6

Automatic Speech Recognition

Gerasimos Potamianos,[1] Lori Lamel,[2] Matthias Wölfel,[3] Jing Huang,[1] Etienne Marcheret,[1] Claude Barras,[2] Xuan Zhu,[2] John McDonough,[3] Javier Hernando,[4] Dusan Macho,[4] Climent Nadeu[4]

[1] IBM T.J. Watson Research Center, Yorktown Heights, NY, USA
[2] LIMSI-CNRS, Orsay, France
[3] Universität Karlsruhe (TH), Interactive Systems Labs, Fakultät für Informatik, Karlsruhe, Germany
[4] TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain

Automatic speech recognition (ASR) is a critical component for CHIL services. For example, it provides the input to higher-level technologies, such as summarization and question answering, as discussed in Chapter 8. In the spirit of ubiquitous computing, the goal of ASR in CHIL is to achieve a high performance using far-field sensors (networks of microphone arrays and distributed far-field microphones). However, close-talking microphones are also of interest, as they are used to benchmark ASR system development by providing a best-case acoustic channel scenario to compare against.

Although ASR is a well-established technology, the CHIL scenario presents significant challenges to state-of-the-art speech recognition systems. This is due to numerous reasons, for example, the presence of speech from multiple speakers with varying accents and frequent periods of overlapping speech, a high level of spontaneity with many hesitations and disfluencies, and a variety of interfering acoustic events, such as knocks, door slams, steps, cough, laughter, and others. Note that the problem of identifying such acoustic events has also been investigated in CHIL, and it is discussed in Chapter 7. In addition, the linguistic content in CHIL scenarios, that of technical seminars, constitutes another challenge, since there exists only a relatively small amount of in-domain acoustic and language modeling data. The focus on handling far-field microphone speech exacerbates these issues, due to the low signal-to-noise ratios and room reverberation.

Of course, these challenges also affect *speech activity detection* (SAD), *speaker identification* (SID), and *speaker diarization* (SPKR) – also known as the "who spoke when" – problems. These technologies jointly address the "what", "when", and "who" of human interaction. In particular, SAD and SPKR constitute crucial components of state-of-the-art ASR systems, and as such they are briefly discussed in this chapter. In addition, they are relevant to other CHIL perceptual technologies, for example, acoustic-based speaker localization and identification, which are discussed in more detail in Chapters 3 and 4, respectively.

Progress of the ASR systems developed for the CHIL domain has been benchmarked by yearly technology evaluations. During the first two years of CHIL, the Consortium internally evaluated ASR with a first dry run held in June 2004, followed by an "official" evaluation in January 2005. Following these two internal campaigns, the CHIL sites involved in the ASR activity (IBM, LIMSI, and UKA-ISL) participated in the Rich Transcription evaluations of 2006 and 2007 – RT06s [13] and RT07 [45]. These international evaluations were sponsored by NIST and attracted a number of additional external participants.

The CHIL partner sites involved in ASR work have made steady progress in this technology in the CHIL domain. For example, in the far-field ASR task, the best system performance improved from a word error rate of over 68% in the 2004 dry run to approximately 52% in the CHIL 2005 internal evaluation, and from 51% in RT06s down to 44% in RT07. These improvements were achieved in spite of the fact that the recognition task became increasingly more challenging: Indeed, from 2005 to 2006, multiple recording sites involving speakers with a wider range of nonnative accents were introduced into the test set. Furthermore, in both 2006 and 2007, the degree of interactivity in the data was significantly increased.

This chapter documents progress in the challenging task of recognizing far-field speech in the CHIL scenarios, and it highlights the main system components and approaches followed by the three CHIL partners involved in this effort. Section 6.1 provides an overview of the ASR problem and its evaluation framework in CHIL, as well as the data resources available for system development. A brief discussion of the SAD and SPKR subsystems appears in Section 6.2. Section 6.3 describes the main components of ASR systems with highlights of specific approaches investigated in CHIL within each of these components. An example of an ASR system implementation is provided in Section 6.4, followed by a brief overview of ASR experimental results in Section 6.5. Finally, the chapter concludes with a summary, a discussion of open problems, and directions for future research (Section 6.6).

