MAKING THE MOST FROM MULTIPLE MICROPHONES IN MEETING RECOGNITION

Andreas Stolcke

Speech Technology and Research Laboratory
SRI International, Menlo Park, CA, USA
and
International Computer Science Institute
Berkeley, CA, USA
stolcke@icsi.berkeley.edu

ABSTRACT

The use of multiple distant microphones has been widely studied for meeting recognition. The two most widely used approaches are 1) combination at the signal level, via blind beamforming, followed by recognition of a single enhanced audio signal, and 2) independent, logically parallel recognition of the multiple audio sources followed by hypothesis-level combination. In this paper we investigate how these two approaches compare for state-of-the-art recognition systems applied to meeting data from the two most recent NIST Rich Transcription evaluations. Our results show that beamforming is the superior approach, giving more accurate results while being inherently less computationally demanding. We then propose a hybrid approach that leverages both beamforming and signal-level diversity for system combination, and show that this approach gives gains over either of the old methods.

***Index Terms*—** Meeting recognition, blind beamforming, system combination.

1. INTRODUCTION

Automatic speech recognition (ASR) from natural multi-person meetings remains one the most difficult recognition tasks formally evaluated by NIST, the American National Institute of Standards and Technology. This is especially true for recognition from distant (e.g., tabletop) microphones. One set of techniques making recognition possible with accuracies approaching those of easier tasks (like telephone speech and broadcast audio) is the use of multiple microphones. Two main approaches have been used for recognition from multiple distant microphones (MDM). The first approach [1, 2] combines information at the signal-level: the time-delay of arrival between the audio channels is locally estimated and the signals are summed after appropriate alignment. As a result the speech signal is enhanced while noise and reverberation are attenuated. A new audio signal is generated that can be recognized by a single recognition run.

An alternative approach is to separately recognize the individual signals, and then combine their results at the hypothesis level (i.e., through confusion network combination (CNC) [3, 4]), which amounts to a weighted voting over individual recognized words. Because recognition decoding is typically slower than beamforming, this approach is inherently more computationally costly, and scales poorly with the number of input channels. The scaling issue can be mitigated at some cost to accuracy by selecting a limited number of "good" channels from among those available [5].

Wölfel, in work presented in a series of papers, has conducted thorough comparisons of these two approaches, while developing variants and refinements of the second, multi-channel approach [6, 5, 7]. The Wölfel results were all based on *lecture meetings*, a specialized form of meeting data dominated by a single speaker and with microphone configurations that are not typical for more common conference-table meetings. Second, owing to the inherent difficulty of the data, the recognition systems exhibited very high word error rates (typically between 40 and 60%), whereas recent NIST meeting evaluations have resulted in substantially lower error rates, of between 25 and 35% [8]. It is well-known that the effectiveness of ASR algorithms (e.g., unsupervised adaptation) may well change as a function of error rate. Also, the reported gains from blind beamforming have been significantly smaller in the earlier studies than what we typically find in MDM conference recognition (possibly due to the special difficulties of the lecture task, where the speaker location might vary substantially). For all these reasons, we found it worthwhile revisiting the question of the relative effectiveness of beamforming versus multiple recognition. In the process we developed additional processing schemes that combine aspects of both approaches, and that resulted in additional accuracy gains.

2. METHOD

2.1. Data

Our data is drawn from the two most recent NIST Rich Transcription (RT) conference meeting evaluation sets, RT-07 and RT-09, as well as from the RT-07 lecture meeting evaluation set (RT-09 did not include any lecture data). Note that each set consists of excerpts of longer meetings, but only the regions defined for evaluation purposes are processed by our systems; while using data outside those regions is legal, little or no benefit was found when doing so. Statistics of these test sets are summarized in Table 1. Note that the number of speakers for lecture meetings is misleading; most of the speech originates from a single speaker, the lecturer.

2.2. Microphone selection

While the RT evaluations define several microphone conditions, here we are only concerned with the MDM condition. For comparison we also give results on the single distant microphone (SDM) condition, which uses a single NIST-defined, "centrally located" microphone as the only input to the recognizer.

