Extending Limabeam with discrimination and coarse gradients

Charles Fox, Thomas Hain

Speech and Hearing group
Department of Computer Science
University of Sheffield, UK
charles.fox@sheffield.ac.uk

Abstract

Limabeam is an approach to multi-microphone array processing for ASR which makes minimal assumptions about system geometry, instead searching for filters to maximise output likelihoods under a speech model. The first results of Limabeam on the AMI meeting corpus are given, then two extensions of the algorithm for this corpus. First, it is shown that the original local gradient following sticks in local minima, and a coarser gradient is used. Second, a new discriminative objective function is provided to handle mis-matched silence models. The extensions are based on examination of 2D receptive fields and 2D likelihood maps which are novel near-field analogs of radial beamformer response patterns, but do not show radial symmetry and have many local minima. The extended Limabeam improves WER on TDOA baselines on the AMI corpus, by 1% rel. when both are adapted with decodes and by 19% rel. when both adapted with ground truth.

Index Terms: ASR, beamforming, discriminative

1. Introduction

ASR in noisy, reverberant, distant-talking environments such as meeting rooms remains a difficult task [1, 2] as additive noise from overlapping speakers and non-speech sources and convolutional noise from reverberation degrade the signal. However in some cases, instrumentation with multi-microphone arrays may be possible, either via static installations or ad-hoc networks using, say, the participants' mobile phones. Such array signals can be processed with weighted delay and sum transforms, $\{w_{ij}\}$ of the multiple input channels $x_i[t]$ to an output channel $y[t]$ for ASR,

$$y[t] = \sum_i \sum_j w_{ij} x_i[t-j]. \quad (1)$$

Traditional beamformers [3] have used geometric assumptions to choose $\{w_{ij}\}$ to optimise criteria. For example, Time Delay Of Arrival (TDOA, [4]) is optimal for a single source in the presence of diffuse white noise; Maximum Variance Distortionless Response (MVDR, [5]) is optimal assuming the target source has the widest variance of any combination of sources. The Multiple Inverse Theorem (MINT, [6]) is optimal for discrete, non-diffuse noise sources. None of these assumptions are perfectly valid, and it has been noted [7] that such mismatch is acute as they all tend to be highly sensitive to low noise in any of their inputs.

Likelihood-maximising beamforming (Limabeam) was introduced to ASR by [8, 9] and tries to make minimal assumptions and instead try to search the $\{w_{ij}\}$ using gradient descent to maximise the likelihood of the signal under an ASR speech model. Using numerical gradient descent with a CMU Sphinx-3 speech model on CMU-8 [8] and CMU-WSJ-PDS corpora [9], gains of 31.4% relative have been reported [9], though with the caveat these significant gains we showed only for longer (>7s) utterances. These corpora contain only single static speakers reading written scripts, so that smaller utterances can easily be chunked together.

The present study tests for independent replication of these results on a standard, unscripted multi-speaker meeting corpus, AMI [1]. It finds no significant improvement over TDOA beamforming using basic Limabeam in this new setting – which has more realistic noise types and utterance characteristics – but suggests two extensions to Limabeam which then do allow it to improve on the TDOA scores. The first is a change of objective function, replacing likelihood with a discriminative likelihood ratio to avoid a problem with silence models in the AMI environment. The second is to replace local gradients with coarser gradients, allowing search to avoid some local minima.

Other improvements to Limabeam and AMI have been suggested, including sub-band and parameter-sharing Limabeam [10] [11], cepstral Limabeam [12],[13], and neural network recognition gains on AMI [14], which could all be combined with the present extensions.

2. Baseline experiments

Unlike CMU-8/WSJ-PDS, AMI consists of unscripted simulated business meetings by groups of four participants around a table. Recorded in three meeting rooms, each is instrumented with a circular, 100mm radius array of omnidirectional microphones in the centre of the table. Training and test sets of 12,000 (15.7 hours) and 1,188 (1.9 hours) non-overlapping, human-segmented utterances are defined, having independent sets of speakers.

