IDIAP RESEARCH REPORT

Microphone Array Speech Recognition: Experiments on Overlapping Speech in Meetings

Darren Moore [a] Iain McCowan [a]

IDIAP–RR 02-41

October 2002

to appear in
2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, April 2003

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

[a] IDIAP

IDIAP Research Report 02-41

Microphone Array Speech Recognition: Experiments on Overlapping Speech in Meetings

Darren Moore Iain McCowan

October 2002

to appear in

2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong, April 2003

Abstract. This paper investigates the use of microphone arrays to acquire and recognise speech in meetings. Meetings pose several interesting problems for speech processing, as they consist of multiple competing speakers within a small space, typically around a table. Due to their ability to provide hands-free acquisition and directional discrimination, microphone arrays present a potential alternative to close-talking microphones in such an application. We first propose an appropriate microphone array geometry and improved processing technique for this scenario, paying particular attention to speaker separation during possible overlap segments. Data collection of a small vocabulary speech recognition corpus (Numbers) was performed in a real meeting room for a single speaker, and several overlapping speech scenarios. In speech recognition experiments on the acquired database, the performance of the microphone array system is compared to that of a close-talking lapel microphone, and a single table-top microphone.

1 Introduction

Meetings are a fundamental human activity, in which speech (along with other modalities) is used to share and develop information between a group of people. For this reason, meetings present an important application domain for speech processing technologies.

One of the problems that arises in meeting speech is that of multiple concurrent speakers. Overlapping speech may occur when someone attempts to take over the main discussion, when someone interjects a brief comment over the main speaker, or when a separate conversation takes place in addition to the main discussion. In [1] it was identified that around 10-15% of words or 50% of speech segments in a meeting or telephone conversation contain some degree of overlapping speech. These overlapped speech segments are problematic for speech recognition, producing an absolute increase in word error rate of between 15-30% using close-talking microphones for a large vocabulary task [1, 2].

In this paper, we investigate the use of a microphone array for acquisition of speech in meetings. Close-talking microphones are used in most applications, as their proximity ensures a high signal level, and also because the speaker constitutes an acoustic barrier that reduces room reverberation effects. The major disadvantage, however, is that meeting participants are required to wear lapel or head-mounted microphones. Microphone arrays offer a potential solution to remove this constraint.

A microphone array has the ability to discriminate between multiple competing speakers based on their location. Recent work has shown that microphone arrays can provide a viable alternative to close-talking microphones for single speaker speech recognition in noisy environments [3, 4]. While the robustness of microphone array recognition systems to background noise and general localised noise sources has been established, as yet, no recognition results have been published investigating the performance of an array system in the presence of genuine competing speech.

This paper makes several contributions. We first propose a microphone array system suitable for use in small meetings. For the enhancement, we present a simplification of the post-filtering approach presented in [5], which incorporates the theoretical noise field coherence in the post-filter estimation. In this framework, we propose a coherence model that takes both the background noise and localised speakers into consideration. A geometry is proposed and analysed to show the theoretical discrimination between speaker locations in a common meeting configuration.

Following this, we describe the collection of experimental data in a real meeting room. The Numbers recognition corpus is played through loudspeakers for various single and competing speaker scenarios, and re-recorded on lapel, table-top and array microphones. To address the current lack of multi-channel speech corpora, the resulting multi-channel database is being distributed [6]. Speech recognition experiments are performed on the database, comparing the performance of the proposed array processing system to that of the close-talking lapel microphones, as well as a single table-top microphone. The results demonstrate both that arrays present a viable alternative to close-talking microphones for single speakers, and that they can be successful in combatting the effects of overlapping speech.

2 Microphone Array System

In this section, we present a microphone array system for small meetings, discussing both the enhancement technique as well as a suitable array geometry.

2.1 Enhancement Technique

A block diagram of the microphone array processing system is shown in Figure 1. It includes a filter-sum beamformer followed by a post-filtering stage.

For the beamformer, we use the superdirective technique to calculate the channel filters w_n maximising the array gain, while maintaining a minimum constraint on the white noise gain. This technique is fully described in [7].

