MICROPHONE ARRAY SUB-BAND SPEECH RECOGNITION

Iain A. McCowan[†] *Sridha Sridharan*

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

ABSTRACT

This paper proposes the integration of sub-band speech recognition with a microphone array. A broadband beamforming microphone array allows for natural integration with sub-band speech recognition as the beamformer is typically implemented as a combination of band-limited sub-arrays. In this paper, rather than recombining the sub-array outputs to give a single enhanced output, we propose the fusion of separate hidden Markov models trained on each sub-array frequency band. In addition, a dynamic sub-band weighting scheme is proposed in which the cross- and auto-spectral densities of the microphone array inputs are used to estimate the reliability of each frequency band. The microphone array sub-band system is evaluated on an isolated digit recognition task and compared to the standard full-band approach. 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.

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]. Research to date has treated the microphone array and speech recognition systems as two distinct components - beamforming is performed on the microphone array inputs to provide a single channel enhanced speech signal which is then passed through to a standard recognition system.

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. In this paper 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 non-uniform manner [3, 4]. 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 [5, 6].

The proposed system integrates the microphone array with sub-band speech recognition in two ways. Firstly, spatial filtering is performed on the input channels to enhance the input to each sub-band recogniser. 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 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.

The second level of integration proposed in this paper is a multi-channel 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 technique to determine the reliability of each sub-band on a word by word basis.

The proposed system is compared to a standard full-band microphone array recognition system, and a single channel sub-band recognition system. To assess the effectiveness of the proposed dynamic sub-band weighting algorithm, we compare the results with those obtained using fixed equal weights, and also with the optimal sub-band results.

2. MULTI-CHANNEL SUB-BAND RECOGNITION SYSTEM

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 recognition models and the sub-band combination.

2.1. Sub-array Beamformer

Previous work has shown that a sub-array near-field superdirective beamformer offers excellent performance in a speech recognition application [7]. Near-field superdirectivity [8] 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 sen-

[†]The author is the recipient of a Motorola Partnerships in Research Grant.

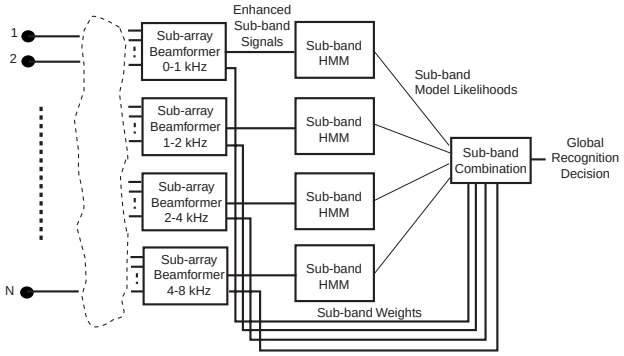

Figure 1: System Block Diagram

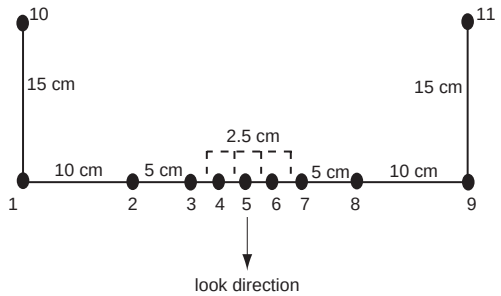

Figure 2: Array Geometry

sors. This is formulated as an optimisation 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 2 and the following sub-array configurations :

1. $0 < f < 1\ kHz$: microphones 1-11
2. $1\ kHz < f < 2\ kHz$: microphones 1, 2, 5, 8 and 9
3. $2\ kHz < f < 4\ kHz$: microphones 2, 3, 5, 7 and 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 a uniform spacing of 10 cm, 5 cm and 2.5 cm respectively.

The sub-array channel filters, $b_i^s(f)$, are calculated using the algorithm detailed by Täger [8], and are band-pass filtered between the specified upper and lower sub-array frequencies for each sub-band. At the output of each channel filter we have

$$v_i^s(f) = b_i^s(f)x_i(f) \tag{1}$$

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 normalised sum across channels as

$$y^s(f) = \frac{1}{\sum_{i=1}^{N} b_i^s(f)} \sum_{i=1}^{N} v_i^s(f) \tag{2}$$

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.