6.1 The ASR Framework in CHIL

Automatic speech recognition in CHIL constitutes a very challenging problem due to the interactive scenario and environment as well as the lack of large corpora fully matched to the CHIL specifics. In the next subsections, we briefly overview these two factors by providing more details of the ASR evaluation framework in CHIL as well as the available data resources utilized in training the ASR systems.

6.1.1 ASR Evaluation Framework

The CHIL interactive seminars were held inside smart rooms equipped with numerous acoustic and visual sensors. The former include a number of microphone arrays located on the walls (typically, at least three T-shaped four-microphone arrays and at least one linear 64-channel array) as well as a number of tabletop microphones. In addition, most meeting participants wore headset microphones to capture individual

speech in the close-talking condition. This setup was installed in five CHIL partner sites and used to record data for the RT06s and RT07 speech technology evaluation campaigns. It also represents a significantly more evolved setup than the initial smart room design installed at the UKA-ISL site to provide data for the 2004 and 2005 CHIL internal evaluations. For example, in the 2004 dry-run evaluation, the far-field audio signals were captured only by a 16-channel linear array and a single tabletop microphone.

The above setup has been designed to allow data collection and ASR evaluation with main emphasis on the use of unobtrusive far-field sensors that fade into the background. These data form the basis of the *multiple distant microphone* (MDM) condition, the designated primary condition in the RT06s and RT07 ASR technology evaluations, where all tabletop microphones – typically ranging from three to five – were utilized to yield a single transcript. Additional conditions are the *single distant microphone* (SDM) one, where only one preselected tabletop microphone is used, as well as a number of conditions that involve using the linear or T-shaped arrays [45].

An interesting contrasting condition is called the *individual headset microphone* (IHM) condition, where the data recorded on the channels from the headsets worn by all lecture participants are decoded, with the purpose of recognizing the wearer's speech. This represents a close-talking ASR task, and putting aside for a moment the challenging task of robust cross-talk removal, it is designed to quantify the ASR performance degradation due to the use of far-field sensors.

Because of the reduced sensory setup in the initial smart room, the conditions were somewhat different in the 2004 and 2005 CHIL internal evaluations. In particular, only the lecturer's speech was decoded in these evaluations, for both the close-talking and far-field conditions. Furthermore, in 2004, only the 16-channel linear array was used in the far field.

6.1.2 Data

A number of seminars and interactive lectures were collected in the five state-of-the-art smart rooms of the CHIL Consortium. Nevertheless, the available amount of CHIL data remains insufficient for training ASR systems, comprising less than 10 hours of speech. This issue was, of course, even more pronounced in the early CHIL evaluations, since the data have been incrementally collected over the duration of the project. To remedy this problem, additional publicly available corpora [31] exhibiting similarities to the CHIL scenarios were utilized for system development. These are the ICSI, ISL, and NIST meeting corpora [24, 7, 14], the additional non-CHIL meeting data from prior RT evaluation runs (2004-2006), including a corpus collected by NIST in 2007, data collected by the AMI Consortium [1], and the TED corpus of lectures collected at the Eurospeech Conference in Berlin in 1993 [50, 30]. Most data sets contain close-talking (headset or lapel) and multiple far-field microphone data, with the exception of TED that contains lapel data only. In total, there are on the order of 250 hours of speech in the combined corpora.

The three CHIL sites used various parts of these sources in their acoustic model-building process, as all these corpora exhibit certain undesirable variations from the

CHIL data scenario and acoustic environment. In particular, since only a portion of the TED corpus is transcribed, the remainder was exploited by some CHIL partners (e.g., LIMSI) via unsupervised training [29]. It is also worth noting that in earlier ASR systems developed for the CHIL internal runs in 2004 and 2005, some partners also used other corpora, such as Broadcast News (LIMSI), or even proprietary data sets such as the ViaVoice and MALACH project corpora (IBM), after applying necessary model adaptation.