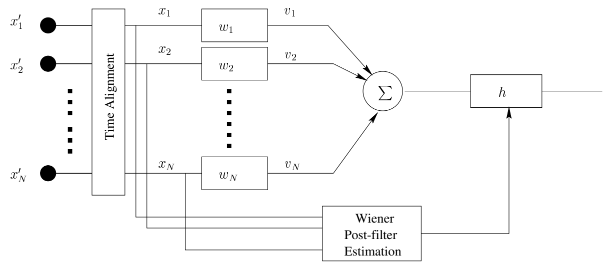

Figure 1: Filter-sum beamformer with post-filter

For the post-filtering stage, we apply the post-filter proposed in [5], with two modifications. This post-filter is a generalisation of the standard Zelinksi post-filter [8, 9] in that the assumption of an incoherent noise field is replaced with that of an assumed noise field coherence model. The first modification we propose is to simplify the estimation procedure described in [5]. The second modification is to incorporate localised competing speakers in the noise field coherence model.

2.1.1 Post-filter Estimation

Use of a post-filter after the beamforming stage helps to further reduce the broadband noise energy [10]. The Wiener post-filter transfer function is given by (omitting the frequency dependence for simplicity) :

$$h = \frac{\phi_{ss}}{\phi_{ss} + \phi_{nn}}$$

where ϕ_{ss} and ϕ_{nn} are the signal and noise power spectral densities, respectively.

In [5] we proposed a technique for estimating a microphone array post-filter based on an assumed noise field coherence model, Γ_{nn}. The formulation required the solution of a quadratic equation to estimate the signal power spectral density, ϕ_{ss}. This has the inconvenience of needing to choose between dual solutions, and also increased computational complexity over the standard Zelinski technique upon which it is based [8, 9]. Here we present a simplification of the approach to address these limitations.

Following application of the time alignment, and under the assumptions of : the same desired signal component across sensors ($\phi_{s_i s_j} = \phi_{ss}$); no correlation between the signal and noise ($\phi_{sn} = 0$); the same noise power spectrum on each channel ($\phi_{n_i n_i} = \phi_{nn}$); and a known noise field coherence model; we can write :

$$\begin{aligned} \phi_{x_i x_j} &= \phi_{ss} + \phi_{n_i n_j} \\ \phi_{x_i x_i} &= \phi_{ss} + \phi_{nn} \\ \Gamma_{n_i n_j} &= \frac{\phi_{n_i n_j}}{\phi_{nn}} \end{aligned}$$

where $\phi_{x_i x_j}$ is the cross spectral density between microphones i and j. From these equations, we obtain

$$\phi_{ss} = \frac{\phi_{x_i x_j} - \Gamma_{n_i n_j} \phi_{x_i x_i}}{1 - \Gamma_{n_i n_j}} \tag{1}$$

As the coherence can vary between $-1 \leq \Gamma_{nn} \leq 1$, the problem of division by zero should be avoided by enforcing an upper bound on the coherence values. The estimated signal power spectral density can then be averaged over all sensor pairs to give a robust estimation of the Wiener post-filter

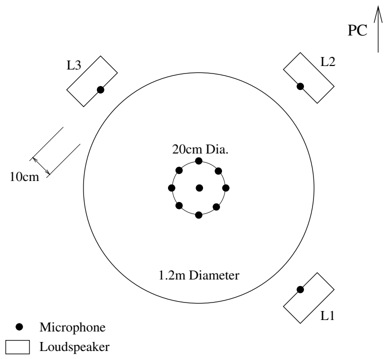

Figure 2: Meeting Room Configuration

numerator. This approach differs from that in [5] by making the assumption of same noise power spectral density on all sensors. The denominator $(\phi_{ss} + \phi_{nn})$ can be estimated by a similar averaging of $\phi_{x_i x_i}$. As well as having a single direct solution, the above formulation gives a significant reduction in computational complexity over that proposed in [5].

2.1.2 Proposed Noise Field Coherence Model

The post-filter estimation requires a coherence model, Γ_{nn}. In the previous work, a diffuse noise field was assumed, as it gives a good approximation of a number of practical noise situations, such as office noise. In the experiments in this paper however, significant noise energy comes from localised noise sources, rendering a purely diffuse noise field model inappropriate. We thus propose using a coherence matrix which is the weighted sum of the diffuse noise coherence and that of the localised noise sources, as follows :

$$\Gamma_{n_i n_j} = \frac{\phi_{dd}\mathrm{sinc}\left(\frac{2\pi f d_{ij}}{c}\right) + \Sigma_{l=1}^{L}\phi_{ll}\exp\left(\frac{j2\pi f d_{ij}}{c}\right)}{\phi_{dd} + \Sigma_{l=1}^{L}\phi_{ll}} \tag{2}$$

where ϕ_{dd} is the component of the noise power due to the diffuse noise field, and ϕ_{ll} that due to the l^{th} localised noise source. These components are not normally known in advance, and in practice, hand-adjusted, frequency-independent values can be used to achieve a compromise between diffuse and localised noise reduction.

