Transcribing Meetings With the AMIDA Systems

Thomas Hain, *Member, IEEE*, Lukáš Burget, *Member, IEEE*, John Dines, *Member, IEEE*, Philip N. Garner, *Senior Member, IEEE*, František Grézl, *Member, IEEE*, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiát, Mike Lincoln, and Vincent Wan

***Abstract*—In this paper, we give an overview of the AMIDA systems for transcription of conference and lecture room meetings. The systems were developed for participation in the Rich Transcription evaluations conducted by the National Institute for Standards and Technology in the years 2007 and 2009 and can process close talking and far field microphone recordings. The paper first discusses fundamental properties of meeting data with special focus on the AMI/AMIDA corpora. This is followed by a description and analysis of improved processing and modeling, with focus on techniques specifically addressing meeting transcription issues such as multi-room recordings or domain variability. In 2007 and 2009, two different strategies of systems building were followed. While in 2007 we used our traditional style system design based on cross adaptation, the 2009 systems were constructed semi-automatically, supported by improved decoders and a new method for system representation. Overall these changes gave a 6%–13% relative reduction in word error rate compared to our 2007 results while at the same time requiring less training material and reducing the real-time factor by five times. The meeting transcription systems are available at www.webasr.org.**

***Index Terms*—AMI corpus, Juicer, meeting transcription, multiple distant microphone, resource optimisation, rich text.**

I. Introduction

OVER the past decades, automatic speech recognition (ASR) research has progressed from work on carefully crafted tasks to real applications that do not require cooperation by the user. Significant progress was made on tasks such as transcription of broadcast news (BN) and conversational telephone speech (CTS), although they remain challenging despite substantial investment from many sides. Aside from advances in core ASR technology these tasks also established the need for additional information apart from the raw sequence of words. Segment information such as timing boundaries or speaker identities are helpful for downstream processing and help readability. This gave rise to new tasks such as diarization ("who spoke when"), automatic capitalization or disfluency removal, yielding *rich transcripts*. The U.S. National Institute for Standards and Technology (NIST) started to benchmark rich transcription (RT) systems in 2002 [1], with further competitions held in 2004, 2005, 2006, 2007, and 2009. In this paper, we present the AMIDA systems developed for these benchmark tests.

Moving on from two-person conversations (as in the CTS task) the past decade saw interest in the processing of meeting speech under a large variety of conditions and scenarios (e.g., [2]). The driving force is not the interest in more generic speech tasks but in the analysis and streamlining of meetings themselves. Many people spend considerable time in meetings and complain about low efficiency, and loss of information—only very formal meetings normally are minuted. So far, computers are still rarely used in streamlining the process or for extracting and retaining the essential information. Efficiency is even poorer when the meeting participants are not in the same room, i.e., using teleconference facilities. In the days of decreasing travel budgets and concerns for the environment, video- and teleconferencing is more widely used and thus the opportunity to record, process, recognize, and categorize the interactions in meetings is recognized even by sceptics of speech and language processing technology. This area was also the focus of the AMI and AMIDA projects [3]: acquisition, multi-modal recognition, and higher level processing of data from distributed or single room meetings.

Meetings are an audio-visual experience by nature, but verbal communication forms the backbone of most meetings, and automatic transcription of speech in meetings is of crucial importance for meeting analysis, content analysis, summarization, and analysis of dialog structure. Enabled by the availability of the ICSI meeting corpus (see Section II) NIST included meeting transcription into RT evaluations from the beginning, even though only as a pilot study for the first years. While ASR is often solely associated with transcription, applications often do not require full transcripts (e.g., content linking [4]). Application specific optimization can yield better results for tasks, but the diverse nature of such application in this domain requires falling back to standard metric-based assessment of transcription and diarization. In the RT'06 evaluation the concept of meetings was cast wider by distinguishing

Manuscript received September 08, 2010; revised June 12, 2011; accepted July 16, 2011. Date of current version January 13, 2012. This work was supported in part by the European Union Sixth FWP IST Integrated Project AMIDA (Augmented Multi-party Interaction with Distance Access) FP6-033812. The works of L. Burget, F. Grézl, and M. Karafiát were supported in part by Czech Ministry of Interior Project VD20072010B16, Grant Agency of Czech Republic Project 102/08/0707, Czech Ministry of Education Project MSM0021630528, and by BUT FIT Grant FIT-10-S-2. The work of F. Grézl was supported in part by Grant Agency of Czech Republic under Project GP102/09/P635. The works of J. Dines and P. N. Garner were supported in part by the Swiss National Science Foundation under the National Center of Competence in Research (NCCR) on Interactive Multi-modal Information Management (IM2). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sadaoki Furui.

T. Hain, A. E. Hannani, and V. Wan are with the SpAndH, University of Sheffield, Sheffield S1 4DP, U.K. (e-mail: t.hain@dcs.shef.ac.uk).

L. Burget, F. Grézl, and M. Karafiát are with the FIT, Brno University of Technology, Brno, 612 66, Czech Republic.

J. Dines and P. N. Garner are with the Idiap Research Institute, CH-1920 Martigny, Switzerland.

M. Lincoln is with the CSTR, University of Edinburgh, Edinburgh EH8 9LW, U.K.

M. Huijbregts is with the HMI, University of Twente, 7500 AE Enschede, The Netherlands.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2163395

between conference and seminar style meetings. The latter separated presentational meetings, where discussions with an audience can occur but are not the norm.

The transcription systems presented here are the joint effort of the AMI/AMIDA group, in a close international collaboration that participated in the RT evaluations every year since 2005 [5] with very competitive results in 2007 and 2009, on far and near-field tasks [6]. The rest of the paper is organized as follows: We first discuss data resources available for meeting processing and their properties, with special focus on the AMI corpus. This is followed by descriptions of processing of acoustic signals, language and acoustic modeling and decoding. We outline the system design strategy and give results on RT test sets in Section VII.

II. Meeting Data

Many meeting resources are available in several languages. The resources available in English are diverse—they rarely include speakers with a single accent. The speakers in several corpora are people involved with research, and include a large proportion of non-native English speakers. Hence, the data used for training of AMIDA systems includes a mixture of U.S. and U.K. speakers and a very high proportion of foreign speakers from French, German, and many other origins. While this naturally makes the task difficult, on the positive side robustness is obtained through multi-accent training.

Another property of the meetings domain is the diversity in recording. While some corpora provide speech recordings with a large number of microphones in different configurations, others simply use one microphone in the middle of the meeting table for acquisition. Even for close-talking recordings the diversity in microphone quality causes substantial differences in signal quality. For example, only some resources use noise cancelling microphones or highly directional microphones [7]–[10]. The impact on signal to noise ratios and crosstalk is substantial. In the case of far field recordings, some corpora provide microphone arrays in well defined configuration while others intentionally place microphones according to convenience. Conference and seminar style meetings differ in the distance of the speaker from the microphone and room acoustics. In all cases, a major concern is synchronous recording of sources. Especially for the far field, sample synchronicity would be desirable but is only provided in a few cases.

The first corpus available was the ICSI Meeting corpus [10], consisting of 70 technical meetings with a total of 73 hours of speech. The number of participants is variable and data is recorded with head-mounted and a total of four table-top microphones. Further meeting corpora were collected by NIST [9] and ISL [7], with 13 and 10 hours, respectively. Both NIST and ISL meetings have unconstrained content (e.g., people playing games or discussing sales issues) and variable numbers of participants. The main additions in 2007 were the completion of the AMI corpus [8] and the second phase release of the NIST corpus. The AMI corpus consists of 100 hours of meetings where 70 hours follow so-called "scenarios" where certain roles are acted by the meeting participants. The meetings are recorded at three different sites and, due to the proximity to research, include a large percentage of non-native English speakers. A further small addition (7 hours of transcribed meetings) was added in 2009 with the AMIDA corpus which consists of meetings held at two locations.

