# An Analysis of Automatic Speech Recognition with Multiple Microphones

Davide Marino and Thomas Hain

Speech and Hearing, Department of Computer Science, Univ. of Sheffield, UK

d.marino@dcs.shef.ac.uk, t.hain@dcs.shef.ac.uk

## Abstract

Automatic speech recognition in real world situations often requires the use of microphones distant from speaker's mouth. One or several microphones are placed in the surroundings to capture many versions of the original signal. Recognition with a single far field microphone yields considerably poorer performance than with person-mounted devices (headset, lapel), with the main causes being reverberation and noise. Acoustic beamforming techniques allow significant improvements over the use of a single microphone, although the overall performance still remains well above the close-talking results. In this paper we investigate the use of beam-forming in the context of speaker movement, together with commonly used adaptation techniques and compare against a naive multi-stream approach. We show that even such a simple approach can yield equivalent results to beam-forming, allowing for far more powerful integration of multiple microphone sources in ASR systems.

**Index Terms**: far field speech recognition, beamforming.

## 1. Introduction

The development of automatic speech recognition technology has come a long way in recent decades and now performs very well in situations where speech recording quality is high. When speech recognition is brought into natural environments to transcribe natural conversations substantial degradation of performance is observed. Two main contributing factors can be identified: The type of speech, e.g. conversational vs planned, and the recording environment. The latter can be attributed to environmental noise, reverberation, and poor recording quality due to the distance of the speaker to the microphone. While even for a microphone close to the mouth of the speaker the noise (and this includes other speakers in the room) still plays an important role. The difference between near and far field based speech recognition is still substantial, typically error rate increases of more then 20% relative are observed.

Microphone arrays can help in two ways. By listening in the speaker's direction the speaker signal is enhanced, while at the same time noise distortions are attenuated. Many ASR systems have made use of microphone arrays for both diarisation and recognition (e.g. [1, 2]), but in almost all cases a speech enhancement based approach was chosen, splitting the task at hand into two parts: enhancement and recognition. This in itself is inconsistent with most paradigms used in standard ASR systems and can be expected to be sub-optimal.

Standard beam-forming combines multiple signals to emphasise the desired source (rather than focus on suppression) by listening in a specific direction. The advantage of such an approach is that the remaining ASR system stays unaltered in structure. Even better, in an ideal situation one would expect to use acoustic models trained from close-talking data (clean condition) for far field processing. The reality however is different. It was shown that training on beam-formed data is important, partly because the beamforming process itself causes distortion by having different characteristic for different frequencies.

In this paper we present an analysis on the effectiveness of beamforming for conversations in meetings, including adaptation and normalisation experiments, and set results into the context of speaker location and movement. Some initial steps towards integration into the ASR system infrastructure are presented. In particular the use of multiple microphones in form of multi-stream input is investigated. In Section 2 beam-forming as used in state-of-the-art systems is discussed and set into relation with feature based schemes. Great care has been taken on data selection, and details are given in Section 3. Section 4 then gives experimental results on beamforming, adaptation and feature concatenation, followed by a summary and conclusions.

## 2. Multiple microphones

Work on large vocabulary far field recognition in recent years has been dominated by the NIST paradigm for meeting recognition as defined the Rich Transcription evaluation campaign[1]. Here a close talking microphone recordings are compared with far field recognition results in diverse recording settings. While greatly aiding in general research on the topic there are two short-comings for our purpose: the diversity of recording facilities is great and hence any algorithm used must follow the lowest specification available; and overlapped speech is mostly discarded from evaluation. The detail on beam-forming algorithm used has to be interpreted in the light of these constraints.

### 2.1. Beamforming

Beam-forming is extensively described in literature (e.g. [3]). Adaptive beamforming, in particular Minimum Variance Distorsionless Response (MVDR) [4], is discussed in this section.

Consider a signal recorded from an array of microphones $x_i(t)$, the signal is an attenuated and delayed version together with noise: $X_i(t) = a_i s(t - \tau) + n_i$. In the frequency domain we have:

$$X(\omega) = S(\omega)d(\omega) + N(\omega)$$

where the vector $d$ denotes the delay vector for $M$ microphones.

$$d(\omega) = [a_1 e^{-j2\pi\omega\tau_1}, a_2 e^{-j2\pi\omega\tau_2}, ..., a_M e^{-j2\pi\omega\tau_M}]$$

$a_i$ and $\tau_i$ are the attenuation and delays respectively that depend on the distance between the signal source and microphones.