2.2 Array Geometry

In this paper, we investigate a scenario where four people are equally spaced around a small meeting table. To give uniform discrimination over all possible speaker locations, we propose a circular array at the centre of the table. In particular, we use 8 equally spaced omnidirectional microphones, with diameter 20 cm. This configuration is shown in Figure 2.

Figure 3 shows the directivity patterns for the beamformer at 100 Hz and 1000 Hz using the proposed geometry. The patterns in all four speaker directions are superimposed, with speaker 1 shown in bold. From 3 (a) we see that, despite poor low frequencey directivity, the theoretical array attenuates the competing speakers to approximately 30% of the level of the desired speaker. As seen in Figure 3 (b), this discrimination improves at higher frequencies, with nulls developing in the direction of competing speakers.

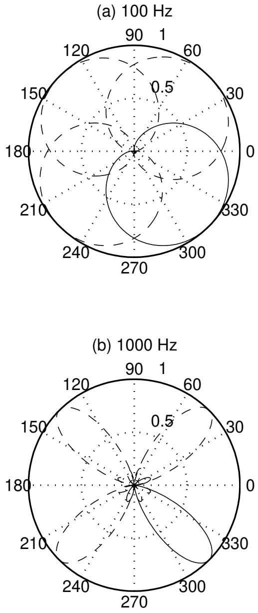

Figure 3: Directivity patterns at 100 Hz and 1000 Hz (speaker 1 in bold)

3 Data Collection

There are currently no publicly available corpora suitable for microphone array speech recognition research. Therefore, the initial focus of this work was to generate a multi-microphone corpus for experimentation and public distribution.

The database was collected by outputting utterances from the Numbers corpus (telephone quality speech, 30-word vocabulary) on one or more loudspeakers, and recording the resulting sound field using a microphone array and various lapel and table-top microphones. The goal of this work was to compare relative speech recognition performance using different microphone configurations in various noise situations, and thus a small vocabulary corpus was considered appropriate.

Three loudspeakers (L1, L2, L3) were placed at 90deg spacings around the circumference of a 1.2m diameter circular table at an elevation of 35cm. The placement of the loudspeakers simulated the presence of a desired speaker (L1) and two competing speakers (L2 and L3) in a realistic meeting room configuration.

An 8-element, equally spaced, circular array of 20cm diameter was placed in the middle of the table, and an additional microphone was placed at the centre of the table. Lapel microphones were attached to t-shirts hanging from each of the loudspeakers. The same type of omnidirectional microphone was used in all locations. The circular table was located at one end of a moderately reverberant, 8.2m$\times$3.6m$\times$2.4m rectangular room. The dominant non-speech noise source was a PC located at the opposite end of the room. The experimental configuration is illustrated in Figure 2.

The energy levels of all utterances in the Numbers corpus were normalised to ensure a relatively constant desired speech level across all recordings. The corpus was then divided into a 6049-utterance training set, a 2026-utterance cross validation set, and a 2061-utterance test set. "Competing-speaker" versions of the cross-validation and test sets were also produced by rearranging the order of their respective utterances.

The cross-validation and test sets were output from L1 with no overlapping speech and recorded on all microphone channels. This was then repeated multiple times in the presence of overlapping speech output from L2 and/or L3. All three possible competing speaker scenarios were considered. The

Scenario	Lapel	Centre	Array
S_1	7.01	10.06	7.00
S_{12}	26.69	60.45	19.37
S_{13}	22.17	54.67	19.26
S_{123}	35.25	73.55	26.64

Table 1: Word error rate results (%)

output levels of L1, L2 and L3 were identical, and were kept constant in all recording scenarios. The multi-loudspeaker data playback and multi-microphone recording were managed by the same equipment, which ensured that all input channels were simultaneously sampled, and that the microphone recordings were synchronised with the loudspeaker outputs. The sampling rate used in all recordings was 8kHz. All subsequent discussion will refer to the recording scenarios as S_1 (no overlapping speech), S_{12} and S_{13} (1 competing speaker), and S_{123} (2 competing speakers).

A multi-microphone corpus containing all recordings detailed above has been compiled, and is now publicly available [6].

For these experiments, the microphone array processing described in Section 2 was applied to the microphone array recordings from each scenario in order to enhance the desired speech signal.

4 Experiments and Results

A baseline speech recognition system was trained using HTK on the clean training set from the original Numbers corpus. The system consisted of 80 tied-state triphone HMM's with 3 emitting states per triphone and 12 mixtures per state. 39-element feature vectors were used, comprising 13 MFCC's (including the 0th cepstral coefficient) with their first and second order derivatives. The baseline system gave a WER of 6.32% using the clean test set from the original Numbers corpus.