Table 1. Comparison of key NIST RT evaluation set properties

Meeting genre	RT-07 lecture	RT-07 conference	RT-09 conference
No. meetings	32	8	7
Avg./max. no. spkrs/mtg	4.41 / 7	4.38 / 6	5.43 / 11
No. of mics per meeting	3-4	3-16	7-16
Total duration	164 mins	180 mins	181 mins
Total speech duration	138 mins	156 mins	162 mins
Total no. of words	25239	36800	36734

As shown in Table 1, the meetings differ in the number (as well as the type) of microphones available. All algorithms discussed here are capable of dealing with a variable number of microphones, but the processing time for approaches based on multiple recognition scales linearly with the number of input channels (since the total processing time is dominated by the ASR). As an expedient, we limit the number of microphones to a maximum of four, sampling from the available ones at constant steps in their nominal order. For example, for meetings with seven microphones, we would only use microphones numbered 1, 3, 5, and 7. This sampling procedure is oblivious to the physical location or type of the microphones. It was known that all microphones were located on the conference tabletop, and we guessed that channel numbers would roughly correspond to physical proximity. For the meetings collected by the Augmented Multi-party Interaction (AMI) consortium, the selection procedure yields two diametrically opposed microphones each from the two 8-channel circular arrays used for meeting recordings. (Following the MDM task specification, the array microphones also available in the lecture rooms were not used in the present study.)

2.3. Error metrics

We report word error rate (WER) as our metric, computed in the standard way as the total number of incorrectly recognized or deleted words, divided by the total number of reference words. However, because conference meetings include a significant amount of overlapping speech (another property that distinguishes them from lectures), NIST introduced an additional free parameter in the WER computation: the maximum allowed number of overlapping speakers. An "overlap-N" WER includes all reference speech segments with up to N speakers talking simultaneously. Since our recognition system does not attempt to model overlapping speech, we will be interested mostly in overlap-1 WER. but we also report overlap-4 WER for completeness. (Computing overlap-N WER becomes prohibitively expensive for larger values of N, and more than four overlapping speakers are very rare even in conference meetings).

3. RECOGNITION SYSTEM

3.1. Preprocessing

Prior to any other processing, the audio signals are individually Wiener-filtered using the ICSI-OGI-Qualcomm Aurora implementation [9] to suppress stationary noise. The wiener filtering (which includes a fast speech/nonspeech detector) runs in less than 0.03 times realtime (xRT) on a single 3.1GHz CPU core. Even though it has to be run for each input channel it is not a significant contributor to overall recognition time, especially on a multicore CPU.

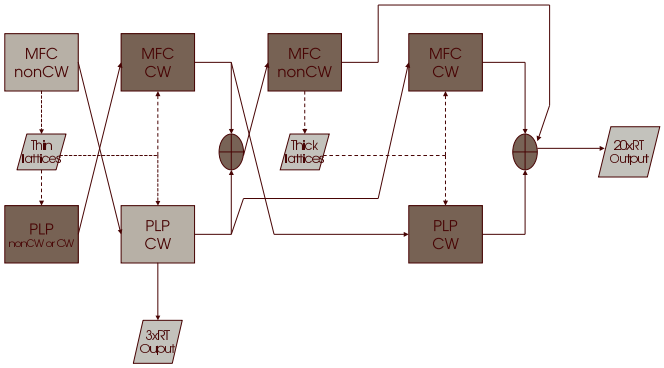

Fig. 1. SRI meeting recognition system. Rectangles represent decoding steps. Parallelograms represent decoding output (lattices or 1-best hypotheses). Solid arrows denote the passing of hypotheses for adaptation or output. Dashed lines denote the generation or use of word lattices for decoding. Crossed ovals denote confusion network system combination steps.

3.2. Beamforming

We use Anguera's freely available blind beamforming implementation *BeamformIt*, version 2.0 [10] to combine multiple audio channels into a single signal. This step runs in less than 0.01xRT on a single 3.1GHz CPU, relative to the combined duration of the input waveforms. Therefore, this step will likewise only contribute a small fraction of the total runtime of the recognition system.

Beamforming can substantially benefit speech segmentation as well as recognition, therefore performing the blind beamforming step prior to segmentation is advisable. On the other hand, beamforming occurs after Wiener filtering, since the noise filtering also enhances beamforming [1].