2.2. Sub-band Recognition Models

Sub-band speech recognition is based upon the work of Fletcher [5] (reviewed by Allen in [6]) which investigated the way in which humans recognise 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 recognised speech unit. In humans, this optimal fusion of articulation band results reduces the overall error rate according to the *product of errors rule*. This rule states that the full-band error rate is equal to the product of the sub-band error rates [5]. 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 [3, 4].

In 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. Of the parameterisation methods examined, sub-band mel-frequency cepstral coefficients (with energy and delta coefficients) were found to give the best results in our experiments. The application for the speech recognition experiments in this paper is an isolated digit recognition task using the single digit utterances from the male adult portion of the TIDIGITS database. Word models were used due to the small vocabulary (’zero’ to ’nine’). The models were trained using the input to the centre microphone in the array.

2.3. Sub-band Combination

To form a global decision, the output log-likelihood scores from each sub-band HMM must be combined in some way. In the proposed system, recombination is performed at the word level. If each sub-band output score is weighted by a factor α^s, we can calculate the global log-likelihood for a particular word m as :

$$L_m = \sum_{s=1}^{4} \alpha^s \log p(\tilde{y}^s|\lambda_m^s) \tag{3}$$

where $\tilde{y}^s$ is the set of sub-band MFCC observation vectors for a word and λ_m^s is the model for the s^{th} sub-band. The values of α^s are positive and are normalised to sum to unity. The decision then consists simply of selecting the word with the maximum global log-likelihood score.

3. CALCULATION OF SUB-BAND WEIGHTS

3.1. Dynamic Sub-band Weighting Algorithm

Clearly the success of the sub-band recognition approach is critically reliant on the sub-band weighting factors, α^s. The reliability of each sub-band recognition result depends upon the amount of speech and noise energy in the given frequency band and so a measure of the relative proportion of speech energy in each sub-band should be an effective means of determining the sub-band weights.

Multi-channel techniques provide us with a convenient means of estimating the signal to noise ratio. Under the assumptions that the noise and speech are uncorrelated, and that the noise has

low correlation between sensors, by averaging the cross- and auto-spectral densities of the input channels we can estimate the ratio of speech to speech-plus-noise energy as follows [9]

$$\hat{W}^s(f) = \frac{\sum_{i=1}^{N} |b_i^s(f)|^2}{\Re\left\{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} b_i^s(f) b_j^{s*}(f)\right\}} \times \frac{\Re\left\{\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} \hat{\Phi}_{v_i^s v_j^s}(f)\right\}}{\sum_{i=1}^{N} \hat{\Phi}_{v_i^s v_i^s}(f)} \quad (4)$$

The values $\hat{\Phi}_{v_i^s v_j^s}(f)$,(respectively $\hat{\Phi}_{v_i^s v_i^s}(f)$) are the estimated cross (power) spectral densities of the channel-filtered signals v_i^s and v_j^s, which are calculated using a simple time recursive formula as

$$\hat{\Phi}^k_{v_i^s v_j^s}(f) = \gamma v_i^s(f) v_j^{s*}(f) + (1-\gamma)\hat{\Phi}^{k-1}_{v_i^s v_j^s}(f) \quad (5)$$

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 4 was thoroughly analysed by Marro *et al* [9] 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

$$\beta^s = \frac{1}{f_h^s - f_l^s} \sum_{f=f_l^s}^{f_h^s} \hat{W}^s(f) \quad (6)$$

where f_h^s and f_l^s are the high and low frequency cut-offs for the sub-band s. From this we take the average across each frame in the word utterance to give $\bar{\beta}^s$, and then determine the normalised sub-band weights as

$$\alpha^s = \frac{\bar{\beta}^s}{\sum_{i=1}^{4} \bar{\beta}^i} \quad (7)$$

3.2. Optimal Sub-band Weights

To measure the effectiveness of the above algorithm, it is desirable to compute the upper bound of the performance that can be obtained using a simple weighted combination of sub-bands. To determine this upper bound we use an iterative minimisation algorithm which uses a dispersion measure as its objective function. The global log-likelihood of each word is first calculated according to Equation 3, and then, given *a priori* knowledge of the correct word, the dispersion is calculated as

$$D = \max_{m \neq c}(L_m) - L_c \quad (8)$$

where the correct word is known to correspond to model c.