TABLE I
Segment Statistics for Meeting Corpora

Meeting resource	Avg Dur (sec.)	Avg. Words/Seg
ICSI	2.11	7.30
NIST	2.26	7.17
ISL	2.36	8.77
AMI	3.29	10.09
VT	2.49	8.27
CHIL	1.80	5.63

TABLE II
Perplexities of Interpolated LMs on the AMI Corpus. $\oplus$ Denotes "Interpolation With"

AMI $\oplus$	Overall	male	female	Scenario	Other
BN	99.8	99.3	100.9	87.9	137.8
CTS	100.5	100.1	101.6	88.2	140.2
Mtg	102.7	101.6	105.4	91.2	138.8
All	92.9	92.8	93.2	84.1	119.7

A. AMI Corpus

Data was collected at three different instrumented meeting rooms in Europe (Edinburgh, Idiap, TNO). The meeting language is English, but many participants are non-native speakers of the language. Each meeting normally has four participants and the corpus is split into a scenario portion and an unconstrained meetings portion. Each scenario in the corpus consists of four meetings with the same participants working on a constrained task. For a full description the reader is referred to [8] or the documentation available with the corpus. For the purpose of ASR the data has some unique properties. Table I shows the raw average segment statistics for several corpora.[1] Segment lengths vary greatly, with AMI corpus recordings having much longer sentences on average, hinting at more controlled speech. The short segments on the CHIL data are surprising given that these are from seminar meetings. The speaking rate however is very similar for all corpora, varying between 3.1 and 3.6 words per second. The higher control is visible when looking at language model (LM) performance. In Table II, perplexities of 3-gram models obtained by linear interpolation of a model derived from the AMI corpus with background material from BN (150 MW), CTS (3 MW), and other meeting sources (1 MW), and all combined. The models are constructed in five-fold cross-validation. It is clear that scenario meetings, making up a larger proportion of the corpus, are much lower in complexity. Note that CTS and meeting data include automatically collected web data [11]. Further analysis of perplexities (Table III) also reveals that the language of origin also has considerable effect. An investigation of out of vocabulary (OOV) words (not shown here) cannot explain the differences, with the speakers of French origin having

[1]A segment here is defined as speech not interrupted with silence of at least 100-ms length.

TABLE III
PERPLEXITIES OF LMS ON THE AMI CORPUS WITH DISTINCTIONS ON LANGUAGE OF ORIGIN

AMI ⊕	English	French	German	EU	Asia	Other
BN	105.2	97.7	128.5	113.3	112.0	102.8
CTS	105.9	100.2	128.9	114.4	115.0	104.0
Mtg	110.3	98.0	126.8	115.9	113.3	103.7
All	96.9	90.8	111.0	103.0	104.7	94.9

TABLE IV
%WER ON TWO 6-HOUR TEST SETS FROM THE AMI CORPUS (WITH AND WITHOUT OVERLAPPED SPEECH)

mic source	#mic	no overlap	with overlap
close talking	1	26.8	33.0
distant	1	60.2	67.2
beamforming	2	54.6	62.8
beamforming	4	52.5	61.2
beamforming	8	50.8	59.4

lowest perplexity (even lower than native English speakers) and German speakers the highest.

Another distinguishing factor is the recording configuration. Circular microphone arrays are used in well defined placement in the meeting room. Table IV shows experimental results on the benefit of multiple microphones. Experiments were conducted using different acoustic data from the same meetings, including beamforming (as described in [5]) with varying numbers of microphones. All acoustic models were trained using standard maximum-likelihood estimation on the AMI corpus only and results were obtained on two different 6-hour test sets, one only with segments with overlapped speech and one without. One can observe considerable degradation between close talking and single distant microphone performance; however, some can be regained from beamforming. Overlapped speech is nonetheless a significant problem which reduces such improvements.

B. Test Sets and Test Conditions

The NIST evaluation data has over the years not only included meetings from sources outlined above, but also from meetings recorded by Virginia Tech University (VT) and the Linguistic Data Consortium (LDC). Equally test sets for seminar style meetings from the CHIL corpora have been provided. In the past we have shown that the performance of the AMIDA systems on seminar data is strongly correlated to conference room meeting transcription [6], [12]. Hence, we have excluded such data from this paper. The AMIDA systems were also tested on the speaker attributed speech to text transcription (STT) task using the diarization output from the ICSI diarization system [13]. Again, it was found that a modest amount of errors are added to the STT results and hence further analysis was excluded. The rest of the paper uses the RT 2005, 2007, and 2009 evaluation test sets, *rt05seval*, *rt07seval*, and *rt09seval*, respectively. Distinctions between the independent head microphone (IHM) and multiple distant microphone (MDM) tasks are explicitly made in the text. In 2009, one new aspect, meetings recorded on two sites, connected with video conferencing, were included.

TABLE V
LM INTERPOLATION WEIGHTS FOR BUILDING OF 4-GRAM LMS

LM corpus source	weight
Fisher web data (Univ. Washington)	0.220
AMI corpus	0.210
Fisher	0.186
Meetings web data (Univ. Washington)	0.103
ISL meeting corpus	0.081
Switchboard Callhome	0.048
Swbd web data (Univ. Washington)	0.045
AMI corpus web data	0.038
Hub4 1996 LM	0.035
NIST meetings phase 2	0.029

III. VOCABULARY AND LANGUAGE MODELING

Meetings cover a wide range of topics, however, both vocabulary and language modeling appears to be mostly driven by the fact that these are conversations that normally make use of small vocabularies. In [14] and [15], padding of meeting vocabularies with the most frequent words from BN sources up to 50 000 words was found to be sufficient to yield out-of-vocabulary (OOV) rates below 1%. This property was retained on *rt07seval* and *rt09seval*; thus, any vocabulary updates where mostly related to LM training corpus changes. Language models (LMs) were constructed in a two-stage process. In the first instance, component models from diverse sources are constructed and optimal interpolation weights are found. Table V shows the 10 out of 15 components with the highest weights. The large set of components is driven by the experience that even BN sources can be very helpful, as outlined in Table II. The set of components include sources from web data search [16]. In a second stage, the interpolated LM serves as basis for a renewed web-data search using the methods outlined in [11]. The approach allows limiting of the search, and hence only 20 MW of web data are collected and used to train an additional LM component for further interpolation. Re-interpolation with only the components that had a weight of more than 1% yield the final model. The final 2007 4 gram LM had a perplexity of 73.1 on the on RT06 evaluation data. Table VI shows perplexity results for two LMs, one optimized for seminar data, one for conference room meetings. In both cases the RT'06 data sets served as estimation sets for interpolation. Modest improvements can be observed from the two-stage approach. However, tuning of LMs to the two domains clearly brings significant improvements.

This LM construction paradigm was used for all AMIDA language models. As OOV rates are generally found to be low using word list padding, only words from the 2007 evaluation data were added to the 2009 word lists. The 2009 LMs were built for a different decoder and LMs using lower n-gram cutoff thresholds were constructed. As the LM sources themselves did not change, no new web-data was added at this point. Table VII shows the impact on OOV, perplexity, and WER performance.

TABLE VI
LM CONSTRUCTION FOR CONFERENCE AND SEMINAR MEETINGS. PERPLEXITIES ON *rt*07seval

Target domain	LM Build Stage	Test	
		Conference	Seminar
Conference	1	73.2	144.5
	2	73.1	140.8
Seminar	1	82.9	120.4
	2	81.9	119.3

TABLE VII
%WER FOR DIFFERENT LMS ON THE *rt*07seval AND *rt*09seval IHM SETS

LM	Perplexity		%OOV		%WER
	rt07seval	*rt09seval*	*rt07seval*	*rt09seval*	*rt07seval*
2007	87.8	73.1	0.74	0.30	37.9
2009	86.4	71.0	0.62	0.29	36.2

The number of n-grams was increased considerably and had to be reduced through entropy pruning [17].

IV. ACOUSTIC PREPROCESSING

Two fundamental modes of operation exist for meetings. Either a microphone is associated with a single speaker (IHM) or one or many microphones are used to record the voices of all speakers. The IHM case is most similar to CTS or BN, where microphones are placed close to the mouth. However, it also includes lapel microphones which often capture a significant amount of speech energy from surrounding speakers. For MDM, different acoustic modeling may be used, for example using channel adaptive training [18]. However, the standard for most systems is to transform the acoustic signal from multiple microphones into a single audio stream and then process the signal in a similar fashion to IHM signals. The ASR systems usually benefit from segment information, i.e., the time boundaries of speech segments. Furthermore, systems are speaker adaptive; thus, speaker cluster information is required. While the speaker identity is clear for IHM data, automatic segment clustering is required for far field input.