Unlike CMU-8/WSJ-PDS, no chunking of short utterances is performed as our tests are intended as a proxy for more general meetings where speakers may move, or where diarisation is unclear. Audio used here is at 16kHz, and converted to PLP features [15]. All processing was per-utterance unless stated otherwise.

Baselines were obtained for Individual Head Microphone (IHM) and Single Distant Microphone (SDM) channels after training 3-state left-to-right hidden Markov Models (HMMs) with state-clustered phonetic decision tree ties states of 16-component GMMs to model output probabilities. Training followed a standard HTK mixup procedure [16]. Word error rates (WERs) are obtained with NIST *sclite* [17]. Decodes are based on HTK HDecode with a 3-gram language model trained from the AMI training set. Decode parameters were fixed at the start of baseline testing (s15p0) and were not changed to overfit later

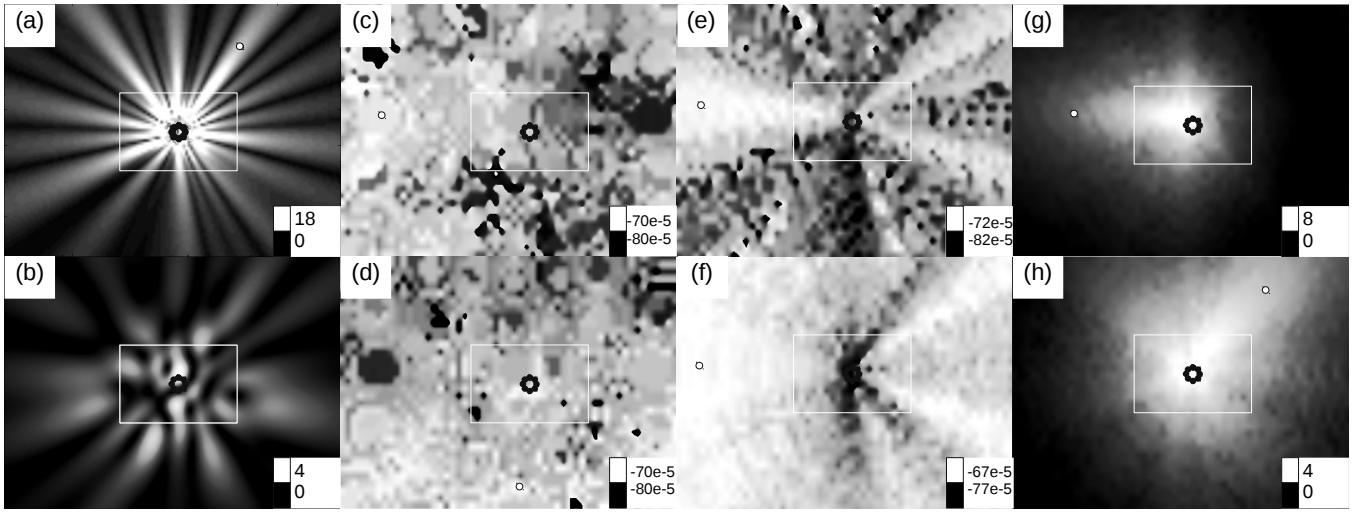

Figure 1: (a) 8kHz Receptive field for TDOA; (b.) 6kHz receptive field for random BF. (c),(d) Likelihood field for utterances. (e),(f) Desilenced likelihood fields. (g),(h) Discriminative likelihood ratio fields. White dots show speaker locations, black dots are microphones. x and y axis are physical coordinates in a 5.5x4m room. Rectangle shows a typical meeting table size.

experiments.