[1] See http://www.itl.nist.gov/iad/mig//tests/rt

To obtain the original signal, the microphone inputs are weighted by a frequency domain coefficient $W(\omega)$, so that the beam-formed output is $Y(\omega) = W^H(\omega)X(\omega)$. In order to find the optimal coefficient values $W$ it is required to minimize the mean squared error between the original signal and the output signal $W = \arg\min_W (W^H(\omega)X(\omega) - S(\omega))^2$. The solution is well known and is usually called MVDR beam-forming:

$$W_{MVDR} = \frac{\Gamma^{-1}d}{d^H\Gamma^{-1}d}$$

where the matrix $\Gamma$ is called the noise coherence matrix. The solution depends on two values: the delay vector and the noise coherence matrix. The exact choice of these leads to different beam-forming design. Many approximation of the $\Gamma$ matrix have been proposed in literature such as [5] [6].

If the distance of the source from the microphone is known, the computation of $d$ is straightforward, but even if this distance is not known, it is possible to estimate both the gain $a_i$ and the delay $\tau_i$. In order to calculate the attenuation factors, the input signals are normalized using a factor $\alpha_i = \sqrt{\frac{E_{ref}}{E_i}}$ where $E_i$ as the average of the $K$ lowest energy frames for each channel $i$, and $E_{ref}$ the highest. The gain factors are computed as the ratio of frame energy between the reference channel and every other channel, for each time step. After the gain calibration the delays are computed with respect the reference channel in the previous step, looking for the Time Delay Of Arrival (TDOA). In order to do that the Generalized Cross Correltion ( GCCPHAT) is used. The GCCPHAT between the microphones i and j is defined as:

$$GCC^{ij}_{PHAT}(\omega) = \frac{x_i(\omega)x^*_j(\omega)}{|x_i(\omega)x^*_j(\omega)|}$$

.
and the TDOA is the maximum of the inverse Fourier Transform of the GCCPHAT.

### 2.2. An Approximation

Summation of time-delayed channels allows simple integration and only requires few easily understandable quantities. However, in practice we do not know the location of the speaker and thus estimation of the speaker direction is necessary, as outlined above. The actual estimation of delays is non-trivial in practice and requires extensive smoothing techniques such as Wiener filtering and Viterbi decoding of multiple delay estimate paths[7]. Robustness thus is hard to achieve. Second, the techniques used are mostly geared to listening to single dominant sources of energy, the detection of two people speaking at the same timing is still very challenging and mostly avoided by selection of a single dominant source. Faced with these difficulties alternative solutions are desirable that are easy to estimate and allow flexible change for multiple speakers.

The key metric for estimation for beam-forming and localisation is based on minimal signal distortion from the source (with some exceptions, e.g. [2]). This however has been shown to yield suboptimal results in other areas of speech technology as it may focus on signal aspects that are not relevant for perception, by human or machine. One would rather prefer to optimise speech recognition metrics directly. If we consider the signals obtained from two microphones, $x_{1,2}(t)$, and the associated spectra $X_{1,2}(f)$, the delayed summation of the signals is given by

$$Y(f) = X_1(f)e^{j\omega d_1} + X_2(f)\,e^{j\omega d_2}$$

| | Standing | Sitting |
|---|---|---|
| Head movement (H+) | 3.6 | 13.7 |
| No Movement (H-) | 10.1 | 56.0 |

Table 1: Data available (hours) from basic movement categories in the AMI corpus.

The front-end of ASR systems are typically based on Mel-frequency cepstra (or are closely related). A magnitude spectrum is computed, followed by application of filter-banks, log compression and cepstral analysis. The conversion to magnitude renders the system unaffected by phase changes which are however a very important effect of acoustic distortion and were addressed in signal based metrics for beamforming. Computing the magnitude spectrum reveals the well-known form

$$|Y|^2 = |X_1|^2 + |X_2|^2 + 2\,|X_1|\,|X_2|\,cos\,(\Delta d + \Delta\varphi)$$

with $\Delta d$ and $\Delta\varphi$ denoting differences in delay and phase between the channels, respectively. When the two microphones are close together the differences in amplitude are in most likely small. If, one ignores such variations in a first order approximation, the adjusted spectrum is given as a product of the recorded magnitude and a multiplicative correction factor that reflects delay and phase differences only. Log compression will turn into additive distortion and hence, if individual feature vectors are computed from both recordings one obtains two strongly correlated observations, based on additive distortion. Joint modeling of the two observations in a Gaussian model then allows to take advantage of the correlation. In practise full covariance modeling is very expensive and can be approximated by Gaussian mixture models with diagonal covariances.