A. Close Talking

Segmentation of meetings is a difficult task due to the considerable amount of cross-talk. The AMIDA systems use a multi-layer perceptron (MLP)-based speech/silence classifier trained on MF-PLPs and features representing correlation and energy of competing channels [19]. Even though classification is binary, a large number of speech and silence training examples are required. Table VIII shows the influence of increasing amounts of training data on recognition performance and compares with reference segments. Larger training sets gave different segment statistics and better results, although not uniformly for all meeting rooms. In the case of the most difficult situation (CMU lapel microphones), the results even degrade due to poorer proportional representation in the new training set.

TABLE VIII
%WER ON *rt*07seval FOR REFERENCE SEGMENTS, AND HOURS OF MLP TRAINING DATA. CMU/EDI/NIST/VT ARE MEETING ROOMS

	#Seg	Tot	CMU	EDI	NIST	VT
Ref	4527	29.3	36.7	24.5	24.5	31.2
30h	2717	32.6	41.2	26.2	29.1	33.3
90h	4541	31.7	42.4	25.3	26.8	31.7

The segmenter itself is implemented as a hidden Markov model (HMM) to allow the inclusion of duration constraints, speech/silence class priors, and an insertion penalty. The output probabilities are given by scaled versions of the MLP posteriors. Originally, mostly the insertion penalty was changed to alter the number of segments. Fig. 1 displays histograms of segment lengths on the *rt07seval* data. The first histogram 1(a) shows reference segmentation. Forced alignment with ASR models yields the second histogram 1(b). Histogram 1(c) was obtained when the main focus was on optimization of segment numbers with the insertion penalty. This approach produced a strong bias towards shorter segments. Appropriate tuning of all hyper parameters yielded a statistic more similar to the target distribution [Fig. 1(d)]. This method in turn gave very good ASR results on *rt07seval*. Table IX shows results on the *rt09seval* set. As can be observed the difference between reference and automatic segmentation is substantial, with 3.5% WER absolute. In addition, the effect of the silence prior on performance does reveal significant influence with the best performance 36.1%. The table also includes oracle results as if the optimal silence prior had been set per meeting or per channel. It is clear that at least the meeting specific settings have significant influence. Further analysis reveals that one meeting has significantly lower energy and signal to noise ratio compared to the standard. The degradation compared to the reference segmentation here is larger than 10% WER absolute. A normalization of energy is likely to improve the situation.

B. Far Field

The number of microphones available for corpora outlined in Section II ranges from 1 to 16. Hence, the approach used needs to allow for variation, even though the NIST paradigm allows systems to make explicit use of meeting room characteristics, the AMIDA systems have not done so. Instead the audio signal is enhanced by beam-forming based on time-delay-of-arrival (TDOA) that can be used with any number of microphones [5]. After noise-filtering of the audio channels the TDOA is estimated. The estimates here, however, are noisy and hence post-filtering is applied. The system presented in [20] has put considerable effort on Viterbi post-filtering to yield smooth trajectories. A comparison of that system with [5] showed 2.2% WER absolute reduction on *rt07seval* and we hence used the toolkit for the RT'09 system.

The second task is segmentation and speaker clustering. In work on the RT'07 system we found a considerable mismatch between requirements for diarization and ASR [6]. ASR systems require clusters of reasonable size and can cope with segments that contain significant amounts of silence. In previous

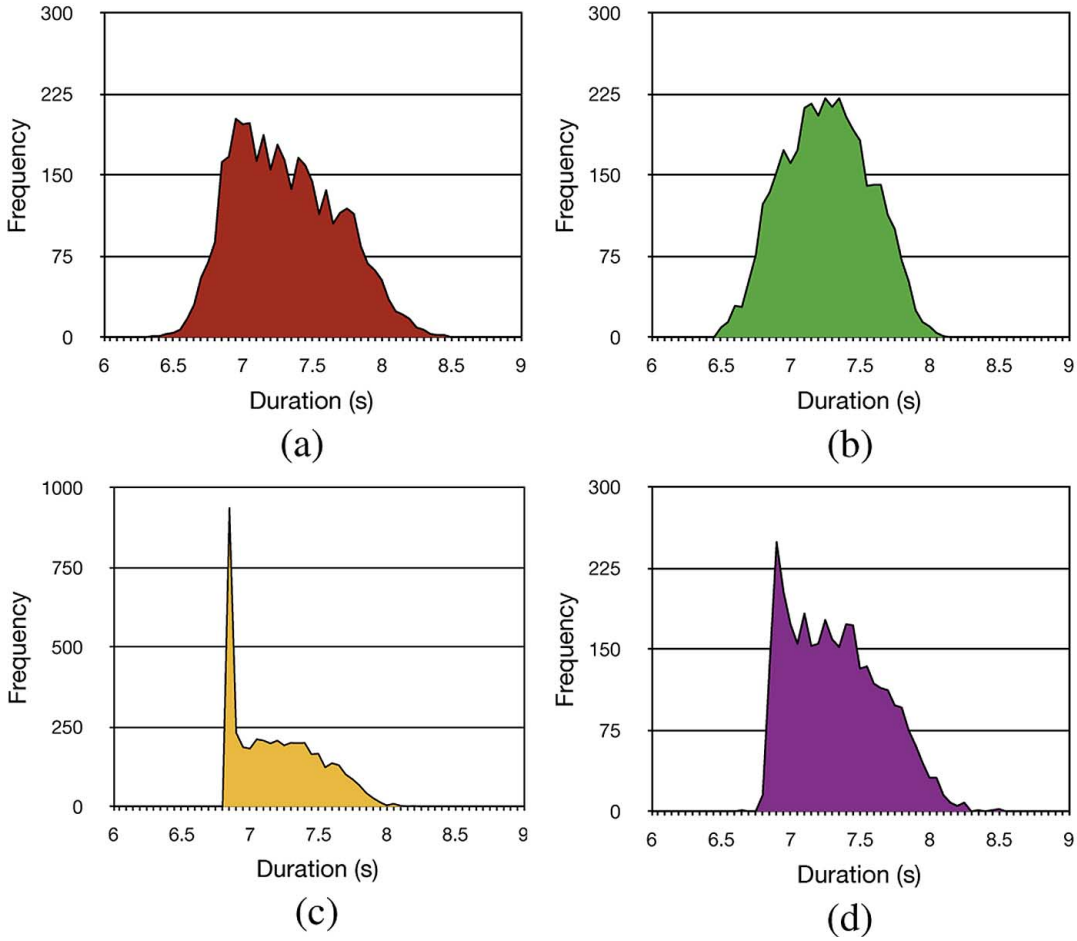

Fig. 1. Segment duration histograms on *rt07seval* for different segment sources. (a) Manual. (b) Forced alignment. (c) #seg optimized. (d) Histogram optimized.

TABLE IX
%WER ON *rt*09seval. CHOOSING THE PROBABILITY OF SILENCE P_{sil}. O STANDS FOR ORACLE RESULT

System	P_{sil}	#Segs	Tot	Sub	Del	Ins
ref	–	5660	32.9	22.1	7.1	3.7
auto	0.80	4809	36.4	20.6	11.9	3.9
auto	0.85	4949	36.1	20.9	10.7	4.4
auto	0.90	5135	36.4	21.0	10.7	4.5
auto	0.95	5504	39.8	22.0	8.5	9.4
auto	O Meeting	-	35.1	21.3	9.0	4.9
auto	O Speaker	-	34.9	21.2	8.9	4.8

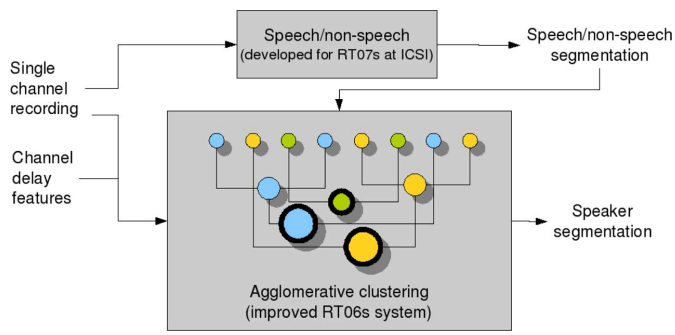

Fig. 2. Clustering of beam-formed audio.

years, we made use of such segmentation and clustering provided by ICSI/SRI (e.g., [15]). For the 2009 system a diarization framework based on [21], specifically adapted for ASR, was included. The acoustic pre-processing is identical to before, i.e., microphone array beamforming with the BeamFormIt toolkit [20]. However, the energy based beamformer now delivers a single audio stream together with the relative time delay estimates between the channels. In contrast to [21], where only the beamformed audio was used for clustering, the delay values now augment standard MFCC features. Fig. 2 illustrates the process. Segment clustering is using the Bayesian information criterion (BIC), with the initial number of clusters based on the amount of data. The Mel-frequency cepstral coefficient (MFCC) and delay feature streams are normalized to yield identical average BIC scores.

