LEVERAGING SPEAKER DIARIZATION FOR MEETING RECOGNITION FROM DISTANT MICROPHONES

Andreas Stolcke[1,2] *Gerald Friedland*[2] *David Imseng*[2]

[1]SRI International, Menlo Park, CA, USA
[2]International Computer Science Institute, Berkeley, CA, USA
stolcke@speech.sri.com, {fractor,imseng}@icsi.berkeley.edu

ABSTRACT

We investigate using state-of-the-art speaker diarization output for speech recognition purposes. While it seems obvious that speech recognition could benefit from the output of speaker diarization ("Who spoke when") for effective feature normalization and model adaptation, such benefits have remained elusive in the very challenging domain of meeting recognition from distant microphones. In this study, we show that recognition gains are possible by careful post-processing of the diarization output. Still, recognition accuracy may suffer when the underlying diarization system performs worse than expected, even compared to far less sophisticated speaker-clustering techniques. We obtain a more accurate and robust overall system by combining recognition output with multiple speaker segmentations and clusterings. We evaluate our methods on data from the 2009 NIST Rich Transcription meeting recognition evaluation.

Index Terms— speech processing, speaker diarization, meeting recognition, rich transcription, system combination.

1. INTRODUCTION

Recognition of free-form, multiparty meetings remains one of the most challenging tasks that is formally studied and evaluated by the speech recognition community, especially when recognition is performed by microphones at a distance from, and shared by, the speakers. This scenario has been studied in a series of NIST evaluations on conference-style meeting data [1], but is also very relevant to what has been described as "broadcast conversation," such as talk and call-in shows. Along with automatic speech-to-text transcription (STT), NIST evaluates speaker diarization (i.e., speech detection and speaker tracking to answer the question "who spoke when").

It seems natural that STT systems should directly benefit from proper diarization. First, most recognizers perform more accurately and efficiently when applied to audio segments that have been trimmed of nonspeech regions (except for short inter-word pauses and a small amount of padding at the edges). Further, speech features are best normalized by speaker, and acoustic model adaptation usually works best on homogeneous speaker clusters.

However, for many reasons, the relationship between diarization quality and STT accuracy is not straightforward. For example, a diarization system that over-hypothesizes speech is not as bad as one that misses it, because the STT system provides another chance to classify nonspeech as such. Also, sub-clustering speakers with sufficient data might well be beneficial, as the STT system can then adapt to variable acoustic conditions (such as environmental noise or position relative to the microphone). The study in [2] found little or no correlation between the standard diarization error metric and word error recognition, and found that properly tuned pause-bridging and padding parameters were the most important factors when utilizing diarization system output for STT.

Table 1. Comparison of key NIST RT evaluation set properties

	RT-07	RT-09
No. meetings	8	7
Avg./max. no. of speakers per meeting	4.38 / 6	5.43 / 11
Total duration	180 mins	181 mins
Total speech duration	147 mins	164 mins
Total no. of words	35882	40110

For many years, as a matter of practical experience, our groups at SRI and ICSI have fielded both speech recognition and diarization systems that had state-of-the-art performance in NIST evaluations. Yet, in previous years, our STT system always worked best when based on a rather simple, ad-hoc diarization system (described below), and no gains were realized when coupling it with the output of our best diarization system. Undaunted, we pursued such a coupling again for the 2009 NIST Rich Transcription (RT-09) evaluation, using the techniques described here.

2. DATA AND METRICS

Our data is drawn from the two most recent NIST Rich Transcription (RT) conference meeting evaluation sets, RT-07 and RT-09. Note that each set contains 20- to 30-minute-long 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 doing so, for either diarization or recognition. Statistics of these test sets are summarized in Table 1.

Diarization and recognition were evaluated under several microphone conditions. The two conditions of interest here are *single distant microphone* (SDM) and *multiple distant microphones* (MDM). In both cases, microphones were placed on tables at which the meeting participants were sitting. Type, placement, and number (for MDM) of microphones was variable across meetings.