Baselines WERs were further obtained for standard beamformers. TDOA audio was created per-utterance via,

$$w_{ij} = \delta(j = \arg_\tau \max r(i, 0)), \quad (2)$$

where $r(i, 0)$ is the cross correlation function between channel i and a reference channel 0 chosen as that with the highest utterance energy, and δ is a Dirac Delta function. We also created baselines for audio output of the standard Beamformit (BFIT, [18]) software using its default settings, which is based on TDOA but including cross-utterance smoothing optimised for meeting rooms.

Table 1: *Baseline results.*

data	model	WER	S	D	I
IHM	xwrd	39.8	24.3	11.5	4.0
SDM	xwrd	66.0	45.5	16.7	3.9
IHM	MLLR(gnd)	23.4	12.0	8,7	2.7
SDM	MLLR(gnd)	50.9	32.4	15.8	2.7
IHM	MLLR(hyp)	37.2	17.4	16.5	3.3
SDM	MLLR(hyp)	62.7	39.4	19.9	3.4
TDOA	SPR	60.6	41.1	15.7	3.8
BFIT	SPR	61.2	40.0	17.8	3.5
TDOA	MLLR(gnd)	51.8	31.8	17.3	2.7
TDOA	MLLR(hyp)	59.4	36.3	19.9	3.2

The basic results in table 1 use Single Pass Retraining (SPR, [16]) to train new TDOA and BFIT models based on the previous IHM alignment. TDOA was found to outperform BFIT in this case.[1]

[1]Following this, GCC-PHAT[19] was also tested and found 11% worse that TDOA, and a static TDOA fixing parameters over all utterances for each speaker was 2.6% worse than TDOA. GCC-PHAT usually improves WER in strong reverberation but does not here. Static TDOA worsening was surprising as AMI speakers are seated and expected to have similar/smoothable TDOA values throughout, and shows the large effects of small head moves.

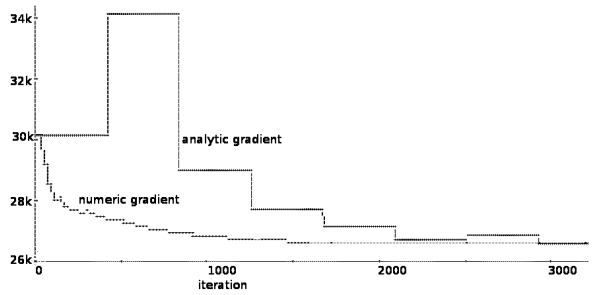

Figure 2: Convergence of gradient descent searches, using exact analytic and numerical estimated gradients, 10s utterance.

Maximum Likelihood Linear Regression (MLLR, [16], from IHM base) adaptation baselines for TDOA are also shown in table 1. MLLR is trained on a per-speaker basis using ground truth (gnd) and TDOA-decodes (hyp) test set transcripts; only means are adapted and 5 regression tree class transforms are used. Ground truth training gives an indication of what adaptation would achieve given large amounts of per-speaker training data.

3. Limabeam experiments

The most basic form of Limabeam models speech as a single GMM on MFCC features (a similar objective to a Wiener filter but optimising GMM feature likelihoods than GMM frequencies) and was tested on AMI. All Limabeam versions in the present paper are initialised to TDOA weights, and work with 8 microphones with 10 positive delay taps each (80 parameters).

Pilot experiments (e.g. fig. 2) found no significant differences in compute time or WER by switching from analytic gradients [9] to numerically computed gradients. Analytical solution makes many evaluations around each point to compute full-dimensioned gradients used to take a few large, accurate steps. Numerical gradient descent takes many smaller steps to give a smoother curve, but the same solution.

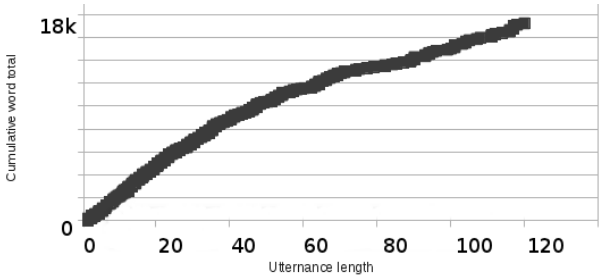

Figure 3: Cumulative distribution of AMI utterance lengths.