3.3. Segmentation and clustering

The audio stream is segmented into speech and nonspeech segments by decoding with a two-class GMM acoustic model based on standard Mel frequency cepstral coefficient (MFCC) features. The HMM structure imposes some minimum duration constraints and penalizes transitions between speech and nonspeech classes. The resulting speech segments are combined and padded to satisfy some duration constraints: no pauses longer than 0.4 s, no segments longer than 60 s, and 0.06 s nonspeech at the beginning and end of segments.

Prior to recognition, the speech segments from a given meeting undergo agglomerative clustering based on acoustic similarity, following a method previously developed for broadcast news recognition [11]. This step results in pseudo-speaker clusters that form the units for cepstral normalization and unsupervised adaptation in the recognition system.

To facilitate the experiments and their analyses, we introduce an additional simplification. In the multiple recognition approach, the segmentation would typically produce quite different results depending on the microphone used (for example, a speaker far removed from the microphone might not be detected well, possibly resulting in all of his or her words being deleted). Further, system combination of ASR outputs using diverging segmentations is cumbersome unless suboptimal algorithms (less than full confusion network combination) are used [8]. Therefore, and because beamforming is fast compared to recognition, we assume that segmentation is always run on the beamformed signal, and a common segmentation used, even if subsequent recognition employs the original, multiple, available

audio signals. This approach is not only practical, it also benefits the multiple-recognition approach, because all recognizers then use higher-quality segmentation. More importantly, the procedure simplifies and focuses the analyses because result differences can be attributed to the recognizer proper, as opposed to different qualities of speech activity detection.

3.4. Recognition system

The ASR system for all our experiments is the meeting recognition system jointly developed by SRI and ICSI for the distant microphone, conference meeting conditions in the NIST RT-07 and RT-09 meeting recognition evaluations [12]. As depicted in Figure 1, the recognizer performs a total of eight decoding passes with alternating acoustic front-ends: one based on telephone-band MFCCs augmented with multilayer-perceptron (MLP) features, and one based on full-band perceptual linear prediction (PLP) features. Acoustic models are cross-adapted during recognition to output from previous recognition stages, and the output of the three final decoding steps is combined via confusion networks. The MFCC models are trained on telephone conversations and then adapted to about 200 hours of meeting data. The PLP models, by contrast, are originally trained on broadcast data. Various discriminative techniques are used in training and adaptation [13]. Language models (LMs) consist of a mixture of genre-specific models for meeting transcripts, telephone conversations, broadcast news, web data, and (for lecture recognition) conference proceedings and lecture transcripts. The recognition system performs vocal tract length normalization, and cepstral mean and variance normalization, and in later recognition passes, unsupervised acoustic adaptation using CMLLR and MLLR on the pseudo-speaker clusters generated by the waveform-clustering step described earlier. Processing time for recognition on a single audio stream is about 3.8 times real time on an 8-core, 3.1-GHz Intel-CPU server, including the processing for waveform segmentation and clustering.

3.5. Acoustic model training

The meeting training data (approximately 200 hours in duration) is prepared for training in a manner that is consistent with single-microphone processing. All recording channels from all training meetings (including close-talking microphones, which are not allowed in SDM and MDM test condition) are pooled. This has the effect that speech is modeled in a range of recording conditions (from close-talking to most distant). Regions of overlapping speech are excluded from training. The same recognition models are used in all test configurations. Therefore, training audio data is noise-filtered, but not beamformed. Thus, model training is actually better matched to the multiple-recognition approach than the beamforming approach.

4. EXPERIMENTS AND RESULTS

4.1. Comparison of standard approaches

All baseline results, as well as results with multiple-recognition and beamforming are listed in Table 2. For reference, the table includes the results with a single microphone (SDM), yielding the worse results, and the MDM processing performed by our evaluation system, based on all microphone channels and beamforming. The relative difference (in overlap-1 WER) between these two systems is between 12 and 20% depending on the test set.

Table 2. WER (%) results with various distant microphone processing methods. All results on nonoverlapping (overlap-1) speech. BF = beamforming, MR = multiple recognition.