Table X shows WER results of automatic approaches, in comparison with the reference, for segmentation and speaker clustering. The loss from automatic segmentation alone is 1.7%, not surprisingly the difference after adaptation is similar. The difference between the unadapted results with or without speaker information originates from cepstral mean and variance normalization (CMN/CVN) as that is also speaker based. Using delay features for clustering brings substantial performance gain, and

TABLE X
%WER ON *rt*07seval USING THE FIRST (UNADAPTED) AND THIRD (ADAPTED) PASS OF THE RT'07 AMIDA MDM SYSTEM

Segmentation	Clustering	Unadapted	Adapted
Ref	-	42.1	36.3
auto	-	43.8	38.1
auto	Ref	40.1	31.1
auto	no delay	42.8	34.5
auto	with delay	42.1	32.7

TABLE XI
%WER RESULTS ON *rt*09seval MDM USING AN ADAPTED SYSTEM AND SEGMENTS FROM ROOM 1, ROOM 2, OR BOTH ROOMS

Description	Segmentation	Tot	Sub	Del	Ins
echo filtering	Auto	33.2	20.6	9.3	3.2
only room1		36.3	20.5	12.7	3.1
only room2		45.1	25.8	14.8	4.4
echo filtering	Ref	30.8	20.1	8.6	2.2
only room1		33.1	21.0	9.9	2.1
only room2		41.0	24.3	14.6	2.1

the final loss from automatic speaker clustering is 1.6%. Experiments indicate that the losses for automatic segmentation and clustering are almost additive.

1) Echo Filtering: One challenge specific to *rt09seval* are meetings held in two rooms. This implies that audio was picked up by microphones in the first room (room 1) and was played through a loud-speaker in the second room (room 2). Naturally the microphones in room 2 pick up speech from that loud-speaker. One could assume that the audio is transported by a conferencing system that inevitably includes echo cancellation. However such systems are not always accessible and echo cancellation itself is far from perfect. Thus automatic filtering of the "echo segments" is required. The conferencing system introduces an unknown and variable audio transfer delay between rooms aside from clock differences, and the assignment of speakers to rooms is unknown.

When performing recognition it is desirable to adapt the ASR system to speakers, and to the recordings that are not distorted by a loudspeaker. This assumption is also verified by the results in Table XI. For both automatic and reference segmentation the best performance is achieved when filtering out echo segments. The filtering itself is performed as follows (a frame is a segment of speech, typically of length 500 ms).

1) For each room generate beam-formed audio.
2) Perform speaker segmentation on room 1 audio.
   a) For each speaker and frame, calculate the delay between the audio from room 1 and room 2 based on maximum cross correlation.
   If delay >0, increment room 1 count, otherwise the room 2 count.
   b) Assign speaker to room with highest count.
   c) Discard segments from speakers in room 2.

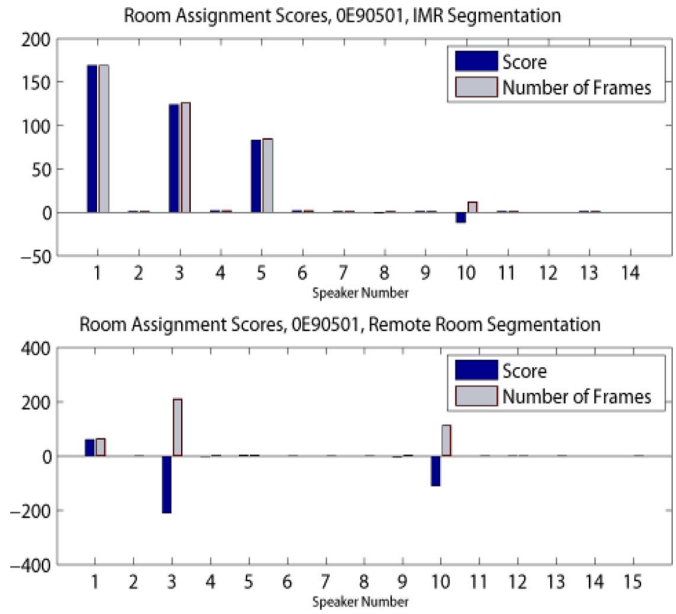

Fig. 3. Example of the echo filtering algorithm.

3) Repeat using segmentation from room 2 audio, discarding segments assigned to room 1.

Fig. 3 shows an example of frame counts. Speaker clustering output yielded too many clusters, but the important ones are clearly visible. With the above algorithm the single speaker in the remote room is clearly identified from each side. Results on the *rt09seval* multi-room meetings reveal that 3.1% WER absolute can be gained from using automatic room filtering compared to using only audio from one (i.e., the best) room (Table XI), but naturally this number depends on the amount spoken in each room. Interestingly, when using reference segmentation, the loss is only 2.4%, which seems to indicate that differences in segmentation are indeed a problem.

V. ACOUSTIC MODELING

Acoustic modeling for meetings differs only in some aspects from other areas such as CTS or BN. The first aspect, addressed here, is data selection, in particular for MDM data. A second issue is the availability of in-domain data. A total of 177 hours of IHM speech was available for training of the RT'09 system, and slightly less before. Hence, efforts to utilize the much larger resources, such as CTS data, can be beneficial. Aside from these, two methods have given gains consistently over many test and training sets: the use of posterior features and discriminative training.

A. Data Selection

Speech overlap varies from corpus to corpus and from meeting to meeting between 10% and 30% of the total meeting duration. Beamforming methods used here cannot separate two concurrent speakers. As back-channel messages are very common, significant distortion can be observed when including overlap speech, even with small amounts of overlap (see Table IV). Our experiments indicated [12] that simply using IHM segments for training of MDM models was sub-optimal.

TABLE XII
%WER ON *rt*07seval USING DIFFERENT THRESHOLDS ON CONFIDENCE SCORES IN LATTICES

%Data retained	80%	90%	95%	100%
ML	42.6	42.2	42.8	42.8
MPE	40.7	40.5	40.7	40.8

Instead we aimed to avoid segments with overlap: Removal of segments that contain any form of overlap would ignore more than 50% of the data, and is thus unacceptable. The segment boundaries given by manual annotation are often crude, and include significant amounts of silence. Hence, automatic methods for finding and removing true overlap need to be used.

The strategy employed requires forced alignment of all IHM segments followed by detection of the nearest word boundary to an overlap region. These are chosen as cutting points. Using this method about 154 hours of training data were retained. However, alignment is often unreliable in boundary regions even for IHM channels, and boundaries have no silence attached. An additional confidence based selection was used to remove an additional 10% of the data. Word lattices were generated for the complete training set and ranked according to the highest word level posterior probability in the lattices. A threshold on the posterior was chosen to remove a given percentage of the data. Table XII shows results for maximum-likelihood(ML) and minimum phone error (MPE) [22] training. While reasonable gain is observed for ML, the impact on discriminative training (1 iteration) is modest.

B. Adaptation From CTS

The Fisher corpus recordings are an extensive resource for conversational speech, and appear to be well suited to the task of meeting adaptation [15], [23]. The corpus data was prepared in the usual fashion, including the deletion of non-uniform amounts of silence at segment boundaries. A total of 170 hours of silence based on the manual segmentation was deleted, leaving a total of 2000 hours of speech. Naturally, CTS data has 4-kHz bandwidth (NB) whereas meetings are recorded typically with 8 kHz (WB). Using wider bandwidth gives significantly better results. In [12] and [24] we have presented a method to retain the benefit from wideband data modeling while retaining the gain from CTS data adaptation using maximum *a posteriori* (MAP) adaptation [25] with inclusion of heteroscedastic linear discriminant analysis (HLDA) [26]. Unfortunately, a detailed description cannot be given here and the interested reader is referred to those papers.

Table XIII shows the performance on the *rt05seval* set using models trained in the mapped narrow-band space on PLP features using VTLN, HLDA and MPE MAP [3]. Note that the baseline performance for CTS models does not change for the 2000 hour models. Solely training on meeting data yields significantly better results. However, after MPE-MAP adaptation, an overall improvement of 1.3% is observed.

