A Posterior Approach for Microphone Array Based Speech Recognition

Dong Wang[1], *Ivan Himawan*[1,2], *Joe Frankel*[1], *Simon King*[1]

[1] Centre for Speech Technology Research, University of Edinburgh, UK
[2] Speech and Audio Research Laboratory, Queensland University of Technology, Brisbane, Australia
dwang2@inf.ed.ac.uk; i.himawan@qut.edu.au; joe@cstr.ed.ac.uk; Simon.King@ed.ac.uk

Abstract

Automatic speech recognition (ASR) is difficult in environments such as multiparty meetings because of adverse acoustic conditions: background noise, reverberation and cross-talk. Microphone arrays can increase ASR accuracy dramatically in such situations. However, most existing beamforming techniques use time-domain signal processing theory and are based on a geometric analysis of the relationship between sources and microphones. This limits their application, and leads to performance degradation when the geometric properties are unavailable, or heterogeneous channels are used. We present a new posterior-based approach for microphone array speech recognition. Instead of enhancing speech signals, we enhance posterior phone probabilities which are used in a tandem ANN-HMM system. Significant improvements were achieved over a single channel baseline. Combining beamforming and our method is significantly better than beamforming alone, especially in a moving speakers scenario.

Index Terms: speech recognition, microphone array, beamforming, tandem approach

1. Introduction

Extending ASR to the meetings domain is challenging because of adverse acoustic conditions. Headset microphones can alleviate this problem, but is inconvenient for users. Microphone arrays are less intrusive and can significantly improve recognition accuracy, via noise suppression [1, 2, 3, 4, 5] and directionality enforcement [6, 7].

Currently, most array processing methods operate on the acoustic signals, and assume that SNR enhancement will lead to ASR error reductions. Classical array processing relies on the time-delay patterns of the wave front reaching every sensor in the array, and applies delay compensation on each channel to *beamform* the array to the desired direction, thus enhancing the SNR. Different noise patterns need different beamformers [4].

Despite the differences in design and implementation, all existing beamformers can be called *acoustic beamformers* because the operate on the acoustic signals. Although powerful and efficient, acoustic beamformers can suffer severe performance degradation in some circumstances, especially when the time-delay is hard to accurately estimate, or when the array channels are heterogeneous. The first scenario happens when speakers move whilst talking, in which case even cross-powerspectrum phase analysis [11] may fail [6]. The second scenario happens when channels with different physical properties coexist in the array system, e.g., both lapel and distant microphones, or if some channels are missing.

This paper presents a novel posterior-based approach for array processing. This is motivated by the tandem ANN-HMM hybrid framework proposed by Hermansky et al. [12], in which posterior probability based features, or PPFs, obtained from a multi-layer perceptron (MLP), are used to augment the conventional acoustic features of a standard HMM-based ASR system. In our method, the posterior probabilities of each frame are calculated via the MLP for each microphone channel, and then accumulated over all channels. These accumulated posteriors are used as PPFs. Compared to acoustic beamforming (which enhances speech signals), the new approach directly enhances posterior probabilities, so might be called *posterior beamforming*, in the sense of channel selection and accumulation. Our approach is an *intermediate* combination approach, somewhere between acoustic beamforming and hypothesis integration (e.g., systems that use ROVER [2]). Posterior beamforming has several potential advantages: (1) time-delay estimation is not required, assuming the posteriors change smoothly; (2) heterogeneous channels can be combined easily; (3) channel selection is more meaningful based on the posteriors. Moreover, posterior beamforming and acoustic beamforming are complementary, as we will demonstrate

We first present the posterior probability features and how they are combined with conventional acoustic features. Then we describe posterior beamforming and hybrid acoustic-posterior beamforming. The results of experiments on the MC-WSJ-AV 8-channel array corpus are provided to demonstrate the effectiveness of our method.