Three recorded "channels" resulting from the data collection and microphone array processing were retained for speech recognition model adaptation and performance evaluation. These channels were:

- Centre tabletop microphone recording (Centre)
- Desired speaker (L1) lapel microphone recording (Lapel)
- Enhanced output of microphone array processing (Array)

MAP adaptation was performed on the baseline models using the cross-validation set for each channel-scenario pair, and then the speech recognition performance of the adapted models was assessed using the corresponding recorded test set. Table 1 shows the word error rate (WER) results for all channel-scenario pairs.

From the S_1 results, the WER's for the Array and Lapel channels were the same, and comparable to that of the baseline system. This shows that the recognition performance of the table-top microphone array is equivalent to the close-talking lapel microphone in low noise conditions. The WER for the centre microphone channel is slightly higher due to its distance from the desired speaker, and greater susceptibility to room reverberation.

The addition of a single competing speaker (S_{12} and S_{13}) (resulting in approximately 0dB SNR at the centre microphone location) had a severe effect on the WER for the centre microphone channel. The lapel microphone channel performed substantially better due to its proximity to the desired speaker. This difference in WER was more pronounced when a second competing speaker was introduced in S_{123} (resulting in approximately -3dB SNR at the centre microphone location).

In all overlapping speech scenarios, the microphone array output gave better word recognition performance than both the centre and lapel microphone channels. These results are put in context

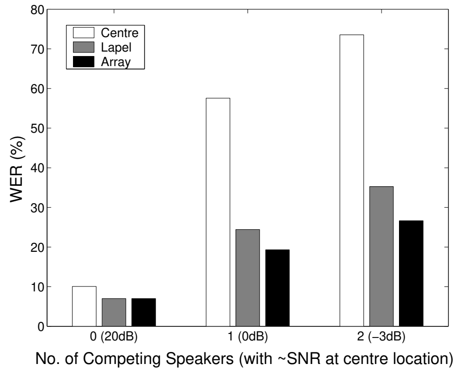

Figure 4: WER results for different numbers of competing speakers

when one considers that the individual microphones in the array were each subjected to essentially the same sound field as the centre microphone. The signal enhancement provided by the array processing overcame the lower SNR and increased reverberation susceptibility, and improved recognition accuracy to a level that exceeded that of the close-talking lapel microphone.

Figure 4 illustrates the WER trends for each channel in scenarios with 0, 1 and 2 competing speakers. The values plotted for the single competing speaker case are the average of the S_{12} and S_{13} WER results shown in Table 1.

5 Conclusions

In this work a table-top microphone array suitable for use in a meeting room was presented. A microphone array postfilter based on an assumed noise field coherence model was used where the coherence model was formulated to account for multiple localised noise sources within an otherwise diffuse noise field.

Speech recognition performance using the output of the microphone array was compared to recognition performance using both a close-talking lapel microphone attached to the desired speaker, and a single microphone in the centre of a meeting table. When no overlapping speech was present, the array output recognition performance was equivalent to that of the lapel microphone. In the presence of overlapping speech, the microphone array successfully enhanced the desired speech, and gave the best recognition performance of all microphone configurations tested.

A microphone array speech recognition database based on the Numbers corpus was recorded during this work, and is now available for public distribution [6].

6 Acknowledgements

The authors wish to thank the Swiss National Science Foundation for supporting this work through the National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management (IM2)".

This work was also funded by the European project "M4: MultiModal Meeting Manager", through the Swiss Federal Office for Education and Science (OFES).

The authors also acknowledge the Center for Spoken Language Understanding at OGI for their cooperation in the distribution of the corpus collected during this work.

References

[1] E. Shriberg, A. Stolcke, and D. Baron. Observations on overlap: findings and implications for automatic processing of multi-party conversation. In *Proceedings of Eurospeech 2001*, volume 2, pages 1359–1362, 2001.

[2] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke. The meeting project at ICSI. In *Proceedings of Human Language Technology Conference*, 2001.

[3] 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.

[4] 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.

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

[6] Multichannel Numbers corpus distribution. http://www.idiap.ch/speech/.

[7] H. Cox, R. Zeskind, and M. Owen. Robust adaptive beamforming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(10):1365–1376, October 1987.

[8] 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.

[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.

[10] K. Uwe Simmer, J. Bitzer, and C. Marro. Post-filtering techniques. In M. Brandstein and D. Ward, editors, *Microphone Arrays*, chapter 3, pages 36–60. Springer, 2001.