Diarization performance is measured by the NIST *diarization error rate* (DER), which is the total audio duration that is mistakenly classified as speech or nonspeech or assigned to the wrong speaker cluster, divided by the total speech duration. Similarly, STT *word error rate* (WER) is computed as the total number of incorrectly recognized or deleted words, divided by the total number of reference words. One additional parameter in error metrics for meeting recognition is the maximum allowed number of overlapping speakers. An "overlap-N" metric includes all reference speech segments with up to N speakers talking simultaneously.

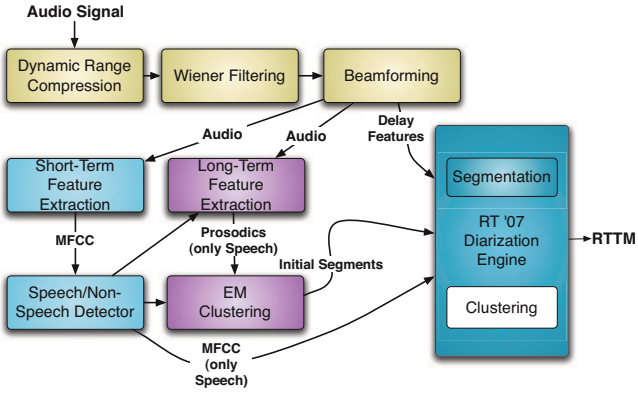

Fig. 1. The ICSI speaker diarization engine as configured for the RT-09 evaluation, MDM condition

3. SPEECH PROCESSING

Prior to diarization and recognition, the audio signal is processed as follows. In the SDM condition, a dynamic range compression is applied to normalize energy variance caused by different microphone distances. Then, the signal is Wiener-filtered [3]. In the MDM condition, the signal is then beamformed using BeamFormIt 2.0 [4].

3.1. Baseline speech detection and clustering

The standard speech detection and clustering process used for STT purposes was developed without regard to diarization metrics and optimized purely with respect to STT accuracy. First, the input is segmented into speech and nonspeech portions 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.

Second, the resulting speech segments from a given meeting undergo agglomerative clustering based on acoustic similarity, following a method previously developed for broadcast news recognition [5]. To define a distance between segments a single GMM is trained per meeting, with a separate mixture weight vector per segment. The distance between two segments is then defined as the weighted-by-counts increase in entropy of the mixture weight distribution due to clustering two segments. As a stopping criterion, a fixed limit on the increase in entropy can be chosen. However, probably due to the typical makeup of RT evaluation meetings, in past evaluations we had obtained best results by fixing the number of clusters at four, and we stuck to this choice.

3.2. ICSI diarization system

In contrast to the baseline approach, we also used the ICSI diarization system as developed for the RT-09 evaluation, which is a "proper" diarization system optimized for the error metric employed by NIST. The system architecture is depicted in Figure 1. From the audio, 19th-order MFCC features are extracted with a frame size of 30 ms and a step size of 10 ms. Speech activity regions are determined using a state-of-the-art speech/nonspeech detector [6]. The nonspeech regions are then excluded from the agglomerative clustering as described below.

In the multiple microphone condition, time-delay-of-arrival features are computed between a reference channel (selected automatically) and each of the other available channels, at a rate of 10 ms, with an analysis window of 500 ms. These delays are input into the clustering system as an extra feature vector and are modeled by an HMM model using the same topology as the cepstral features, using one Gaussian per cluster. In both the Viterbi decoding and the BIC comparison, we used a weighted combination of the two models.

The algorithm is initialized using the prosodic-features initialization scheme presented in [7, 8]. Each cluster is modeled with a Gaussian mixture model (GMM). The algorithm then performs the following iterations:

Re-Segmentation: Run Viterbi alignment to find the optimal path of frames and models. As classifications based on 10 ms frames are very noisy, a minimum duration of 2.5 s is assumed for each speech segment.