C. Posterior Features

The use of posterior features consistently gave good improvements on meeting data. In original implementations the state

TABLE XIII
%WER RESULTS ON *rt*05seval ADAPTING CTS MODELS TO MEETING DATA INCLUDING NB/WB TRANSFORMS AND JOINT HLDA ESTIMATION. CTS/MTG DENOTE THE AMOUNT OF HOURS OF TRAINING DATA

CTS/Mtg	MAP	TOT	AMI	CMU	ICSI	NIST	VT
270/-	-	30.4	31.4	33.0	26.4	32.5	28.3
2000/-	-	30.4	30.7	31.3	27.9	32.2	30.1
-/170	-	23.4	20.5	21.2	20.2	29.0	26.6
2000/170	Std	23.8	22.8	23.0	20.8	27.1	25.5
2000/170	MPE	22.1	20.4	20.2	19.7	25.7	24.8

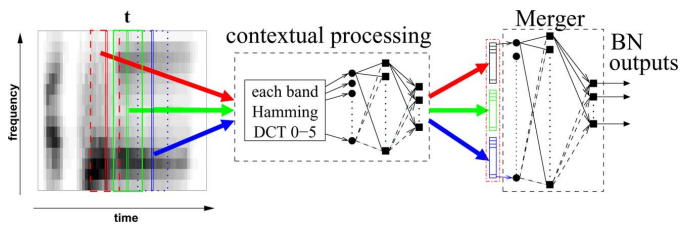

Fig. 4. Stacked bottleneck feature computation.

level posterior probabilities were estimates using two MLPs, one associated with times before the current time (left context, LC) and the other with the right context (RC) [12], [27]. This implementation required the use of HLDA for decorrelation and reduction of feature sizes. Later developments found that bottleneck (BN) MLPs gave similar or better performance while avoiding the need for substantial dimensionality reductions. It is straight-forward to extend BN features with the LCRC paradigm. In this case the output of LC and RC BN MLPs forms the input to a "merger" MLP, again with BN output, resulting in LCRCBN features [27].

The training of three MLPs is time consuming and requires considerable tuning. While conceptually the inputs to the final merger MLP are estimates of posterior probabilities of the current speech frame x_t with respect to left, right, and central context, the BN framework does not make explicit use of these. The basic concept in stacked BN (SBN) features is to replace the conditional MLPs with MLPs that always focus on a central frame. Given such an MLP it can be used to provide features derived from the left, the right, or in the case of SBN features, the middle of the current frame by simply taking its output at relevant time distance to the current frame. Fig. 4 illustrates the process, starting from a filterbank output illustrated in time and frequency. A Hamming window and discrete cosine transform (DCT) are applied to each filterbank band, retaining the coefficients 0–5. Three areas in the input spectra form the input to three identical (i.e., with same parameters) MLPs yielding three vectors that then form the input to the merger MLP. The size of the input window for the first stage MLP can be smaller than for LCRCBN features as an additional central vector is added.

Surprisingly, even the modification of the condition in MLP training has a positive effect. Table XIV shows a comparison of the feature types. Results are obtained using vocal tract length normalization (VTLN); CMN/CVN and BN features augment the PLP standard feature vectors. The resulting feature vector dimensionality ranges from 69 to 80 with configurations yielding

TABLE XIV
%WER ON *rt*07seval USING REFERENCE SEGMENTATION

HLDA-PLP	+BN	+LCRCBN	+SBN
36.0	31.7	30.6	29.4

TABLE XV
%WER ON *rt*07seval. USE OF DISCRIMINATIVE TRAINING TECHNIQUES WITH DIFFERENT FEATURES

HLDA-PLP+	ML	MPE	fMPE	fMPE+MPE
-	35.6	32.6	31.4	29.7
+LCRCBN	30.4	28.1	26.7	26.3
+SBN	29.4	27.5	26.9	26.1

the best result being displayed. SBN features clearly outperform all other variants. The number of trainable parameters in the system is reduced, allowing for training of larger MLPs to reach the same number of trainable parameters in the whole architecture.

D. Discriminative Training

Aside from VTLN and the use of posterior features, equivalent performance gains can be obtained from discriminative training such as MPE [22]. In the context of adaptation from CTS (Section V-B) discriminative training is difficult and MPE-MAP only partially allows transfer of gains from CTS to meeting data. Even more powerful training techniques suffer from the same problem. Hence, for the RT'09 systems all models were trained on meeting data only. In that way, the considerable complexity due to the use of the Fisher corpus for training was avoided and the simpler training setup allowed use of fMPE training [28].

fMPE is implemented using the region-dependent linear transform (RDLT) framework [29]. Posterior probabilities of the Gaussians are computed for each frame and these are spliced with the averages of posteriors for adjacent frames 1–2, 3–5, and 6–9 on the right and likewise for the left context. This means that 7 groups spanning 19 frames in total were used. All Gaussians from an ML trained HMM model are pooled and clustered using agglomerative clustering to create a Gaussian mixture model with 1000 components. Only offset features (not the posteriors) are used. Table XV shows results for use of fMPE in conjunction with BN features as outlined in Section V-C. fMPE is always applied to the full feature vector. One can observe that the gains are not additive. The 6.2% WER absolute gain from using SBN features and ML training is reduced to 3.6% after the joint use of fMPE and MPE. At this level, the improvement from MPE is 1.9% WER absolute and an additional 1.4% absolute (5% relative) is obtained with fMPE. The advantage of SBN over LCRC BN features in the end is small despite lower feature vector dimensionality.

VI. DECODING

Juicer, a weighted finite-state transducer (WFST)-based decoder [30] was used in previous systems, including 2007. However significant changes where made to its core allowing much faster and more accurate operation. The main changes

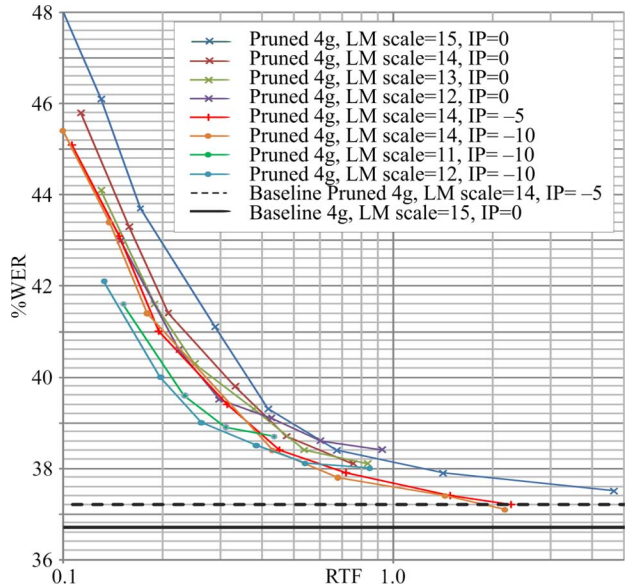

Fig. 5. Real-time factor (RTF) against %WER on *rt07seval* IHM for various decoding arrangements with HTK HDecode and Juicer. "Baseline" refers to lattice generation with a bigram LM and rescoring with a 4-gram LM.

are the inclusion of a modular front-end called Tracter [31], improved token passing and histogram pruning, use of acoustic model independent WFSTs and integration of the Hidden Markov Model Toolkit (HTK) for acoustic model computations. These changes had a considerable impact on processing speed against performance. Due to size constraints, and in contrast to lexical tree based decoders such as HTK HDecode, static WFST-based decoding requires language models to be pruned before WFSTs can be constructed, thus incurring performance degradation. However, on the positive side, higher order n-grams can be included with limited impact on real time factors.

Fig. 5 shows performance against decoding speed on the *rt07seval* IHM test data. The solid horizontal line marks HDecode performance by generation of bi-gram lattices, followed by 4-gram lattice re-scoring. The dashed line marks an identical configuration, but using a language model generated with entropy pruning as described in [17], and shows a degradation in performance by 0.5%. The remaining curves show results obtained with Juicer and the same pruned LM. The WER results obtained with the 2-pass strategy (at about 30 RTF) are matched in a single decoding pass at around 2 RTF. Equivalently, the 0.5% degradation from the best performance is also observed when using HDecode with the equivalent trigram LM. Table XVI shows WER results and RTF at different pruning levels. Surprisingly, little impact on WER is observed with larger LMs when RTF is roughly kept constant. This implies that the best strategy is to build the largest WFST possible regardless of target speed. Equivalently, Table XVII shows impact of dictionary sizes and n-gram order. The dictionary size can be reduced substantially while retaining modest degradation, allowing very compact decoding.