Genre	RT-07 lecture		RT-07 conference		RT-09 conference	
Overlap	1	4	1	4	1	4
Sngle mic (SDM)	50.6	54.5	33.1	45.2	41.3	49.9
BF, ≤ 4 mics	44.6	49.1	28.1	40.8	37.2	45.5
MR, ≤ 4 mics	47.9	52.5	31.6	45.7	39.7	49.6
BF + MR, ≤ 4 mics	44.0	48.8	28.2	41.5	36.8	45.9
BF, all mics	44.6	49.1	26.5	39.3	33.6	42.7

Table 3. WER (%) results with leave-one-out beamforming (LOO-BF). BF = beamforming. Number of microphones ≤ 4.

Genre	RT-07 lecture		RT-07 conference		RT-09 conference	
Overlap	1	4	1	4	1	4
BF	44.6	49.1	28.1	40.8	37.2	45.5
LOO-BF	42.8	47.9	28.1	41.5	36.4	45.4
BF + LOO-BF	42.7	47.7	27.5	40.8	36.2	45.2

The 4-microphone beamforming achieves only between 10 and 15% relative improvement in overlap-1 WER, showing that blind beamforming can effectively combine a fairly large number of channels for incremental gains.

Independent recognition and CNC of the same four channels, however, gives substantial less gain, generally less than half the relative gain of beamforming. We also tried to enhance the multiple-recognition approach by cross-adapting the acoustic models using preliminary hypotheses obtained from other channels. The results were worse, presumably because such cross-adaptation causes the different parallel systems to converge on similar hypotheses, which then tends to leave less room for improvement in the final combination.

We also tried combining the final outputs from multiple recognition (MR) with the output from beamforming (BF), as was suggested in [5]. This indeed gives small additional gains over beamforming alone, but the differences are small and inconsistent across test sets. (For RT-09, the overlap-4 WER is actually slightly worse as a result of the BF+MR combination.)

4.2. Leave-one-out beamforming

The initial experiments raise the question if beamforming and system combination approach can somehow be combined for additional gains. The question is how channels should be beamformed in multiple, distinct ways to achieve both good recognition from each beam, as well as to preserve diversity among the different recognition runs. For example one might partition the available audio channels. Here we choose a simple strategy that beamforms all N channels minus one, where the left-out channel is changed in a round-robin manner. This again yields N different signals, which are then recognized separately for eventual CNC. We call this "leave-one-out beamforming". It is depicted in Figure 2, contrasting it with the traditional multiple recognition and beamforming processing schemas.

The results appear in Table 3. The gains in overlap-1 WER over simple beamforming are between 0 and 4% relative. The results can be improved somewhat if the N-way beamformed system is added to the final CNC.

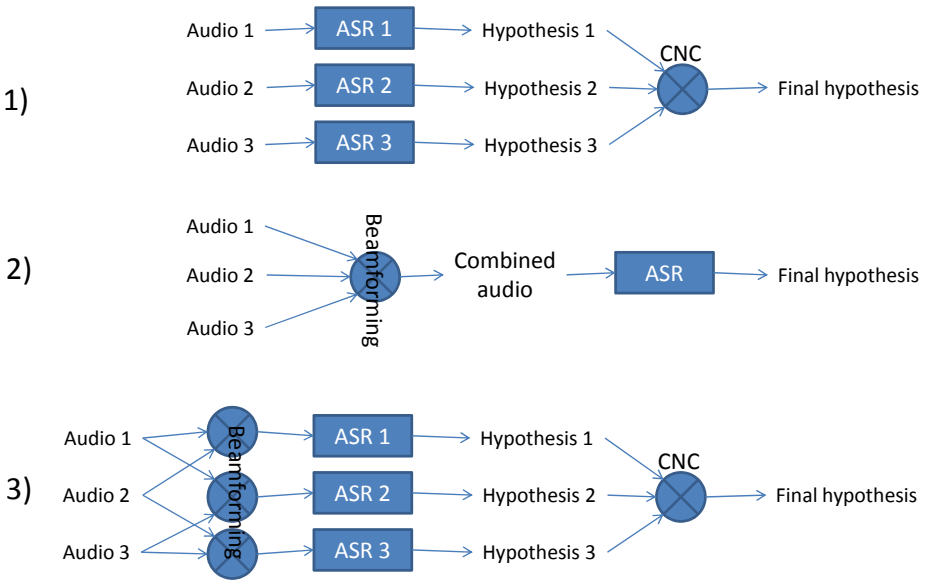

Fig. 2. MDM processing schemes. (1) multiple recognition, (2) beamforming, (3) leave-one-out beamforming with multiple recognition.