For language model training, transcripts from the CHIL data sets as well as the above-mentioned meeting corpora were used. In addition, LIMSI generated a set of cleaned texts from published conference proceedings; these are also very relevant to the task due to the technical nature of CHIL lectures, which were employed in some of the developed ASR systems. Furthermore, some sites used Web data, for example, data available from the EARS program [31], and possibly additional sources such as data from conversational telephone speech, e.g., the Fisher data [31].

6.2 ASR Preprocessing Steps

Two important preprocessing stages in all CHIL partners' ASR systems are the speech activity detection (SAD) and speaker diarization (SPKR) components. These locate the speech segments that are processed by the ASR systems and attempt to cluster homogeneous sections, which is crucial for efficient signal normalization and speaker adaptation techniques. As mentioned in the introduction, these components have been evaluated in separate CHIL internal and NIST-run evaluation campaigns. Within the Consortium, they have attracted significant interest among the CHIL partners in addition to the three sites involved in ASR system development.

6.2.1 Speech Activity Detection

Speech activity detection has long been an important topic as a front-end step to the ASR process, having a positive impact on ASR systems in terms of both CPU usage and ASR accuracy. This is due to the fact that the decoder is not required to operate on nonspeech segments, thus reducing the processing effort and word insertion error rate. In addition, SAD systems provide the segments used as input to the speaker diarization component (see ahead). Robust performance of SAD is also important in other technologies of interest in CHIL, such as acoustic speaker localization.

Not surprisingly, the SAD technology has been investigated by many CHIL partners. At some stage during the CHIL project, AIT, FBK-irst, IBM, INRIA, LIMSI, UKA-ISL, and UPC developed SAD systems for CHIL. For instance, all these partners participated in the 2005 CHIL internal evaluation of the technology, a campaign that followed the dry run conducted in the summer of 2004. More recently, the technology was evaluated stand alone in RT06s (AIT, FBK-irst, IBM, INRIA, LIMSI, and UPC participated), and as a component of SPKR systems in RT07 (IBM, LIMSI, and UPC took part); both campaigns were organized by NIST. At the RT06s evaluation, the best systems achieved very encouraging error rates of 4% and 8% for the

conference and lecture subtasks, respectively, when calculated by the NIST diarization metric.

For SAD system development, CHIL partners followed various approaches that differed in a number of factors, for example, in feature selection [Mel frequency cepstrum coefficients (MFCCs), energy-based, combined acoustic and energy-based features, etc.], the type of classifier used [Gaussian mixture model classifiers (GMMs), support vector machines, linear discriminants, decision trees], the classes of interest (IBM initially used three broad classes), and channel combination techniques (based on signal-to-noise ratios, voting, etc.). Details can be found in a number of partner site papers, for example, [35, 38, 22].

6.2.2 Speaker Diarization

Speaker diarization, also referred to as the "who spoke when" task, is the process of partitioning an input audio stream into homogeneous segments according to speaker identity. It is useful as an ASR preprocessing step because it facilitates unsupervised adaptation of speech models to the data, which improves transcription quality. It also constitutes an interesting task per se, since structuring the audio stream into speaker turns improves the readability of automatic transcripts. A review of activities in speaker diarization can be found in [51].

Historically, SPKR systems were evaluated by NIST on Broadcast News data in English up to 2004; following that, the meeting domain became the main focus of the RT evaluations. These included CHIL lecture seminars and multisite meetings in the 2006 and 2007 evaluations [13, 45]. Similarly to ASR, a number of evaluation conditions have been defined in these campaigns (e.g., MDM and SDM conditions). Notice that in the adopted evaluation framework, the number of speakers in the recording or their voice characteristics are not known a priori; therefore, they must be determined automatically. Also, SPKR systems were evaluated as independent components. Clearly, the use of other sources of knowledge, such as output of multimodal person tracking and identification, could dramatically improve system accuracy. Without such additional information, diarization of audio data recorded by distant microphones remains a very challenging task, for the same reasons that apply to the far-field ASR problem in CHIL, as discussed in the introduction.