4. 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. As discussed in Section 2.2, 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.

sub-band	WRR
full-band	99.7%
1	92.7%
2	85.3%
3	85.5%
4	58.5%
combined	98.6%

Table 1: Sub-band Word Recognition Rates (clean speech)

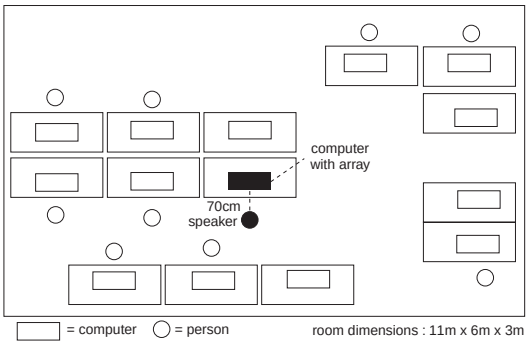

Figure 3: Experimental Setup

The experimental context is the computer room shown in Figure 3. 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.

The purpose of the experiments was to 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. As the advantages of the sub-band recognition technique will be most pronounced for band-limited noise, in the experiments we used white noise that was band-pass filtered so that only one sub-band was corrupted 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 results for different noise levels are given in Table 2 and are plotted in Figure 4. Results are given for the following cases :

- 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**)

The results clearly demonstrate several interesting trends. Firstly, the results for **single** and **BF** and also **single-SB** and **BF-SB** indicate the level of performance improvement to be obtained by using a multi-channel technique rather than a single channel system.

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	84.8%	80.3%	75.4%	70.3%
BF-SB	91.6%	86.5%	83.5%	76.7%
BF-SB-DW	97.2%	95.7%	94.9%	91.6%
BF-SB-OPT	98.8%	98.6%	97.6%	96.6%

Table 2: Word Recognition Rates

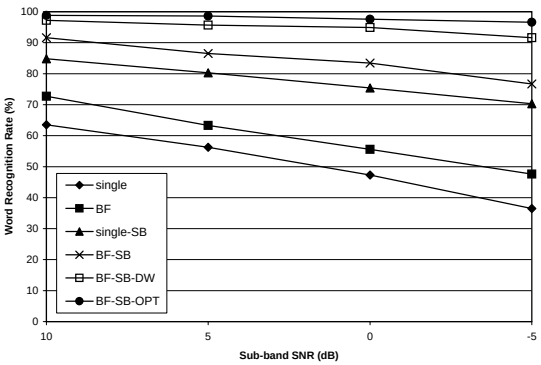

Figure 4: Speech Recognition Results

Due to the product of error rates between sub-bands, by reducing the error rate in each frequency band by beamforming, the overall error rate after recombination is significantly reduced. It is expected that this beneficial effect of the beamformer would be even more clearly observable in experiments where multiple sub-bands were corrupted with different noise levels.

Secondly, 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 it is clear that the multi-channel input provides an accurate and robust method for its estimation. While the sub-band system with fixed equal weighting gives excellent performance, it is seen that the proposed dynamic weighting algorithm is successful in providing significant further improvement in the results. While the optimal results indicate that some room for further improvement still exists, it is seen that the proposed system performs at a level comparable to the theoretical upper bound obtained using *a priori* knowledge of the correct results.

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. However, they serve to demonstrate the relative improvements that can be attained by integrating a sub-array beamformer with a sub-band recognition system, when compared to a single microphone system. Future work will investigate application of the proposed technique in more realistic noise scenarios.

5. CONCLUSIONS

An integration of microphone array beamforming and sub-band recognition techniques has been proposed. This integration is two-fold. Firstly, the sub-array beamformer provides enhanced inputs to each sub-band recogniser, considerably improving the overall performance by reducing the recognition errors in each sub-band. Secondly, the cross- and auto-spectral densities of the multi-channel input are used to estimate the signal to signal-plus-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 it is important to note that the proposed algorithm can be applied at lower levels as the sub-band weights can effectively be calculated on a frame by frame basis.

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 maximise the use of the available multi-channel input, the high levels of performance required for real applications are achievable in adverse conditions.

6. REFERENCES

[1] 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)*, pages 171–174, Tampere, Finland, May 1999.

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

[3] H. Bourland and S. Dupont. Subband-based speech recognition. In *Proceedings of ICASSP 97*, pages 1251–1254, 1997.

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

[5] H. Fletcher. The nature of speech and its interpretation. *J. Franklin Inst.*, 193(6):729–747, 1922.

[6] J. B. Allen. How do humans process and recognise speech? *IEEE Transactions on Speech and Audio Processing*, 2(4):567–577, 1994.

[7] 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.

[8] W. Täger. Near field superdirectivity (NFSD). In *Proceedings of ICASSP '98*, pages 2045–2048, 1998.

[9] 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.