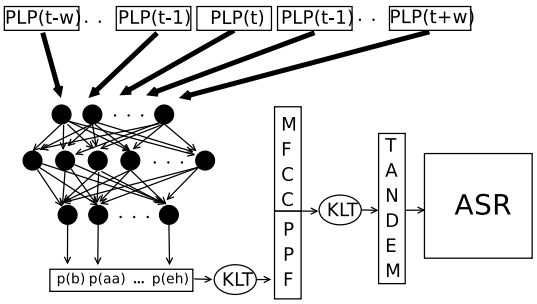

Figure 1: Generating PPFs and Tandem features. In our experiments, the MLP network uses PLP input features over a 9-frame window (351 input units), has 2000 hidden units and 46 output units corresponding to the 46 phones used in the ASR system.

2. Posterior probability based features

For classification problems, a decision based on posterior probability maximisation is optimal in the sense of discrimination, and leads to minimum classification errors. Here we follow the tandem framework introduced by Hermansky et al. [12] in which framewise phone class posterior probabilities are trans-

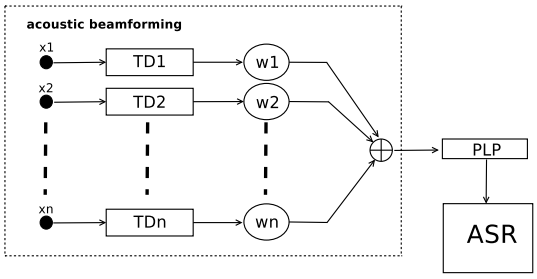

Figure 2: An array-based ASR system using an acoustic delay-sum beamformer. x1, x2, .. represent microphone elements in the array, and the TD blocks represent time delay compensation for each channel. w1, w2, ... are weights applied to the channels.

formed into PPF, via a logarithm and Karhunen-Loeve transform (KLT). In our system, the PPFs are appended to MFCCs, creating so called *Tandem features*, then passed to a conventional HMM-based recogniser (Figure 1). With PPFs or Tandem features, channel combination is straightforward at the feature level, since posterior probabilities can be simply summed additive: this is the basis of our method.

3. Posterior-based array processing

3.1. Posterior beamforming

Figure 2 shows a traditional array-based ASR system using a delay-sum beamformer, which first compensates each channel with proper time-delay, and then accumulates the speech signal to generate enhanced speech. The enhanced speech sounds much clearer to listeners, and performs much better in ASR, than any single-channel signal.

In the posterior-based approach (Figure 3), we do not accumulate speech signals, but rather posterior probabilities. Posterior beamforming does not consider any geometric properties of the environment and requires no physical similarity among channels. Posterior beamforming obviously has some shortcomings compared with acoustic beamforming; e.g., no use is made of the SNR enhancement possible if working in the time domain. Each MLP operates on an unenhanced signal. Posterior beamforming can not really beamform the array to a

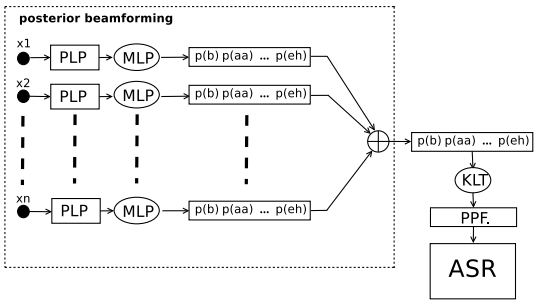

Figure 3: An array-based ASR system using the posterior accumulation approach. x1,x2, .. represent microphone elements in the array, and $p(b)$, $p(aa)$, .. represent the raw posterior probabilities of this frame belonging to phone b, aa, etc. The posterior probabilities from all channels are accumulated and averaged. A logarithm and KLT transforms them into PPFs.

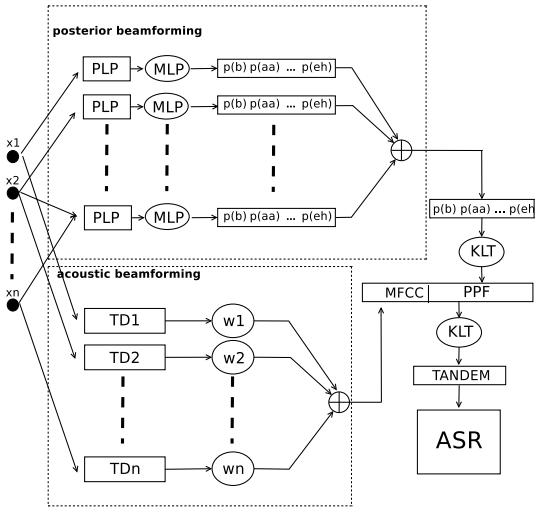

Figure 4: The diagram of the parallel approach to hybrid acoustic-posterior breamfoming. The acoustic beamformed audio is used to generate the acoustic components in the Tandem feature, and the posterior beamformed PPFs form the posterior components.