A number of CHIL partner sites developed SPKR systems (AIT [43], IBM [22, 20], LIMSI [59, 60], and UPC [34]). The following specific research directions were addressed in these systems:

- *Exploiting multiple distant microphone channels:* Acoustic beamforming was performed on the input channel after Wiener filtering, using a delay-and-sum technique [34, 4]. Up to 20% relative gain was obtained by using the beamformed audio, compared to using a single channel on conference data, even if the gain on lecture data was less significant [60]. Using the delays between the acoustic channels as features appears also to be a very promising direction [39].
- *Speech parameterization:* Frequency filtering, which showed good results in the CLEAR 2007 evaluation in the acoustic person identification task, was used as

an alternative to the classical MFCC features [34]. Derivative parameters were tested but did not seem of much benefit [60].

- *Speech activity detection:* SAD errors directly affect SPKR system performance. Therefore, additional effort was placed toward improving SAD models, as discussed in a previous section. When the diarization task is considered standalone, a different balance has to be chosen between missed speech and false-alarm speech than when SAD is used as an ASR preprocessing step. An explored solution was a purification of the acoustic segments using an automatic word-level alignment in order to reduce the amount of silence or noise portions, which are potentially harmful during clustering [20].
- *Segmentation and clustering:* An initial step provides an overestimated number of clusters; each cluster is modeled with a Gaussian model (typically a single Gaussian with a full covariance matrix or a GMM with diagonal covariance matrices). Clusters are further grouped following a distance (Mahalanobis, likelihood gain [20], a Bayesian information criterion (BIC) measure [34], cross log-likelihood ratio [60]) until some threshold is reached.

Results in the RT07 evaluation campaign demonstrate that the speaker diarization problem is far from being solved in the CHIL scenarios. In particular, the best SPKR system (developed by LIMSI) was benchmarked at a 26% diarization error rate for the MDM condition. This is significantly worse than the diarization error typically achieved in the Broadcast News task – about 10% [5].

6.3 Main ASR Techniques and Highlights

Although the three CHIL sites have developed their ASR systems independently of each other, all systems contain a number of standard important components, which are summarized in this section. In addition, in the spirit of collaboration through competition in technology evaluation – the so-called co-opetition paradigm that has been adopted in the CHIL project as a means to promote progress – CHIL partners have shared certain components, such as the UKA-ISL beamforming algorithm for far-field acoustic channel combination or the LIMSI proceedings text corpora for language model training.

6.3.1 Feature Extraction

Acoustic modeling requires the speech waveform to be processed in such a way that it produces a sequence of feature vectors with a relatively small dimensionality in order to overcome the statistical modeling problem associated with high-dimensional feature spaces, called the *curse of dimensionality* [6]. Feature extraction in ASR systems aims to preserve the information needed to determine the phonetic class, while being invariant to other factors including speaker differences such as accent, emotion, fundamental frequency, or speaking rate, as well as other distortions, for example, background noise, channel distortion, reverberation, or room modes. Clearly,

this step is crucial to the ASR system, as any loss of useful information cannot be recovered in later processing.

Over the years, many different speech feature extraction methods have been proposed. The methods are distinguished by the extent to which they incorporate information about the human auditory processing and perception, robustness to distortions, and length of the observation window. Within the CHIL framework, different state-of-the-art feature extraction methods have been investigated. For example, ASR systems have utilized MFCCs [9] or perceptual linear prediction (PLP) [18] features. Feature extraction in CHIL partner systems often involved additional processing steps, for example, linear discriminant analysis (LDA) [17] or a maximum likelihood linear transform (MLLT) [16]. Feature normalization steps, such as variance normalization and vocal tract length normalization (VTLN) [3] were also employed by some sites.