Standard BFGS Quasi-Newton was used to perform gradient-based optimisation in both cases (Octave *fminunc*, whose numerical gradients use w_{ij} moves of 6e-6). Numerical gradient computation allowed simple switching from MFCC to PLP features which gave 7.1% relative WER improvement. However even with this, the WER of table 2 did not improve on the relevant (TDOA, SPR, 60.6%) baseline, suggesting that LIMA-GMM likelihood is not a good proxy for WER optimisation.

A full HMM-based Limabeam was then implemented using HMMs. Alignment was performed at every parameter evaluation (80 for each gradient descent iteration) using HVite with ground truth transcripts. MLLR adaptation was applied and results are shown as LIMA-HMM in the table above. Again this standard Limabeam underperformed its baseline (TDOA, MLLR(gnd)).

Table 2: *Limabeam results.*

data	model	WER	S	D	I
LIMA-GMM	SPR	60.8	41.3	15.7	3.9
LIMA-HMM	MLLR	64.4	36.1	26.1	2.2

4. Inspection of corpus and beams

Previous work [9] showed on other corpora that Limabeam gave no improvement on short utterances, such as those less than 7s duration. To explore this for the AMI corpus, figs 3 and 4 show the word length distribution of AMI utterances and the WER of LIMA-HMM. These suggest that some of the overall poor performance is due to a large number of short utterances, confirming the findings of [9].

Inspection of the w_{ij} during optimization suggested that most utterances' parameters were shifting only by small amounts away from the initial TDOA solutions – comparable to the search step size of the optimiser. This could occur if TDOA solutions are already local (or global) optima, making it impossible for the optimizer to escape from them. To gain some intuition about the shape of the search space, we examined the theoretical (no reverb or noise) spatial receptive field patterns for various w_{ij} sets. Traditionally, beamformer responses are plotted only as functions of angle, not radius, under far-field assumption. This assumption might not hold for the distances in AMI corpus, where speakers sit <1m from the array. Fig. 1(a) shows a typical receptive field over one AMI meeting room (at 6kHz) for a TDOA filter focused on a speaker location. For TDOA filters it can be seen that the near-field effects are limited to a small region around the mic array so the usual radial plots are appropriate. However Limabeam can search a much

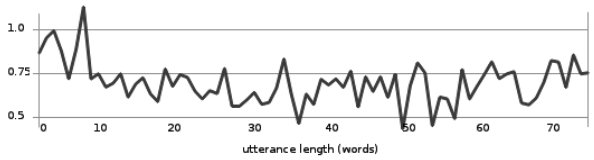

Figure 4: WER of standard LIMA-HMM by utterance length

larger, in our case 80D, space than the 3D manifold of TDOA solutions. Just one example of a receptive field given random weights is shown in 1(b), which indicates the ability to produce fields very unlike these classical side-lobed shapes, and where near-field effects do dominate the area where the AMI speakers sit. Such plots suggest many local minima where Limabeam could stick.

Single frequency receptive fields differ from *likelihood fields* however, which are illustrated in fig. 1(c) and (d). These show the effect of deliberately choosing $\{w_{ij}\}$ to focus on a speaker at the white dot, then moving the speaker around in simulation. At each pixel, the same utterance is placed there and the likelihood of the beamformed signal measured under the HMM model. This likelihood is plotted as the pixel color. The resulting likelihood fields are again very non-smooth and show many localised minima.

It must be emphasised that this 2D (actually a slice of a 3D room) field is not the space searched by the 80-dimensional LIMA search. But it does give a suggestion of the shape of the latter space and the types of minima found in it, and therefore that Limabeam is likely to get stuck often in such minima. The likelihood fields are highly diffuse, lacking clear single peaks at speaker locations. (This is the first publication to show such fields.)