## 3. Data

The AMI corpus [8] is one of the most extensive collections of speech data recorded with multiple microphones in consistent configuration to date. Approximately 100 hours of recordings from three different rooms are available, and different recording streams (including individual headset microphones, lapel and multiple distant microphones) all aligned using a same timeline. For each meeting a manually produced orthographic transcription is provided at both segment and word level. These transcriptions are then aligned at the word level.

Aside from segment, word and speaker annotation additional basic information about the speaker location and movements are included in the corpus. For each segment the speaker location is labeled as sitting (at the table) or standing (near the whiteboard). Head turning is also labeled. For the purpose of consistent data construction meetings with more than four speakers and fewer than 8 microphones have been discarded. Table 1 gives details on the availability of data from each category. The majority of speech is generated in a still sitting position.

### 3.1. Selection

For the purpose of analysis of far-field processing the data was split into a range of different data sets with the objectives to have allow determination of significance 6 hour test sets were derived, and the natural distribution between movement classes as outlined in Table 1 would imply low amounts of data for some categories. Hence we derived two test configurations, one with equal distribution of data across classes, and one representative of the overall corpus statistic. Third, a differentiation

| Set | Sitting | | Standing | | | | |
|---|---|---|---|---|---|---|---|
| | H- | H+ | H- | H+ | #spk | #mtg | #seg |
| NE | 1.5 | 1.5 | 1.5 | 1.5 | 123 | 105 | 6333 |
| OE | 1.5 | 1.5 | 1.5 | 1.5 | 103 | 90 | 6017 |
| NR | 3.6 | 1.0 | 1.0 | 0.4 | 97 | 74 | 6923 |
| OR | 4.5 | 0.9 | 0.5 | 0.1 | 37 | 25 | 5963 |

Table 2: Test set statistics for four test sets. The naming represents inclusion of overlap and distribution, H± denotes presence of head movement. On the right hand side number of speakers, meetings, and segments are shown.

| | Sitting | | Standing | | |
|---|---|---|---|---|---|
| Set | H- | H+ | H- | H+ | Overall |
| NE | 25.4 | 26.1 | 29.8 | 25.9 | 26.8 |
| OE | 33.9 | 31.7 | 32.4 | 24.7 | 30.9 |
| NR | 27.1 | 26.6 | 29.2 | 26.2 | 27.3 |
| OR | 32.6 | 35.2 | 33.0 | 36.5 | 33.1 |

Table 3: %Word error rates using close talking microphones with CMN CVN

between speech in overlap (i.e. two or more speakers talking at the same time) and non-overlap is desirable. Finally, in order to allow good coverage across a large set of meetings, data selection per segment basis is preferable. These objectives gave rise to 4 test set definitions, with associated training set defined as the remainder of the corpus.

The test sets were created using a random process, that selects segments split in four different classes (sit H+/-, standing H+/-), choosing segment in two different ways either representative, where the amount of time is proportional to the size of a class, or equal, the same amount of time is picked for each class. The test sets are Non overlap Equal (NE), Non overlap Representative (NR), Overlap Equal (OE) and Overlap Representative (OR).

Table 2 shows the characteristics for each data set. For the OR data set one can observe that the least common data category is overlapped speech when standing and moving the head. For each of the test sets recordings from 8 microphones in circular array configuration as well as close talking (head-mounted) microphones. Table 3 shows baseline results for the 4 data sets. One can observer significant degradation for speech with overlap, however the effect is less pronounced when standing, presumably because the others speakers are further away.

# 4. Experiments

In the following experiments our standard configuration for the training and testing of acoustics models is used[1]. The front-end extracts twelve MF-PLP features at a rate of 100 Hz together with a zeroth cepstral coefficient, augmented with first and second order differential coefficients. Maximum likelihood phonetic decision tree state clustered triphone models based on 3-state left-to-right Hidden Markov Models (HMMs) are trained from scratch in each experiment to avoid any bias from bootstrapping with specific acoustic models. Models have typically around 3000 states and use 16 mixture components per state. For decoding purposes the AMI RT'07 50k vocabulary and an interpolated trigram language model is used (rttgint07).

Adaptation experiments explore the use of two related techniques, cepstral normalisation and transform based adaptation. The experiments are conducted in the form of speaker adaptation experiments, however these would be expected to capture

| Set | Headmounted | Far-field | | | |
|---|---|---|---|---|---|
| | 1 | 1 | 2 | 4 | 8 |
| NE | 26.8 | 60.2 | 54.6 | 52.5 | 50.8 |
| OE | 33.0 | 67.2 | 62.8 | 61.2 | 59.4 |

Table 4: %WER on different microphones using two different test set definitions.