In addition to the above, a novel feature extraction technique has been developed by UKA-ISL that is particularly robust to changes in fundamental frequency, f_0. This is important in the CHIL scenario, as public speeches have a higher variance in f_0 than do private conversations [19]. Additional advantages of the proposed approach, based on a warped-minimum variance distortionless response spectral estimation [56], are an increase in resolution in low-frequency regions relative to the traditionally used Mel filter banks, and the dissimilar modeling of spectral peaks and valleys to improve noise robustness, given that noise is present mainly in low-energy regions. To further increase the robustness to noise, a signal adaptive front end has been proposed [52], that emphasizes classification of relevant characteristics, while classification-irrelevant characteristics are alleviated according to the characteristics of the input signal; for example, vowels and fricatives have different characteristics and should therefore be treated differently. Experiments conducted by UKA-ISL have demonstrated that the proposed front ends reduce the *word error rate* (WER) on close-talking microphone data by up to 4% relative, and on distant speech by up to 6% relative, as compared to the widely used MFCC features [58].

6.3.2 Feature Enhancement

Feature enhancement manipulates speech features in order to retrieve features that are more similar to the ones observed in clean training data of the acoustic model. Thus, the mismatch between the unknown, noisy environment and the clean training data is reduced. Speech feature enhancement can be realized either as an independent preprocessing step or on the features within the front end of the ASR system. In both cases, it is not necessary to modify the decoding stage or acoustic models of the ASR system.

Most popular feature enhancement techniques for speech recognition operate in the frequency domain. Simple methods such as spectral subtraction are limited to removing stationary noise, where the spectral noise floor is estimated on noise-only regions. More advanced methods attempt to track either the clean speech or the noise for later subtraction. First approaches in this direction used Kalman filters (KFs) [25], which assume the relationship between the observations and the inner state to be

linear and Gaussian, which does not hold in practice. To overcome this constraint, variants to the KF, such as the extended KF, have been proposed.

Research in CHIL has focused on the enhancement of features in the logarithmic Mel-spectra domain by tracking the noise with a particle filter (a.k.a. sequential Monte Carlo method) [44, 48]. Feedback of the ASR system into the feature enhancement process results in further improvements by establishing a coupling between the particle filter and the ASR system, which had been treated as independent components in the past [11]. Experiments conducted by UKA-ISL using a novel feature enhancement technique that is able to jointly track and remove nonstationary additive distortions and late reverberations have demonstrated that word accuracy improvements on distant recordings by more than 20% relative are possible [54], independent of whether the acoustic models of the speech recognition system are adapted.

6.3.3 Acoustic Modeling

For acoustic modeling in CHIL, *hidden Markov models* (HMMs) were exclusively used. Most sites estimated system parameters by the expectation maximization (EM) algorithm (maximum-likelihood training) [10], followed by discriminative model training using the maximum mutual information (MMI) [41] or the minimum phone error (MPE) approach [42]. A number of adaptation techniques were also employed, ranging from maximum a posteriori estimation (MAP) [15] to maximum-likelihood linear regression (MLLR) [32], feature space MLLR (fMLLR), or speaker adaptive training (SAT) [2]. The above approaches require a multipass decoding strategy, where a word hypothesis is used for unsupervised model adaptation prior to the next decoding pass. Finally, some sites developed systems with slight variations to improve final system performance through combination or cross-system adaptation.

An additional area of interest in acoustic modeling is that of pronunciation modeling. The pronunciation dictionary is the link between the acoustic and language models. All CHIL sites used phone representations in their systems, with about 40 to 50 phoneme-like units. Special phone symbols were also sometimes used to model nonspeech events such as hesitation, cough, silence, etc. Each lexical entry of the word dictionary can then be associated with one or more pronunciations to explicitly model frequent variants. It is common practice to include some acronyms, compound words, or word sequences in order to capture some of the coarticulation in spontaneous speech. These multiwords typically represent only a small part of the vocabulary. Some of the CHIL sites (e.g., LIMSI) also explored explicitly including pronunciation variants for nonnative accented speech; however, while the variants better represented the foreign accents, the overall recognition performance did not improve.