specific direction in the way that acoustic beamforming does, unless super-directive microphones and channel selection are used.

3.2. Hybrid acoustic-posterior beamforming

Acoustic and posterior beamforming appear to have complementary properties, making combination, either serially or in parallel, into a hybrid system an obvious next step.

The serial hybrid processes the beamformed signal as just another channel, and combines its posteriors with those of the other (real) channels as in Figure 2. This approach did not improve performance very much. The parallel hybrid (Figure 4) uses the beamformed signal to provide the conventional acoustic features (MFCCs in our case) to which the posterior beamformer appends PPFs to make tandem features. This method performed the best in our experiments.

4. Experiments

4.1. Corpus

Experiments were performed on the MC-WSJ-AV corpus [1] which contains sentences from the wsjcam0 development set and evaluation sets recorded by new talkers in instrumented meeting rooms at several sites including CSTR, IDIAP and NTO. The speech was recorded simultaneously with individual headset microphones (IHM), lapel microphones (Lapel), a single channel distant microphone (SDM) and two 8-channel circular microphone arrays. In our experiments, we tested two scenarios: the stationary-talker scenario in which talkers speak from six static positions, and the moving-talker scenario in which talkers move as they speak. For the stationary-talker scenario, we used one of the circular arrays, which is relatively small (diameter 20cm), while for the moving-talker scenario, we used an *ah hoc array* which contains 8 microphones scattered around the table in unknown positions (Figure 5).

The standard wsjcam0 training set was used for training the MLP and HMMs, and the wsjcam0 development set was

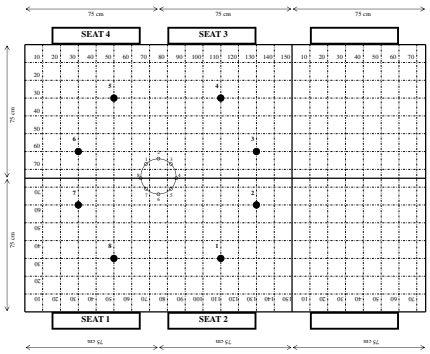

Figure 5: The microphone placements used in our experiments. The small circle represents the calibrated 8-microphone array and the 8 black dots represent the ad-hoc array.

Word error rate (WER) in %				
	IHM	Lapel	SDM	A-beam
MFCC	7.8	13.1	65.9	48.2
PLP	7.7	12.3	65.0	44.2

Table 1: Performance of the baseline system with MFCCs or PLPs under various conditions: individual headset microphone (IHM), lapel microphone (Lapel), single distant microphone (SDM) or acoustic beamforming (A-beam).

used to tune decoding parameters. Evaluation was performed on the MC-WSJ-AV evaluation data which were recorded in CSTR, University of Edinburgh. For the stationary-talker scenario, there are 188 sentences in the evaluation set, and for the moving-talker scenario, 179 sentences were used. Each scenario contains 5 talkers.

Throughout the experiments, the evaluation was on a 5k-vocabulary task. The dictionary used is generated from the one developed for the AMI NIST RT05S system [13] and the standard MIT-Lincoln Labs 5k Wall Street Journal trigram language model was used for decoding. Neither channel- nor speaker-adaptation was performed, though speaker-based CMN/CVN was applied wherever acoustic features appear. The HTK toolkit [14] was used for HMM training and decoding, and the tool *QuickNet* from ICSI for manipulating MLPs.