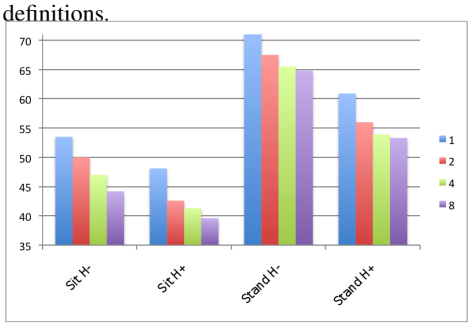


Figure 1: Implication of movement in WER for beam-forming experiments on NE data set.

location specific information as most speakers are in static position for the majority of the meetings. Cepstral mean normalisation (CMN) [9] was of course intended to combat static acoustic channel effects and has been used in a variety of tasks. In far field meeting recognition again the statistic is typically computed on a speaker basis as the concept of a channel is less clear. Cepstral variance normalisation (CVN) was applied in the same way[10]. The standard in model based adaptation is maximum likelihood linear regression (MLLR) and its constrained variant (CMLLR)[11, 12]. Using adaptation in this form for normalisation prior to adaptation leads to Speaker Adaptive Training (SAT)[13]. In this paper all adaptation and SAT experiments use CMLLR with one transform for speech models and one for silence.

### 4.1. Beamforming

As outlined in section 2.1, using a single distant microphone can yield significant performance improvements over a single microphone. Experiments were conducted using identical beamforming as used in the AMI evaluation systems[1] instead of a MDVR beacuse as discussed on the section 2.1 the distance between microphones is needed to find the optimal solution.

Table 3 presents the results for IHM overall the test sets using cepstral normalisation. The results highlight an important point: the effect of the movement on close talk, the performances decrease especially when the speaker in standing, that is a problem for close talk condition as well as for the far field. Table 4 presents the decoding results including overlap segments and their effect for both close talk and far-field, showing an average 6% absolute WER decrease.

Hidden in these results is the variation by movement. Figure 1 shows the results split by speaker movement. Notably the lowest word error rate is achieved when a person is sitting, the distance to the microphones is the most likely contributor to performance decrease when standing. Head movement surprisingly appears to have an overall positive effect and improvement from beam-forming is far better in the sitting condition. This can partly be caused by the fact that spatial separation between speakers is better in that case.

| 8 microphones BF | | | Sitting | | Standing | | |
|---|---|---|---|---|---|---|---|
| Set | CN | Adapt | H- | H+ | H- | H+ | Overall |
| NE | | | 43.5 | 39.7 | 66.5 | 53.7 | 50.8 |
| | × | | 44.2 | 39.6 | 64.8 | 53.3 | 50.4 |
| | × | S | 39.2 | 36.5 | 57.1 | 49.0 | 45.4 |
| | × | U | 43.9 | 38.8 | 64.5 | 53.3 | 50.1 |
| | × | SAT U | 43.4 | 38.6 | 64.6 | 52.8 | 49.8 |
| OE | | | 55.3 | 53.2 | 71.2 | 58.1 | 59.4 |
| | × | | 55.5 | 52.4 | 69.1 | 56.1 | 58.2 |
| | × | S | 51.3 | 49.3 | 63.0 | 51.2 | 53.7 |
| | × | U | 53.7 | 51.4 | 69.2 | 56.1 | 57.6 |
| | × | SAT U | 54.9 | 51.4 | 68.8 | 55.5 | 57.6 |

Table 5: %WER results on two different training and test set configurations. (CN) is cepstral mean and variance normalisation, (S) denotes supervised adaptation, (U) unsupervised.

### 4.2. Adaptation

Adaptation experiments investigated the utility of adaptation techniques on overlapped and non-overlapped data (the NE and OE data sets). Table 5 shows results with and without the use of CMN/CVN, supervised and unsupervised adaptation and SAT training on beam-formed data with 8 microphones. CMN/CVN appears to be modestly effective, but more so for overlapped speech. Overall, unsupervised adaptation gives only very modest gains and SAT training allows only little further improvement overall. Split by categories, unsupervised adaptation and SAT can even degrade performance, the highest gains are obtained in the most difficult category. Overall the high error rates still contribute to a marked difference between supervised and unsupervised results. Additional experiments (not shown here for lack of space) increasing the number of microphones used indicate that adaptation gains get smaller the more microphones are included.