Re-Training: Given the new segmentation of the audio track, compute new GMMs for each of the clusters.

Cluster Merging: Given the new GMMs, try to find the two clusters that most likely represent the same speaker. This is done by computing a score based on the Bayesian information criterion (BIC) of each of the clusters and the BIC score of a new GMM trained on the merged segments for two clusters. If the BIC score of the merged GMM is larger than or equal to the sum of the individual BIC scores, the two models are merged and the algorithm continues at the re-segmentation step using the merged GMM. If no pair such is found, the algorithm stops.

A more detailed description can be found in [9, 10]. As a result of various optimization approaches [11], our current implementation runs at about 0.6 times real time.

3.3. Other diarization systems

For analysis purposes, we also experimented with two additional diarization outputs. As an upper bound on what could be achieved with perfect diarization, we use the NIST scoring references as a "cheating" diarization system, providing close-to-perfect speech/nonspeech and speaker assignment. Second, we also tested our algorithms on output provided by the Institute for Infocomm Research/Nanyang Technological University (IIR/NTU), Singapore [12], which had outstanding performance on the RT-09 diarization tasks.

3.4. Speech Recognition

The STT 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 meeting recognition evaluation [13]. 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 [14]. Language models (LMs) consist of a mixture of genre-specific models for meeting transcripts, telephone conversations, broadcast news, and web data.

Table 2. DER and WER results for different diarization methods

Diarization	RT07 MDM			RT07 SDM			RT09 MDM			RT09 SDM		
metric	DER	WER	WER	DER	WER	WER	DER	WER	WER	DER	WER	WER
overlap		1	4		1	4		1	4		1	4
Baseline	30.9	26.2	40.5	53.9	33.1	45.2	37.3	34.0	42.9	37.3	41.3	49.9
ICSI	8.6	25.9	37.8	17.2	32.5	44.2	17.2	35.9	43.0	31.3	44.6	51.6
IIR/NTU	n/a	n/a	n/a	n/a	n/a	n/a	9.2	34.7	43.4	16.0	40.9	49.4
Reference	0	24.6	39.2	0	30.7	44.6	0	31.7	43.9	0	39.5	50.0

Table 3. WER (overlap-1) results for different diarization front ends

Speech detection	Clustering	RT07	
		MDM	SDM
Baseline	Baseline	26.2	33.1
ICSI	Baseline	26.4	33.1
ICSI	ICSI	25.9	32.5

The recognition system assumes that speech is first separated from nonspeech and segmented into segments of moderate length (up to 60 seconds). Further, the system assumes that the speech segments are clustered into equivalence classes that, ideally, each pertain to a single speaker. These *pseudo speaker* clusters are then used to perform vocal tract length normalization, and cepstral mean and variance normalization. In later recognition passes, the pseudo speaker clusters also form the units on which unsupervised acoustic adaptation is performed. In each MLLR step of the multi-pass recognition systems, separate adapted models are computed for each of the speaker clusters.

Recognition for single and multiple distant microphone conditions differs only in the preprocessing of the signal, using Wiener filtering and beamforming as for the diarization system. Processing time is about 3.8 times real time on an 8-core, 3.1-GHz Intel-based server.

4. COMBINING DIARIZATION AND RECOGNITION

From prior experience (our own as well as others' [2]), good STT performance requires postprocessing of the speech regions detected by the diarization system. The parameters of this postprocessing were optimized on a previous NIST evaluation set (RT-06). The resulting steps for segmentation postprocessing were as follows:

1. Merge segments assigned to the same speaker if they are separated by less than 0.4 s nonspeech, and as long as the resulting segment is shorter than 60 s long.
2. Discard segments with less than 0.2 s of speech
3. Pad segments with 0.2 s of nonspeech at the beginning and end.

After optimizing these steps based on the ICSI diarization output on RT-06 data, we kept them constant for all subsequent experiments with other diarization systems and reference outputs. The speaker assignments of the diarization system can be transferred to the STT system because the segment postprocessing never combines segments from different speakers.