4.2. Stationary-speaker scenario

4.2.1. Baseline system with acoustic beamforming

Our baseline system uses conventional acoustic features and acoustic beamforming. We tried both MFCC and PLP-based features, each consisting of 13 cepstral coefficients (including zero-order coefficient C0), plus first and second derivatives, leading to a 39-dim vector. The blind delay-sum acoustic beamformer estimates the time-delay of each channel to a reference channel by peak detection on the general cross-correlation (GCC) [11].

Experimental results are presented in Table 1. A substantial and significant WER reduction (about 20% absolute) was obtained by acoustic beamforming over the SDM condition. The PLP-based system outperformed MFCCs in all conditions, so was selected for our baseline system in all remaining experiments.

WER					
	IHM	Lapel	SDM	A-beam	P-beam
Baseline	7.7	12.3	65.0	44.2	-
MPPF	7.7	18.1	59.8	52.8	58.0
PPPF	8.8	15.8	59.3	50.7	57.1

Table 2: Performance of posterior beamforming with an ASR system based on 20-dim PPFs under various conditions. *MPPF* denotes PPF features generated from a MLP with MFCC features as the input, and *PPPF* denotes PPF features generated from a MLP with PLP input. *A-beam* denotes acoustic beamforming. *P-beam* denotes posterior beamforming.

4.2.2. Posterior beamforming

Sentences in the wsjcam0 training corpus were phone labelled by forced alignment, and then used to train the MLP. 70% phone accuracy was obtained on the cross-validation set of 700 sentences held out from the training set. On our task, KLT reduction of the log posteriors to the first 20 principal components gave the best performance. We tried both PLP and MFCC features as input to the MLP. Table 2 presents results.

From Table 2, we can observe that the posterior beamforming does significantly improved the recognition performance compared to the SDM condition, but it is not as good as acoustic beamforming. The PPF-based systems work significantly better than the baseline in the SDM condition, suggesting that PPFs are more robust than acoustic features in noisy environments, which is consistent with Hermansky's results [12]. PLP-based PPFs work better than the MFCC-based ones, so we use PLP-based PPFs in all subsequent experiments.

4.2.3. Hybrid acoustic-posterior beamforming

As mentioned in section 3.2, we can combine acoustic beamforming and posterior beamforming into a hybrid system. This section reports experiments on this hybrid approach, using Tandem features (Section 2).

In our experiments, 46-dim PPFs are generated from a PLP-based MLP, which then are concatenated with 39-dim MFCCs, leading to 85-dim vectors. After KLT, the first 30 principal components are retained, giving 30-dim tandem features. Hybrid beamforming is implemented by using the acoustic beamformed signal to generate the MFCCs and the 8-channel accumulated posteriors to generate the PPFs.

Table 3 gives the results of various beamforming approaches with systems based on tandem features. The system based on tandem features performs significantly better than the baseline in the SDM condition, although a little bit worse in the IHM and Lapel conditions. The hybrid system is better than acoustic beamforming alone.

4.3. Moving-speaker scenario

In the moving-speaker scenario, acoustic beamforming is expected to be error-prone because time-delay estimation in this time-delay-varying condition is more difficult. For the posterior approach, the impact is expected to be less, since this approach does not rely on time synchronisation.

Results are shown in Table 4. Compared to the stationary-speaker scenario, the performance in all conditions is worse. However, whilst the baseline system WER is increased by nearly 8%, the tandem system is less affected. The hybrid beamforming approach only suffers a 3.3% increase in WER

	WER					
	IHM	Lapel	SDM	A-beam	P-beam	AP-beam
Baseline	7.7	12.3	65.0	44.2	-	-
Tandem	8.9	14.0	56.1	43.6	55.5	**42.5**

Table 3: The experimental results of the parallel hybrid beamforming approach based on tandem features. For the Tandem system, in the conditions *IHM*, *Lapel* and *SDM*, both the MFCC and PPF components of the tandem features come from the same single audio channel. In *A-beam*, both these two components come from acoustic beamformed speech. In *P-beam*, the MFCC components come from the SDM channel, and the PPFs come from posterior accumulation over the 8 array channels. Finally, *AP-beam* represents the hybrid acoustic-posterior beamforming, for which the acoustic beamformed speech generates the MFCCs for the tandem features, and posterior accumulation generates the PPFs.