### 4.3. Concatenation of channels

In section 2.2 we outlined that beamforming could be interpreted as causing (in rough approximation) linear scaling and shift of observed signals. Adaptation experiments for single microphone channels do not give the expected gains as the scale is inter-microphone delay dependent and thus is not available. In order to allow to learn these concatenated feature systems were constructed. Due to high dimensionality of the resulting systems experiments with 2 and 4 microphones only were conducted, yielding feature vector sized of 78 and 156 respectively. In contrast to the training procedure outlined above single pass retraining (SPR) from beam-formed models was used for initialisation of the concatenated feature models. The number of clustered states remains the same for these models. Using the SPR models for initialisation, a further set of models is trained (2M). The number of states increases from 2885 to 3473 and 5289 for the 2 and 4 feature systems respectively. Table 6 compares the results for the concatenated systems with beamforming. One can observe not only at least equivalent results to the use of beam-forming, but for 2 microphones the concatenated system outperforms the beamformed models. Compared to beam-forming the H- results are always lower, the expected detrimental effect of head movement is visible in the results.

## 5. Conclusions

In this paper analysis on the effectiveness of beam-forming for far field recognition in a meeting scenario was carried out.

| | | | Sitting | | Standing | | |
|---|---|---|---|---|---|---|---|
| | Train | # | H- | H+ | H- | H+ | Overall |
| | | 1 | 53.5 | 48.1 | 71.6 | 60.9 | 58.5 |
| BF | | 2 | 50.0 | 42.6 | 67.5 | 56.0 | 54.0 |
| C | SPR | 2 | 47.6 | 43.8 | 67.5 | 57.3 | 54.0 |
| C | 2M | 2 | 47.3 | 43.0 | 67.5 | 56.6 | 53.5 |
| BF | - | 4 | 47.0 | 41.3 | 65.5 | 53.9 | 51.9 |
| C | SPR | 4 | 47.0 | 43.0 | 65.4 | 55.1 | 52.6 |
| C | 2M | 4 | 45.7 | 43.1 | 64.6 | 54.8 | 52.0 |

Table 6: %WER results on the NE data sets using concatenation of features from multiple microphones. (SPR) denotes single pass retraining, (2M) retraining using 2-model re-estimation.

Large test sets were defined to give good coverage of available speaker movement classes and to allow separation of effects associated with speech overlap. We show that beam-forming is effective, but adaptation in addition only gives modest gains. Initial experiments on an alternative strategy, the concatenation of microphone channels, are conducted and found to yield at least equivalent performances. These encouraging results will allow the investigation into novel schemes for integration of beamforming into the speech recognition algorithms.

## 6. Acknowledgements

The research leading to these results has received funding from the EU-FP7/2007-2013 under grant agreement n° [213850].

## 7. References

[1] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, D. van Leeuwen, and V. Wan, "The 2007 AMI(DA) System for Meeting Transcription." Springer, 2008, pp. 414–428.

[2] M. L. Seltzer and R. M. Stern, "Subband Likelihood-Maximizing Beamforming for Speech Recognition in Reverberant Environments," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 6, pp. 2109–2121, 2006.

[3] I. Himawan, I. McCowan, and S. Sridharan, "Clustered blind beamforming from ad-hoc microphone arrays," *IEEE Trans. on Audio, S & L Proc.*, vol. 19, no. 4, pp. 661–676, 2011.

[4] V. T. H. L., *Optimum array processing*. John Wiley and Sons, 2002.

[5] J. B. Allen, D. A. Berkley, and J. Blauert, "A multimicrophone signal-processing technique to remove room reverberation from speech signals," *JASA*, vol. 62, pp. 912–915, Oct. 1977.

[6] L. Griffiths and C. Jim, "An alternative approach to linearly constrained adaptive beamforming," *Antennas and Propagation, IEEE Transactions on*, vol. 30, no. 1, pp. 27–34, 1982.

[7] X. Anguera, C. Woofers, and J. Hernando, "Purity Algorithms for Speaker Diarization of Meetings Data," in *ICASSP'06*, 2006, pp. I–I.

[8] J. Carletta, S. Ashby, S. Bourban, M. Guillemot, M. Kronenthal, G. Lathoud, M. Lincoln, I. McCowan, T. Hain, W. Kraaij, W. Post, J. Kadlec, P. Wellner, M. Flynn, and D. Reidsma, "The AMI Meeting Corpus: A Pre-announcement." Edinburgh: Springer, 2005, pp. 28–39.

[9] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. on Acoustics Speech and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[10] S. Tibrewala and H. Hermansky, "Multi-Band and Adaptation Approaches to Robust Speech Recognition ," in *Proc. Eurospeech'97*, 1997, pp. 2619–2622.

[11] M. Gales and P. Woodland, "Mean and variance adaptation within the mllr framework," *Computer Speech & Language*, vol. 10, pp. 249–264, 1996.

[12] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer S & L*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[13] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker-adaptive training," in *in Proc. ICSLP*, 1996, pp. 1137–1140.