Table 3 shows the STT results with various combinations of baseline and ICSI processing for speech segmentation and clustering. A small gain (0.3% for MDM, 0.6% for SDM) is seen with ICSI diarization compared to the baseline. However, we also see that this gain can be credited fully to the improved speaker clustering from the diarization system. The updated segmentation does not seem to yield an improvement, and it might even hurt STT (0.2% degradation for MDM). Note that the ICSI speaker clustering was optimized to operate on ICSI segmentation; therefore, we found it safest to adopt both segmentation and clustering from the diarization system.

Table 2 shows DER and WER results for different diarization methods. (The IIR/NTU diarization output was only available for the RT-09 test set.) The main observation is that the RT-09 data was much more challenging for diarization, and using its output for STT leads to a loss compared to the baseline. A likely reason for the greater difficulty is that the RT-09 evaluation set contains more speakers (cf. Table 1) and, most importantly, a much larger amount of overlapping speech (up to 37% in one meeting) than all previous evaluation and development sets. Even using the IIR/NTU system, with much lower DER than ICSI, gives no gain in the MDM condition, even though a 0.4% improvement is found for SDM. Also note that, as expected, the baseline system has very high DER, mostly because of poor speaker clustering (by diarization standards).

The reference diarization provides substantial benefits for overlap-1 WER: 2.3% lower WER for MDM and 1.8% for SDM. Note that the overlap-4 WER results with reference diarization are not easily interpreted since the reference output contains overlapping speaker segments, but the STT system has no special processing for such overlaps (it assumes the speakers are strictly nonoverlapping). None of the actual diarization systems produce overlapping speaker labelings.

5. COMBINING MULTIPLE DIARIZATIONS

To extract larger STT gains from diarization output, we may combine STT outputs from multiple STT systems, each based on different diarizations. This approach is motivated by two prior results: Cambridge U. reported gains from combining hypotheses obtained based on different broadcast news segmentations [15], and in the 2007 evaluation, we had seen gains by combining systems that differed only in their pseudo-speaker clustering parameters, according to the baseline algorithm [13]. In this study, we combined two STT systems at a time. One was always based on the baseline segmentation and clustering method (in hindsight this was a good choice because the baseline system seems quite robust even to data that is difficult to diarize). The other STT systems made use of one of the diarization outputs, as described before.

Hypotheses from the component systems were then combined using one of two methods. The first was the NIST ROVER algorithm [16], which aligns the hypothesized 1-best words and resolves disagreements based on word confidences. We used the word posterior probabilities generated during the final confusion network (CN) combination stage in each of the STT systems as confidence estimates.

The second method was confusion network combination (CNC), whereby the CNs from both STT systems are aligned, and all word hypotheses (not just the 1-best) vote with their posterior probabilities toward a new best hypothesis [17]. CNC requires a consistent seg-

Fig. 2. Consensus segmentation for CNC from differing diarizations

Table 4. STT WER results combining diarization methods

Diarization overlap	RT09 MDM 1	RT09 MDM 4	RT09 SDM 1	RT09 SDM 4
Baseline	34.0	42.9	41.3	49.9
ICSI	35.9	43.0	44.6	51.6
Base+ICSI (ROVER)	34.2	43.8	42.2	51.1
Base+ICSI (CNC)	33.3	43.0	40.8	50.1
IIR/NTU	34.7	43.4	40.9	49.4
Base+IIR/NTU (CNC)	32.7	41.5	40.0	48.8

mentation of hypotheses across the combined system, and is therefore not directly applicable to our case. Stringing all the segment-level CNs from one system together would not be feasible, as the CNC algorithm is quadratic in the length of the inputs. As a solution, we concatenate CNs according to a *consensus segmentation* (i.e., a set of segments bounded by nonspeech regions that are shared among all input systems, as depicted in Figure 2), and then perform CNC on the matching concatenated CNs. Finally, the new 1-best words are force-aligned to the consensus waveform segments to obtain word times.