	WER			
	SDM	A-beam	P-beam	AP-beam
Baseline	72.7	52.0	-	-
Tandem	65.0	48.3	63.0	**45.8**

Table 4: Results for the moving-speaker scenario. The meaning of each column is the same as Table 3.

and outperforms the acoustic beamforming system by 6.2%. A pairwise t-test shows that the performance improvements of tandem features and hybrid beamforming over acoustic beamforming is statistically significant ($p < 0.01$).

5. Conclusions

In this paper, we presented a new posterior beamforming approach for microphone array-based speech recognition. In this approach, posterior probabilities calculated from each array channel are accumulated to form robust posterior probability-based features. Our experimental results demonstrate a significant performance improvement over baseline in the SDM condition. Posterior beamforming can be combined with conventional acoustic beamforming, leading to a hybrid acoustic-posterior beamforming approach which is significantly better than acoustic beamforming alone. We believe that MLPs trained with normal speech signals are far from optimal when used with beamformed signals, so a proper adaptation scheme should be used. This is future work.

6. Acknowledgements

DW holds a fellowship from the Marie Curie Early Stage Research Training scheme "Edinburgh Speech Science and Technology." IH is a visiting researcher at the Centre for Speech Technology Research, University of Edinburgh. JF was funded by Scottish Enterprise under the Edinburgh-Stanford Link programme. SK holds an EPSRC Advanced Research Fellowship. We are grateful to Mike Lincoln for help with the corpora we used in these experiments.

7. References

[1] M. Lincoln, I. McCowan, J. Vepa, and H. K. Maganti, "The multi-channel wall street journal audio visual corpus (mc-wsj-av): specification and initial experiments," in *ASRU 2005*, December 2005, pp. 357–362.

[2] S. M. Chu, E. Marcheret, , and G. Potamianos, "Automatic speech recognition and speech activity detection in the chil smart room," in *MLMI 2005*, 2005.

[3] H. Francois, D. Pearce, and J. Rex, "Dual-microphone robust front-end for arm's-length speech recognition," in *2006 International Workshop on Acoustic Echo and Noise Control*, Paris, France, September 2006.

[4] J. Bitzer, K. Simmer, and K. Kammeyer, "Multimicrophone 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.

[5] D. Giuliani, M. Matassoni, M. Omologo, and P. Svaizer, "Hands free continuous speech recognition in noisy environment using a four microphone array," in *ICASSP 95*, vol. 1, Detroit, MI, USA, May 1995, pp. 860–863.

[6] H. K. Maganti and D. Gatica-Perez, "Speaker localization for microphone array-based asr: the effects of accuracy on overlapping speech," in *Proceedings of the 8th international conference on Multimodal interfaces (ICMI 06)*, Banff, Alberta, Canada, November 2006, pp. 35 – 38.

[7] D. C. Moore and I. A. McCowan, "Microphone array speech recognition: experiments on overlapping speech in meetings," in *ICASSP 2003*, vol. 5, Hong Kong, China, April 2003, pp. 497–500.

[8] H. Silverman, "Some analysis of microphone arrays for speech data acquisition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, pp. 1699–1712, December 1987.

[9] M. Seltzer, B. Raj, and R. Stern, "Speech recognizer-based microphone array processing for robust hands-free speech recognition," in *ICASSP 02*, vol. 1, Orlando, FL, USA, May 2002, pp. 897–900.

[10] D. Raub, J. McDonough, and M. Wofel, "A cepstral domain maximum likelihood beamformer for speech recognition," in *Interspeech 04*, Jeju Island, Korea, October 2004, pp. 817–820.

[11] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Signal Processing*, vol. 24, no. 4, pp. 320 – 327, August 1976.

[12] H. Hermansky, D. P. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP 2000*, Istanbul, Jun. 2000.

[13] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, I. McCowan, D. Moore, V. Wan, R. Ordelman, and S. Renals, "The 2005 ami system for the transcription of speech in meetings," in *MLMI 2005*, Edinburgh, UK, July 2005.

[14] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. amnd Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*, Microsoft Corp. and Cambridge University Engineering Department, 2006.