We again tried to improve the leave-one-out beamformed system further by introducing cross-adaptation between the parallel systems, but this, too, gave no additional win.

5. DISCUSSION AND CONCLUSIONS

We may summarize the results as follows. When comparing signal-level combination via beamforming versus system combination at the word-level, the beamforming approach is consistently and substantially superior. Because beamforming is also much faster than parallel recognition beamforming is clearly the recommended approach for leveraging multiple microphones for recognition. It is noteworthy that this conclusion is supported by results across different meeting genres (lectures and conferences), and for very different operating points (word error rates in the 20s, 30s, and 40s).

If runtime is not an issue and multiple parallel recognition runs are affordable, we find that the best strategy is to create multiple beamformed signals by leaving out one channel at a time. Adding recognition from the fully formed beam into the combination helps too.

6. ACKNOWLEDGMENTS

Thanks to Xavi Anguera for fruitful discussions and for making his beamformer software [10] available, and to Adam Janin and ICSI colleagues for discussions and assistance with the evaluation data.

7. REFERENCES

[1] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 15, pp. 2011–2022, Sep. 2007.

[2] T. Hain, L. Burget, J. Dines, G. Garau, V. Wan, M. Karafiat, J. Vepa, and M. Lincoln, "The AMI system for the transcription of speech in meetings", *in Proc. ICASSP*, vol. 4, pp. 357–360, Honolulu, Apr. 2007.

[3] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation, and system combination", *in Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[4] A. Stolcke, H. Bratt, J. Butzberger, H. Franco, V. R. Rao Gadde, M. Plauché, C. Richey, E. Shriberg, K. Sönmez, F. Weng, and J. Zheng, "The SRI March 2000 Hub-5 conversational speech transcription system", *in Proceedings NIST Speech Transcription Workshop*, College Park, MD, May 2000.

[5] M. Wölfel, , C. Fügen, S. Ikbal, and J. W. McDonough, "Multi-source far-distance microphone selection and combination for automatic transcription of lectures", *in Proc. ICSLP*, pp. 361–364, Pittsburgh, PA, Sep. 2006.

[6] M. Wölfel and J. McDonough, "Combining multi-source far distance speech recognition strategies: Beamforming, blind channel and confusion network combination", *in Proc. Interspeech*, pp. 3149–3152, Lisbon, Sep. 2005.

[7] M. Wölfel, "Channel selection by class separability measures for automatic transcriptions on distant microphones", *in Proc. Interspeech*, pp. 582–585, Antwerp, Aug. 2007.

[8] A. Stolcke, G. Friedland, and D. Imseng, "Leveraging speaker diarization for meeting recognition from distant microphones", *in Proc. ICASSP*, pp. 4390–4393, Dallas, Mar. 2010.

[9] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR", in J. H. L. Hansen and B. Pellom, editors, *Proc. ICSLP*, vol. 1, pp. 4–7, Denver, Sep. 2002.

[10] X. Anguera, "Beamformit (the fast and robust acoustic beamformer)", http://www.icsi.berkeley.edu/~xanguera/beamformit/, 2006.

[11] A. Sankar, F. Weng, Z. Rivlin, A. Stolcke, and R. R. Gadde, "The development of SRI's 1997 Broadcast News transcription system", *in Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 91–96, Lansdowne, VA, Feb. 1998. Morgan Kaufmann.

[12] A. Stolcke, K. Boakye, Özgür Çetin, A. Janin, M. Magimai-Doss, C. Wooters, and J. Zheng, "The SRI-ICSI Spring 2007 meeting and lecture recognition system", in R. Stiefelhagen, R. Bowers, and J. Fiscus, editors, *Multimodal Technologies for Perception of Humans. International Evaluation Workshops CLEAR 2007 and RT 2007*, vol. 4625 of *Lecture Notes in Computer Science*, pp. 450–463, Berlin, 2008. Springer.

[13] J. Zheng and A. Stolcke, "fMPE-MAP: Improved discriminative adaptation for modeling new domains", *in Proc. Interspeech*, pp. 1573–1576, Antwerp, Aug. 2007.