5%
BF-SB-DW	95.7%	92.0%	86.9%	69.1%
BF-SB-OPT	99.5%	99.5%	97.5%	93.6%

5.3. *Discussion of results*

The results demonstrate several interesting trends. First, the results show the performance improvement obtained by using multi-channel beamforming rather than a single channel system. In both sets of results, the beamformed system (BF) offers improved performance over the standard single channel system (*single*). In fact, while the *single-SB* system performs better than the BF system for the first noise condition (which is ideal for sub-band recognition), the beamformer proves to be more robust in the more realistic noise scenario. In both sets of results the baseline beamformed sub-band system (BF-SB) demonstrates a clear improvement over the standard single channel sub-band system (*single-SB*). The error rate reduction in each individual frequency sub-band provided by the beamformer translates into more significant improvements following the fusion of the sub-band results.

Second, the results demonstrate the effectiveness of the proposed dynamic sub-band weighting algorithm. We can conclude that the proportion of speech energy in each sub-band is a meaningful measure of the sub-band reliability, and that the multi-channel input provides an accurate and robust method for its estimation. While the sub-band system with fixed equal weighting gives good performance, the proposed dynamic weighting algorithm is successful in providing further improvement in the results. In fact, in the first noise condition the proposed system performs at a level comparable to the theoretical upper bound obtained using a priori knowledge of the correct results. The results for the second noise configuration also show that the proposed algorithm is robust to real noise environments, although it is apparent that some room for further improvement exists given the theoretical upper bound represented by BF-SB-OPT.

Once again, it is worthwhile noting that the two different noise scenarios examined in the experiments represent favorable conditions for a sub-band recognition approach. In situations where the noise corrupts all frequency bands uniformly, a sub-band system offers no significant benefits over a standard full-band system.

6. CONCLUSIONS

An integration of microphone array beamforming and sub-band recognition techniques has been proposed. This integration is two-fold. First, the sub-array beamformer provides enhanced inputs to each sub-band recognizer, considerably improving the overall performance by reducing the recognition errors in each sub-band. Second, the cross- and auto-spectral densities of the multi-channel input are used to give a measure of the signal to noise ratio, which is in turn used to calculate the weights to use in the sub-band recognition recombination. Experiments conducted with high levels of band-limited noise show that both levels of integration successfully improve the noise robustness of the recognition performance. In this paper we have examined sub-band recombination at the word level, however the proposed algorithm can be applied at lower levels as the sub-band weights can effectively be calculated on a frame by frame basis.

While clearly successful in the experiments, it is pertinent to note that the proposed system is limited in its application to noise environments which are approximately diffuse and band-limited in nature. A diffuse noise field closely obeys the assumptions made for the dynamic sub-band weighting algorithm in Section 4.1, while the general sub-band speech recognition approach is ideal for the case of band-limited noise.

In summary, the proposed system serves to demonstrate the advantage of fully *integrating* a microphone array with other robust speech recognition techniques, rather than simply using the array as a front-end enhancement module. By taking care to maximize the use of the available multi-channel input, the high levels of performance required for real applications are achievable in adverse conditions.

REFERENCES

[1] K. Kiyohara, Y. Kaneda, S. Takahashi, H. Nomura, and J. Kojima, "A microphone array system for speech recognition," in *Proceedings of ICASSP 97*, April 1997, pp. 215–218.

[2] J. Bitzer, K. U. Simmer, and K. Kammeyer, "Multi-microphone noise reduction techniques for hands-free speech recognition—a comparative study," in *Robust Methods for Speech Recognition in Adverse Conditions (ROBUST-99)*, Tampere, Finland, May 1999, pp. 171–174.

[3] M. Brandstein, "On the use of explicit speech modeling in microphone array applications," in *Proceedings of ICASSP 98*, May 1998, vol. 6, pp. 3613–3616.

[4] H. Bourland and S. Dupont, "Subband-based speech recognition," in *Proceedings of ICASSP 97*, 1997, pp. 1251–1254.

[5] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards ASR on partially corrupted speech," in *Proceedings of ICSLP 96*, October 1996.

[6] H. Fletcher, "The nature of speech and its interpretation," *J. Franklin Inst.*, vol. 193, no. 6, pp. 729–747, 1922.

[7] J. B. Allen, "How do humans process and recognise speech?," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, 1994.

[8] H. Cox, R. Zeskind, and M. Owen, "Robust adaptive beamforming," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 35, no. 10, pp. 1365–1376, October 1987.

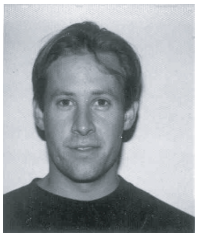
[9] W. Täger, "Near field superdirectivity (nfsd)," in *Proceedings of ICASSP '98*, 1998, pp. 2045–2048 (French).

[10] I. McCowan, C. Marro, and L. Mauuary, "Robust speech recognition using near-field superdirective beamforming with post-filtering," in *Proceedings of ICASSP 2000*, 2000, vol. 3, pp. 1723–1726.

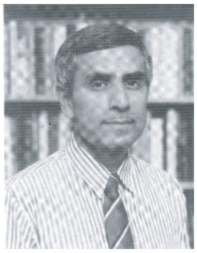
[11] A. Morris, A. Hagen, and H. Bourland, "The full combination sub-bands approach to noise robust HMM/ANN based ASR," in *Proceedings of Eurospeech '99*, 1999, pp. 599–602.

[12] J. Kittler, "Combining classifiers: a theoretical framework," *Pattern Analysis and Application*, vol. 1, no. 1, pp. 18–27, 1998.

[13] C. Marro, Y. Mahieux, and K. U. Simmer, "Analysis of noise reduction and dereverberation techniques based on microphone arrays with postfiltering," *IEEE Transactions on Speech and Audio Processing*, vol. 6, no. 3, pp. 240–259, May 1998.



Iain A. McCowan received the B. Eng(Hons) and B. InfoTech degrees from the Queensland University of Technology, Brisbane, in 1996. In February 1998 he joined the Research Concentration in Speech, Audio and Video Technology at the Queensland University of Technology where he is currently completing his Ph.D. His main research interests are in the fields of robust speech recognition and speech enhancement using microphone arrays. Mr McCowan is a student member of the Institute of Electrical and Electronic Engineers.



Sridha Sridharan obtained his B.Sc. (Electrical Engineering) and M.Sc. (Communication Engineering) from the University of Manchester Institute of Science and Technology, United Kingdom and Ph.D. (Signal Processing) from the University of New South Wales, Australia. Dr. Sridharan is Senior Member of the IEEE, USA and a Corporate Member of IEE, United Kingdom and IEAust of Australia. He is currently Professor in the School of Electrical and Electronic Systems Engineering of the Queensland University of Technology (QUT), and is also the Head of the Research Concentration in Speech, Audio and Video Technology at QUT.