Juicer has a total of six pruning parameters that can be changed to influence processing speed. As grid search for such parameters is very costly we have made use of gradient based

TABLE XVI
EFFECT OF ENTROPY PRUNING AND USING FIXED BEAM SETTINGS. %WER AND RTF PERFORMANCE ON *rt*07seval MDM

Total n-grams in 4g LM	Arcs in WFST	WER	RTF
3.5M	15.6M	46.8	0.579
4.4M	19.5M	46.5	0.591
6.1M	26.6M	46.6	0.597
8.0M	35.2M	46.7	0.606

TABLE XVII
CHANGE OF LEXICON SIZE AND N-GRAM ORDER. %WER AND RTF PERFORMANCE ON *rt*07seval MDM

Lexicon size	LM order	Arcs in WFST	WER	RTF
2K	7	11.8M	55.3	0.827
6K	7	12.5M	48.2	0.625
10K	7	13.8M	47.2	0.582
16K	7	14.7M	46.8	0.589
50K	4	15.6M	46.8	0.579

search methods described in [32] to find the best possible configurations. As we found clear dependence of such parameters on the acoustic and language models used, optimal operating curves were generated for each acoustic model/language model combination.

VII. SYSTEM DESIGN AND RESULTS

The construction of offline ASR systems is mostly governed by raw performance and much less focus is put on processing speed. The objective in NIST RT evaluations again was on optimal WER result and hence multi-pass recognition strategies are standard. Two fundamental concepts govern the design of such ASR systems: adaptation and the use of complimentary system output. The systems outlined in this paper make use of adaptation in the form of cepstral mean and variance normalization per speaker, VTLN, and maximum-likelihood linear regression (MLLR) [33]. The AMIDA systems use complementary system output created in several ways, by employing different language models and vocabularies, acoustic training data and acoustic model training as well as different features. The various outputs can then be used in cross-adaptation, i.e., making use of the output of one pass to adapt another model set. The second option is system combination using confusion network combination or ROVER.

Fig. 6 shows an outline of the RT'07 system where the main design paradigm was cross-adaptation. A total of four different acoustic model configurations are used. Standard PLP/HLDA models (M1) are required for the initial stages to allow estimation of VTLN warp factors. Models constructed on PLP/LCRC (M2) and MFCC/BN (M3) features are designed to yield slightly different output operating at similar word error rates. The fourth model set is based on CTS adaptation as outlined in Section V-B whereas all other models are trained on meeting data only. The figure also highlights the fact that the first best output, lattices and confusion networks are passed between the stages. Table XVIII shows results for several key passes. The

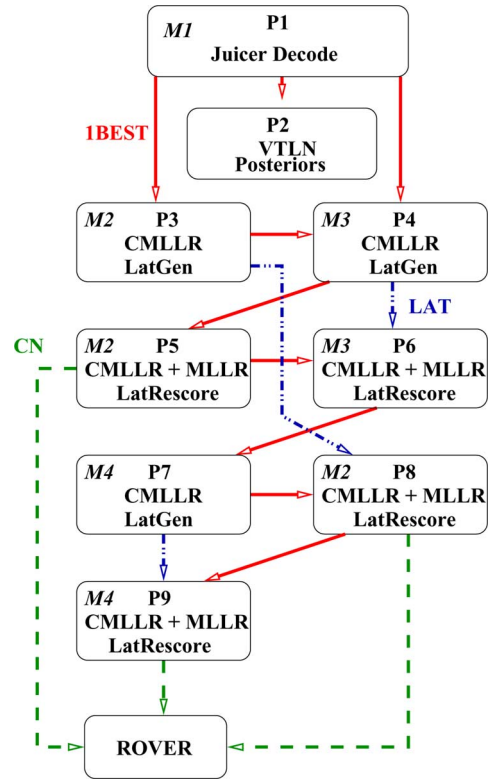

Fig. 6. AMIDA RT'07 system schematic. `LatGen` denotes lattice generation, `LatRescore` acoustic rescoring.

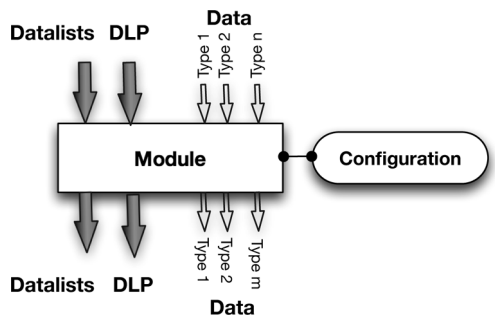

Fig. 7. ROTK modules. Data lists hold metadata such as speaker or segment information, DLP files list data lists and processing information.

TABLE XVIII
%WER RESULTS ON *rt*07seval FOR SEVERAL PASSES OF THE AMIDA 2007 AND 2009 IHM SYSTEMS

System	Pass	Tot	CMU	EDI	NIST	VT
2007	P1	37.4	47.7	29.3	33.8	38.4
	P5 CN	25.9	35.1	20.4	21.8	25.7
	Final	24.9	33.9	19.8	20.9	24.7
2009	Initial	38.1	51.7	30.2	32.5	37.5
	SBN/fMPE	24.5	35.1	18.1	20.5	23.9
	Final	23.4	33.8	17.5	19.1	22.8

difference between the initial and final passes is considerable but most of the gain is achieved in the first stages of cross adaptation. Without cross-adaptation, 2% absolute poorer performance is obtained [6]. Most notably the expensive training