5. Extensions

The AMI corpus contains strong, localised noise sources in at least one of the rooms having a loud server rack. It was found that 27% of segment time is assigned to silence. Together this could lead to optimising the filter to local minima that transform silence in new utterances to sound like silence in trained models, perhaps dominating any transform of speech sound. To test this hypothesis, two alternative optimisation objectives were constructed.

The *first* objective function is a per-frame likelihood average, but with silence phones excluded,

$$obj_1(\{w_{ij}\}) = \frac{1}{M}\sum_{n=1}^{N} b(n)\log P(x(n)|M,\{w_{ij}\}), \quad (3)$$

where there are N frames in the utterance and M contain non-silence phones in the current alignment indicated by the indicator function b, λ is the likelihood and M is the HMM model giving a best alignment.

Instead of simply excluding silence, the *second* objective function actively penalises the transformation of model silence into new-utterance silence by using the discriminative function,

$$obj_2(\{w_{ij}\}) = \log \frac{\prod_n P(x(n)|M,\{w_{ij}\})}{\prod_n P(x(n)|S,\{w_{ij}\})}, \quad (4)$$

where S is an HMM model consisting only of a single GMM trained on silence from the training set.

Sample likelihood fields for these two objective functions are shown in fig.1(e),(f) for obj_1 and fig.1(g),(h) for obj_2. Manual inspection of 10 utterances each suggested that these forms are typical, and that obj_2 tends to have a smoother form. This objective was thus selected for full testing.

While these 3D spatial fields look considerably smoother than the basic likelihoods, it might still be possible that local minima exist in the higher dimensional weight space, and so a second extension aims to help escape from any such minima by using less localised gradient estimates. Analytic gradients are perfectly local, measuring the slope at an infinitesimal point. Standard numerical gradient descent approximates this by measuring the gradient between finitely but closely spaced point. However for very non-smooth surfaces, such as fractals, local gradients are of little use and larger steps should be taken to escape from very small minima.

Two methods were tested to do this. Firstly, a simulated annealing search [20] (SA) and secondly, gradient descent searches using coarse gradient estimates, obtained by sampling points 2000 times further away from the current solution than used by the standard optimiser (GDx2000). Limabeam searches of most forms are computationally expensive to run (e.g. 400 days of 3GHz core time for the AMI test set) so these alternatives to basic gradient descent were used to give just an indication of alternative methods rather than an exhaustive search.

Results are shown in table 3. All runs here use Discriminative LIMA-HMM and MLLR training on either ground truth or SPR-TDOA decodes, and in TDOA model space.

Table 3: *Extended Limabeam results. 'gnd'=MLLR adaptation performed using ground truth data; 'hyp'=MLLR adaptation performed on decoded hypotheses. GD=standard gradient descent search; SA=simulated annealing search; GDx2000=coarse gradient descent search.*

search	MLLR	WER	S	D	I
GD	gnd	48.2	25.0	21.1	2.0
GD	hyp	59.3	36.4	19.8	3.1
SA	gnd	49.9	29.2	19.0	1.7
SA	hyp	63.0	39.0	21.3	2.6
GDx2000	gnd	41.8	24.0	15.8	2.0
GDx2000	hyp	58.7	37.6	18.0	3.1

The hyp-based experiments give marginal improvements over TDOA, (0.1% abs. for GD, no improvement for SA, and 1% rel. for GDx2000). Ground-truth LIMA-MLLR results give much more impressive improvement (19% relative, still not as large as in [9]) than ground truth TDOA-MLLR, suggesting if sufficient Lima-processed per-speaker training data was available then such improvements would also occur.

6. Conclusion

No significant WER gains for found on AMI with standard Limabeam. However by extending with discriminative objective and coarse gradient descent we obtained a 1% relative improvement, and a suggestion from gnd MLLR that larger gains up to 19% over gnd-MLLR standard Limabeam would be available given large per-speaker training data.