As shown in Table 4, combining diarizations is generally beneficial. CNC gives results that are between 0.8% and 1.4% better than ROVER combination. Using the best-available diarization (from IIR/NTU), we now see substantial gains over the baseline, of between 1.1% and 1.4% lower WER. Even for the ICSI diarization system, CNC results in 0.7% to 0.5% overlap-1 WER reduction (and a very small loss in overlap-4 WER).

6. CONCLUSIONS

We have shown that meeting diarization system output can give modest gains in STT accuracy over a simple, but relatively robust baseline speech segmentation and clustering algorithm as found in our state-of-the-art meeting recognizer. Potential gains from better diarization are substantial (around 2%), as assessed by utilizing gold-standard diarization references. However, even state-of-the-art meeting diarization systems are presently not accurate enough to give substantial and consistent STT gains, as seen on the more difficult RT-09 evaluation set. Finally, we do see more than 1% absolute WER reduction over baseline when the outputs of STT systems based on different diarizations are combined. To this end, we developed a confusion network combination algorithm that can deal with diverging waveform segmentations in the component systems.

7. ACKNOWLEDGMENTS

We are grateful to Haizhou Li and Trung Hieu from IIR/NTU for making their diarization system output available for our study. We would also like to thank Jon Fiscus and Jerome Ajot at NIST for assistance with the ROVER and evaluation software, and Adam Janin at ICSI for invaluable help with data processing. Gerald Friedland and David Imseng were sponsored by the Swiss-funded IM2 program. Gerald Friedland has also been sponsored by the European project AMIDA.

8. REFERENCES

[1] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The Rich Transcription 2007 meeting recognition evaluation", In Stiefelhagen et al. [18], pp. 373–389.

[2] S. E. Tranter, K. Yu, D. A. Reynolds, G. Evermann, D. Y. Kim, and P. C. Woodland, "An investigation into the interactions between speaker diarisation systems and automatic speech transcription", Technical Report CUED/F-INFENG/TR-464, Cambridge University Engineering Department, Oct. 2003.

[3] 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.

[4] A. Miró, *Robust Speaker Diarization for Meetings*, PhD thesis, Universitat Politècnica de Catalunya, Oct. 2006.

[5] 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.

[6] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system", In Stiefelhagen et al. [18], pp. 509–519.

[7] D. Imseng and G. Friedland, "Tuning-Robust Initialization Methods for Speaker Diarization", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. to appear, 2010.

[8] D. Imseng, "Novel initialization methods for speaker diarization", Idiap-RR Idiap-RR-07-2009, IDIAP, May 2009, Master's thesis.

[9] J. Ajmera, "A robust speaker clustering algorithm", *in Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 411–416, St. Thomas, U. S. Virgin Islands, Dec. 2003.

[10] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló, "Robust speaker segmentation for meetings: The ICSI-SRI Spring 2005 diarization system", in S. Renals and S. Bengio, editors, *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, vol. 3869 of *Lecture Notes in Computer Science*, pp. 402–414. Springer, 2006.

[11] Y. Huang, O. Vinyals, G. Friedland, C. Müller, N. Mirghafori, and C. Wooters, "A fast-match approach for robust, faster than real-time speaker diarization", *in Proceedings IEEE Workshop on Speech Recognition and Understanding*, pp. 693–698, Kyoto, Dec. 2007.

[12] H. Sun, T. L. Nwe, B. Ma, and H. Li, "Speaker diarization for meeting room audio", *in Proc. Interspeech*, pp. 900–903, Brighton, U.K., Sep. 2009.

[13] 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 Stiefelhagen et al. [18], pp. 450–463.

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

[15] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK broadcast news transcription system", *IEEE Trans. Audio, Speech, and Lang. Process.*, vol. 4, pp. 1513–1525, Sep. 2006.

[16] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER)", *in Proceedings IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 347–352, Santa Barbara, CA, 1997.

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

[18] 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*, Berlin, 2008. Springer.