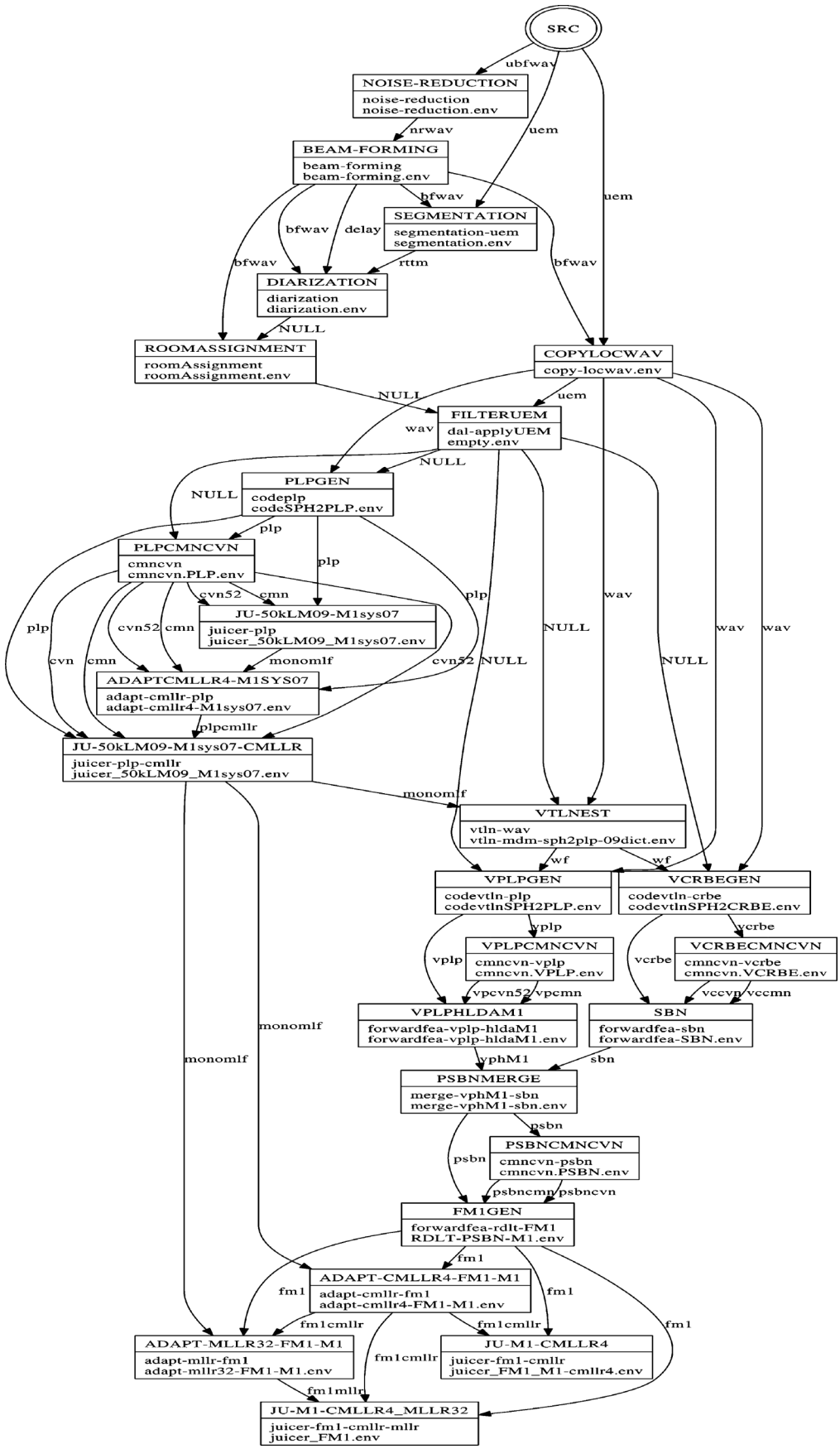

Fig. 8. Graph of the 2009 MDM system. Arcs describe data types; NULL indicates transport of metadata only. Boxes hold module instance, module, and configuration names.

TABLE XIX
%WER ON *rt*09seval IHM FOR THE AMIDA RT'09 SYSTEM. IDI/EDI/NIST ARE MEETING ROOMS

Description			Automatic				Manual			
LM	AM	Notes	Tot	IDI	EDI	NIST	Tot	IDI	EDI	NIST
6kLM09-7g	M2		41.3	45.1	32.3	44.9	38.3	44.0	31.9	38.3
50kLM09-4g	M1		45.9	50.9	36.8	48.3	43.7	50.2	36.8	43.3
50kLM09-4g	M2	CMLLR	36.4	38.8	28.5	40.2	32.9	37.9	27.7	32.5
50kLM09-4g	M3	CMLLR	28.3	28.5	21.4	33.2	24.2	27.8	21.1	23.5
50kLM09-4g	M4	Lattices / MLLR	27.6	28.3	20.9	31.9	23.9	27.9	20.6	22.8
50kLM09-4g	M3	Rescore / MLLR	27.2	28.0	20.3	31.9	23.5	27.5	20.0	22.6
Confusion network			27.4	28.6	20.4	31.6	23.8	28.0	20.7	22.5

of CTS adapted models only gave small further improvements. These results gave rise to a completely different strategy developed for the AMIDA 2009 systems. First, the system software was cast in the form of the resource optimization toolkit (ROTK). Here, the system is compiled from a set of fundamental modules, for example: PLP computation; decoding using a specific configuration; adaptation; or segmentation. Fig. 7 illustrates the input and output of a module. Apart from processing raw data, modules also provide metadata, e.g., speaker labels, in a systematic way and each module has certain data types it can process, e.g., raw audio files. The actual processing is specified by a module configuration file. Hence, the same module can operate differently, for example by using different acoustic or language models for decoding. System implementation then can be performed by simply writing a module graph of the form displayed in Fig. 8 which displays the 2009 AMIDA MDM system. Most branches in this graph deal with front-end processing. In contrast, The IHM system graph (not shown) is considerably larger and contains cross-adaptation and lattice rescoring.

Aside from the data flow the toolkit also automatically organizes distribution of processing on a computing grid. While these are implementation details the representation also allows to build graphs semi-automatically by finding modules that would fit into the graph. Multiple module combinations can then be tested automatically, replacing the manual design process. This approach was adopted for parts of the graph generation for the RT'09 systems to find optimal parameter configurations. Aside from the new modeling outlined in previous sections, adaptation was also slightly altered. Under normal NIST RT evaluation regimes, only 10 minutes of a meeting form a test set. However the complete meeting, typically longer than 30 minutes can be used for adaptation purposes. Little gain was found by direct use of more data. Instead, complementary decoding was performed to find segments of agreement, to only use those for adaptation purposes. In this process, typically half of the data is discarded and small but consistent gains were obtained. For IHM the acoustic models developed were: HLDA-PLP/ML (M1) and HLDA-PLP/MPE (M2), VTLN/SBN/MPE/fMPE (M3), VTLN/LCRCBN/MPE/fMPE (M4). For MDM, no LCRCBN models were created. The language models used are a 4-gram LM based on 50 K vocabulary, and a 7-gram LM with 6 K vocabulary. Table XVIII compares results for the IHM system on the *rt07seval* data with the 2007 system. An overall improvement of 1.5% WER absolute can be observed with a much simpler system that also now operates in about a fifth of the time, at 19.5 RTF. Table XIX shows the results in more detail for the various decoding passes for automatic and manual segmentation. One can clearly observe the difficulty with one NIST meeting as described in Section IV-A while the IDI meeting room recordings give poor results due to the use of non-noise cancelling microphones. Overall it is found that (in contrast to *rt07seval*) neither lattice rescoring nor confusion networks give significant gains. The M3 output is available at 9.8 RTF. MDM performance is shown Table XX. The difference between IHM and MDM is $\sim$20% WER relative, which is similar to that for manual segmentation. However, the cause for the two microphone cases is not the same: The difficulties of the NIST meeting for IHM are due to sound levels. By contrast, those for MDM are caused by the large number of very active speakers. Note that MDM scoring excludes overlapped speech.

TABLE XX
%WER ON MDM FOR THE AMIDA RT'09 SYSTEM. SEG DENOTES AUTOMATIC (AUT) OR REFERENCE (REF) SEGMENTATION AND I/F ARE INITIAL AND FINAL PASSES

		rt07seval				*rt09seval*			
Seg		Tot	Sub	Del	Ins	Tot	Sub	Del	Ins
Aut	I	40.3	25.1	11.1	4.2	44.2	28.7	10.8	4.7
	F	29.3	17.0	9.0	3.3	33.2	20.6	9.3	3.2
Ref	I	37.8	24.4	11.1	2.3	42.3	28.8	10.3	3.2
	F	26.5	16.4	8.4	1.6	30.7	20.3	8.3	2.1

VIII. CONCLUSION

In this paper, we have presented our work on meeting transcription for the AMIDA systems. We have outlined the basic properties of conference room meeting data and the core technologies tested and developed for the complete AMIDA systems. We put special emphasis on the specific aspects of the data, such as far field recognition, reverberant and noisy environments, but also language and lexical issues. While vocabulary and language modeling show some influence from diversity in topics, the main source of difficulty appears to originate from the complex acoustic environment. Acoustic modeling techniques are found to behave in similar way to the CTS task, with strong

gains from techniques such as VTLN, HLDA, MPE, and posterior features. However, adaptation from CTS models was found to only bring modest gains, partly due to bandwidth issues. We have outlined the benefit of essential technologies on NIST evaluation test sets in detail. Two approaches to system design are explained and the results for both systems are compared. Both systems are available for tests at www.webasr.org. For comparison to other state-of-the-art systems the interested reader is referred to publications by NIST on the outcome of RT evaluations which can be found on their web pages.

Although the results on NIST evaluation data are among the best achieved (see [34] or publications in this special issue), there is clearly substantial room for improvement. While more data for training will undoubtedly help, more fundamental issues on room acoustics and noise robustness need to be addressed for significant progress. The NIST paradigm is still somewhat optimistic, for example in terms of the number of speakers, the recording settings (e.g., quality and location of microphones) or the language and interaction style used in real-world meetings (most recordings are based on scenarios!). These are issues of robustness. The gap between close-talking and far-field performance is still wide, and much worse when using a single microphone.

ACKNOWLEDGMENT

The authors would like to thank A. Stolcke (SRI) and D. van Leeuwen (TNO) for their help with segmentation and D. Marino (University of Sheffield) for providing AMI corpus analysis.

REFERENCES

[1] J. Fiscus and J. Garofolo, "RT-2002 Evaluation Plan (Version 1.0)," NIST, 2002 [Online]. Available: http://www.itl.nist.gov/iad/mig//tests/rt/2002/index.html.