Unusual shapes formed by arbitrary parameter values emphasise there is more to beamforming than shown in traditional angular responses plots under far-field assumptions. For AMI, the space where speakers sit is susceptible to dominating near-field effects, which can produce non-radially symmetric local minima in both beamformer receptive fields and likelihood maps. Even conservatively quantising each weight to 10 possible values, gives a 10^{80} sized search space – comparable to the number of atoms in the universe – and impossible to search with any current computer. The space contains all known linear beamformers and many more. So any Limabeam-like search is heuristic. Adapting the coarseness of gradient descent to better fit intuition about minima distribution gives a small WER improvement. This suggests that future work could quantify such prior knowledge and use it to create custom search algorithms to better exploit it.

7. References

[1] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al., "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*, pp. 28–39. Springer, 2006.

[2] Charles Fox, Yulan Liu, Erich Zwyssig, and Thomas Hain, "The Sheffield Wargames Corpus," *Proceedings of Interspeech.*, 2013.

[3] Robert J Mailloux, *Phased array antenna handbook*, Artech House Boston, 2005.

[4] Ralph O Schmidt, "Multiple emitter location and signal parameter estimation," *Antennas and Propagation, IEEE Transactions on*, vol. 34, no. 3, pp. 276–280, 1986.

[5] Jack Capon, "High-resolution frequency-wavenumber spectrum analysis," *Proceedings of the IEEE*, vol. 57, no. 8, pp. 1408–1418, 1969.

[6] Jacob Benesty, Jingdong Chen, Yiteng Arden Huang, and Jacek Dmochowski, "On microphone-array beamforming from a mimo acoustic signal processing perspective," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 3, pp. 1053–1065, 2007.

[7] Felicia Lim, Mark RP Thomas, and Patrick A Naylor, "Mintformer: A spatially aware channel equalizer," in *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*. IEEE, 2013, pp. 1–4.

[8] Michael L Seltzer, *Microphone array processing for robust speech recognition*, Ph.D. thesis, Carnegie Mellon University, 2003.

[9] Michael L Seltzer, Bhiksha Raj, and Richard M Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 489–498, 2004.

[10] Michael L Seltzer and Richard M Stern, "Subband likelihood-maximizing beamforming for speech recognition in reverberant environments," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 6, pp. 2109–2121, 2006.

[11] Michael L Seltzer and Richard M Stern, "Parameter sharing in subband likelihood-maximizing beamforming for speech recognition using microphone arrays," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 1, pp. I–881.

[12] Dominik Raub, John W McDonough, and Matthias Wölfel, "A cepstral domain maximum likelihod beamformer for speech recognition.," in *Interspeech*, 2004.

[13] Kshitiz Kumar and Richard M Stern, "Maximum-likelihood-based cepstral inverse filtering for blind speech dereverberation," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4282–4285.

[14] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 285–290.

[15] Hynek Hermansky, "Perceptual linear prediction (PLP) analysis of speech," vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[16] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, XA Liu, G. Moore, J. Odell, D. Ollason, D. Povey, et al., "The HTK book," 2006.

[17] National Institute of Standards and Technology (NIST), *Speech Recognition Scoring Toolkit (SCTK) Version 2.4.0.*, web resource: http://www.itl.nist.gov/iad/mig/tools,, 2010.

[18] X Anguera, "Beamformit, the fast and robust acoustic beamformer. in http://www. icsi. berkeley. edu/xanguera," 2006.

[19] Michael S Brandstein and Harvey F Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*. IEEE, 1997, vol. 1, pp. 375–378.

[20] William L Goffe, "Simann: a global optimization algorithm using simulated annealing," *Studies in Nonlinear Dynamics and Econometrics*, vol. 1, no. 3, pp. 169–176, 1996.