6.3.4 Language Modeling

For language modeling, different n-gram language models (LM), with $n = 3$ or 4, have been employed by the CHIL sites. These LMs were typically developed separately for various data sources and were subsequently linearly interpolated to give

rise to a single model. Most often, CHIL and other meeting corpora were employed for this task. In addition, sometimes text from scientific proceedings (close to the CHIL lecture subjects) or data mined from the Web were also used. Based on these LMs, typical perplexities of the CHIL test data ranged in the order of 105 to 140. In terms of vocabulary size, CHIL sites used anywhere from 20k to 60k vocabularies, achieving out-of-vocabulary (OOV) rates in the order of 0.5 to 2.0%.

The use of a connectionist LM [47], shown to be performant when LM training data are limited, was explored at LIMSI. The basic idea is to project the word indices onto a continuous space and to use a probability estimator operating on this space. Since the resulting probability densities are continuous functions of the word representation, better generalization to unknown *n*-grams can be expected. A neural network LM was trained on the transcriptions of the audio data and proceedings texts and interpolated with a standard back-off LM. A significant word error reduction of 1.6% absolute was obtained when rescoring word lattices in under $0.3\times$RT.

An important part relevant to the language modeling work is the determination of the recognition vocabulary. The recognizer word list is usually determined by combining all the distinct words in the available audio transcriptions with the most frequent words in the relevant text sources. It is common practice to require a minimum number of word observations to be included in the word list. This ensures that the word occurs often enough to warrant modeling and also reduces the number of "false words" arising from typographical errors. Some text preprocessing is generally carried out to ensure conformity of the various text sources, removing undesirable data (email, addresses, mathematical formulas and symbols, figures, tables, references), formatting characters, and ill-formed lines. Acronyms, numbers, and compound words are also processed to ensure consistency and to approximate spoken language. The word list is typically selected so as to minimize the OOV rate on a set of development data. It was recently proposed to select the most probable words by linear interpolation of the unigram language models obtained from individual data sources.

6.3.5 Multiple Microphone Processing

The use of multiple microphones is an important component of far-field ASR in CHIL, as the sound pick-up quality might vary at different spatial locations and directions. An appropriate selection or combination of the different microphones can improve the recognition accuracy. The degree of success depends on the quality and variance of information provided by the microphones and the combination method used.

Speech recognition channel combination techniques can be broadly classified into signal and word-based combination methods. *Signal combination algorithms*, such as *blind source separation* and *beamforming* techniques, exploit the spatial diversity resulting from the fact that the desired and interfering signal sources are located at different points in space. This diversity can be taken advantage of by suppressing signals coming from directions other than the desired source direction. Those approaches assume that the speaker's position (time delay of arrival between

different microphones) can be reliably estimated, and it might employ knowledge about the microphone positions relative to each other. Correct speaker localization is crucial for optimal recognition accuracy [57]. Due to the reduction of multiple channels into one channel, the decoding time is not significantly changed, compared to that of a single-microphone approach.

In contrast to signal combination, *word based-combination techniques*, such as ROVER [12] and confusion network combination (CNC) [36], fuse information from the recognition output of different systems that can be represented as a one-best, *n*-best, or lattice word sequence, augmented by confidence scores. Word-based approaches assume that the transcription of different microphone channels leads to different word hypotheses. Their advantage is that no spatial information of the speaker or microphones is required. However, since each microphone channel is decoded independently, these approaches are computationally expensive. A hybrid approach, where the beamformed channel is augmented by additional channels and combined with CNC, has been shown to lead to additional improvements over either of the other approaches [55].

Due to the broad variance of the different microphone channels, it may not be optimal to blindly consider all channels for combination (e.g., if a microphone is directly placed near a sound source). It may instead be preferable to measure the quality of the different microphones and select only "good" channels. A traditional measure to achieve such selection is the *signal-to-noise ratio* (SNR). More reliable measures consider the properties of the human auditory system and/or operate on the features of the recognition system. One promising approach in this direction is based on class separability [53], which shows significant improvements over SNR-based channel selection methods.

Employing multiple microphones has been shown to improve word accuracy by up to 10% absolute, which compensates for approximately one third of the reduction in WER observed when moving a single microphone from the mouth region of the speaker (close talk) into the room.