[2] A. Waibel, M. Bett, M. Finke, and R. Stiefelhagen, "Meeting browser: Tracking and summarizing meetings," in *Proc. DARPA Broadcast News Transcript. Understand. Workshop*, Lansdowne, VA, 1998, pp. 281–286.

[3] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings: The AMI and AMIDA projects," in *Proc. ASRU'07*, 2007, pp. 238–247.

[4] A. Popescu-Belis, J. Carletta, J. Kilgour, and P. Poller, "Accessing a large multimodal corpus using an automatic content linking device," in *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*. New York: Springer, 2009, vol. 5509, pp. 189–206.

[5] 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 *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science. Edinburgh, U.K.: Springer-Verlag, 2005, vol. 3869, pp. 450–462.

[6] 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," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science. New York: Springer-Verlag, 2007.

[7] S. Burger, V. MacLaren, and H. Yu, "The ISL meeting corpus: The impact of meeting type on speech style," in *Proc. ICSLP*, 2002.

[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," in *Proc. MLMI'05*, Edinburgh, U.K., 2005.

[9] J. Garofolo, C. Laprun, M. Miche, V. Stanford, and E. Tabassi, "The NIST meeting room pilot corpus," in *Proc. LREC'04*, 2004.

[10] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," in *Proc. IEEE ICASSP*, 2003, pp. 364–367.

[11] V. Wan and T. Hain, "Strategies for language model web-data collection," in *Proc. ICASSP'06*, 2006, pp. 1069–1072.

[12] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The AMI meeting transcription system: Progress and performance," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science. New York: Springer, 2006, pp. 419–431.

[13] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Machine Learning for Multimodal Interaction*, ser. LNCS. New York: Springer-Verlag, 2007, vol. 4625, pp. 509–519.

[14] T. Hain, G. G. J. Dines, M. Karafiat, D. Moore, V. Wan, R. Ordelman, and S. Renals, "Transcription of conference room meetings: An investigation," in *Proc. Interspeech'05*, 2005, pp. 1661–1664.

[15] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, F. Grezl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Machine Learning for Multimodal Interaction*, ser. LNCS. New York: Springer Verlag, 2005, pp. 463–475.

[16] I. Bulyko, M. Ostendorf, and A. Stolcke, "Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures," in *Proc. Human Lang. Technol. Conf.*, 2003.

[17] A. Stolcke, "Entropoy-based pruning of backoff language models," in *Proc. DARPA Broadcast News Transcript. Understand. Workshop*, 1998, pp. 270–274.

[18] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, V. Wan, and J. Vepa, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP'07*, 2007, vol. 1, pp. 357–360.

[19] J. Dines, J. Vepa, and T. Hain, "The segmentation of multi-channel meeting recordings for automatic speech recognition," *Proc. Interspeech'06*, 2006.

[20] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, UPC, Barcelona, Spain, 2006.

[21] D. A. van Leeuwen and M. Huijbregts, "The AMI speaker diarization system for IST RT06s meeting data," in *Proc. MLMI'06*, 2006, pp. 371–384.

[22] D. Povey, "Discriminative Training for Large Vocabulary Speech, Recognition," Ph.D. dissertation, Cambridge Univ., Cambridge, U.K., 2004.

[23] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. LREC'04: 4th Inte. Conf. Lang. Resources Eval.*, Lisbon, Portugal, 2004.

[24] M. Karafiat, L. Burget, T. Hain, and J. Cernocky, "Application of CMLLR in narrow band wide band adapted systems," in *Proc 8th Int. Conf. Interspeech'07*, Antwerp, Belgium, 2007, p. 4.

[25] J.-L. Gauvain and C.-H. Lee, "Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 291–298, Apr. 1994.

[26] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, John Hopkins Univ., Baltimore, MD, 1997.

[27] F. Grezl, M. Karafiat, and L. Burget, "Investigation into bottle-neck features for meeting speech recognition," in *Proc. Interspeech'09*, 2009, no. 9, pp. 2947–2950.

[28] D. Povey, "Improvements to fMPE for discriminative training of features," in *Proc. Interspeech*, 2005, pp. 2977–2980.

[29] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. Interspeech'06*, 2006.

[30] P. N. Garner, J. Dines, T. Hain, A. E. Hannani, M. Karafiát, D. Korchagin, M. Lincoln, V. Wan, and L. Zhang, "Real-time ASR from meetings," in *Proc. Interspeech*, 2009.

[31] P. N. Garner and J. Dines, "Tracter: A lightweight dataflow framework," in *Proc. Interspeech*, Makuhari, Japan, Sep. 2010.

[32] A. E. Hannani and T. Hain, "Automatic optimisation of speech decoder parameters," *IEEE Signal Processing Letters*, vol. 17, pp. 95–98, 2010.

[33] M. J. Gales and P. Woodland, "Mean and variance adaptation within the MLLR framework," *Comput. Speech Lang.*, vol. 10, pp. 249–264, 1996.

[34] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Machine Learning for Multimodal Interaction*, ser. Lecture Notes in Computer Science. New York: Springer-Verlag, 2007.

Thomas Hain (M'02) received the Dipl.-Ing. degree in electrical engineering from the Vienna University of Technology, Vienna, Austria, and the Ph.D. degree from the University of Cambridge, Cambridge, U.K.

He is a Senior Lecturer at the Speech and Hearing Group, University of Sheffield, Sheffield, U.K., and has worked on speech processing for more than 15 years. His main interests are in speech processing, machine learning, and natural man/machine interfaces.

Lukáš Burget (M'03) received the Ph.D. degree from the Brno University of Technology, Brno, Czech Republic.

He is an Assistant Professor in the Faculty of Information Technology, Brno University of Technology. He serves as Scientific Director of the Speech@FIT research group. His scientific interests are in acoustic modeling for speech, speaker, and language recognition.

John Dines (M'03) received the B.Eng. degree from the University of Southern Queensland, Brisbane, Australia, and the Ph.D. degree from the Queensland University of Technology, Brisbane.

Since 2003 he has been a Researcher at the Idiap Research Institute, Martigny, Switzerland. His main research interests include parametric speech synthesis, diarization, and speech recognition.

Philip N. Garner (M'96–SM'05) received the M.Eng. degree in electronic engineering from the University of Southampton, Southampton, U.K., in 1991.

He is a Senior Researcher at the Idiap Research Institute, Martigny, Switzerland. He has a mixed academic and industrial background in a broad range of speech and signal processing topics. His main research interests are in the overlap of speech signal processing, recognition, and synthesis.

František Grézl (M'03) received the Ph.D. degree from the Brno University of Technology, Brno, Czech Republic.

He has been with the ASP Group of OGI Portland, OR, the Speech Processing Group at Idiap Research Institute, Martigny, Switzerland, and ICSI, Berkeley, CA. He worked on probabilistic features based on TempoRAl Patterns and led the research on Bottle-Neck features.

Asmaa El Hannani received the M.Sc. degree from the University of Fribourg, Fribourg, Switzerland, and the Ph.D. degree from the Institut National des Telecommunication, Evry, France.

She joined the Speech and Hearing Research Group, University of Sheffield, Sheffield, U.K., in 2007. Her interests include biometrics technologies and speech processing.

Marijn Huijbregts received the Ph.D. degree in spoken document retrieval from the University of Twente, Enschede, The Netherlands.

He is a Post-Doctoral Researcher at Radboud University, Nijmegen, The Netherlands. His main interests are in speech/non-speech segmentation, speaker diarization, and multimedia retrieval.

Martin Karafiat received the Ph.D. degree from the Brno University of Technology, Brno, Czech Republic.

He is a Post-Doctoral Researcher in the Speech Group, Brno University of Technology. His main interests are in systems for speech recognition and acoustics modeling.

Mike Lincoln received the B.S. and Ph.D. degrees from the University of East Anglia, Norwich, U.K.

He is a Research Fellow at the Center for Speech Technology Research, University of Edinburgh, Edinburgh, U.K., and has worked on speech processing for more than ten years. His main interests are in microphone array processing, audio capture for speech recognition, and applications of speech technology.

Vincent Wan received the B.A. degree from the University of Oxford, Oxford, U.K., and the Ph.D. degree from the University of Sheffield, Sheffield, U.K.

He worked on speech recognition at the University of Sheffield, the Motorola Human Interface Labs, Palo Alto, and is now a member of the speech technology group at Toshiba Research Europe. His interests include machine learning, biometrics, and speech processing.