6.4 An ASR System Example

Following the overview of the main approaches used in CHIL for ASR, we proceed with a more detailed description of the ASR systems developed by one of the three CHIL partners, IBM.

The IBM ASR systems for CHIL have progressed significantly over the duration of the project. In particular, during the first two project-internal evaluations (2004 and 2005), the IBM team focused on combining in-house available ASR systems, appropriately adapted to the available CHIL data [8]. However, it soon became apparent that this approach yielded a poor performance in the CHIL task; as a result, new systems trained exclusively on meeting-like corpora were developed for the RT06s and RT07 evaluations [23, 21]. The new approach was based on developing a small number of parallel far-field ASR systems (typically three or four) with minor variations in their acoustic modeling, and combining them using ROVER (for the close-talking

condition, a single system was developed). Additional work has been carried out for language modeling in order to create larger and richer LMs, suitable for the CHIL tasks. More details follow.

6.4.1 Acoustic Modeling

For acoustic modeling, first a speaker-independent (SI) model is trained, based on 40-dimensional acoustic features generated by an LDA projection of nine consecutive frames of 13-dimensional perceptual linear prediction (PLP) features, extracted at 100 Hz. The features are mean-normalized on a per-speaker basis. The SI model uses continuous-density, left-to-right HMMs with Gaussian mixture emission distributions and uniform transition probabilities. In addition, the model uses a global semi-tied covariance linear transformation [46], updated at every EM training stage. The system uses 45 phones; namely, 41 speech phones, one silence phone, and three noise phones. The final HMMs have 6k context-dependent tied states and 200k Gaussians. Since only a small part of the training data is from CHIL, MAP-adaptation of the SI model was deemed necessary to improve performance on CHIL data.

The SI features are further normalized with a voicing model (VTLN) with no variance normalization. The most likely frequency warping is estimated among 21 candidate warping factors ranging from 0.8 to 1.2. A VTLN model is subsequently trained on features in the VTLN warped space. The resulting HMMs have 10k tied states and 320k Gaussians. Following VTLN, a SAT system is trained on features in a linearly transformed feature space resulting from applying speaker-dependent fMLLR transforms to the VTLN-normalized features. Following SAT, feature space minimum phone error (fMPE) transforms are estimated [40], followed by MPE training [42] and MAP-MPE on the available amount of CHIL-only data [23, 21].

Following the above training procedure, two systems are built, one with the VTLN step present, and one with VTLN removed. Based on the latter, two additional SAT systems are built using a randomized decision-tree approach [49].

In contrast to the far field, only one system has been developed for the close-talking condition. This is identical in both RT06s and RT07 evaluations and is a 5k-state, 240k Gaussian mixture HMM system with both VTLN and variance normalization present [23].

6.4.2 Language Modeling

Five separate four-gram LMs were built. The first four were also used in the IBM RT06s system and were based on CHIL data (0.15M words), non-CHIL meetings (2.7M), scientific conference proceedings (37M), and Fisher data (3M words) [31]. A novel fifth LM used 525M words of Web data available from the EARS program [31]. For decoding, two interpolated LMs were used based on these five models. A reduced-size model was pruned to about 5M n-grams and was employed for static decoding, whereas a larger 152M n-gram model was used in conjunction with an on-the-fly dynamic graph expansion decoding. A 37k-word vocabulary was used.

6.4.3 Recognition Process

After speech segmentation and speaker clustering, a final system output was obtained in three decoding passes for each microphone: (a) an initial SI pass using MAP-adapted SI models to decode; (b) employing output from (a), warp factors using the voicing model and fMLLR transforms are estimated for each cluster using the SAT model. The VTLN features after applying the fMLLR transforms are subjected to the fMPE transform, and a new transcript is obtained by decoding, using the MAP-adapted MPE model and the fMPE features. (c) The output transcripts from step (b) are used in a cross-system fashion to estimate MLLR transforms on the MPE model. The adapted MPE model together with the large LM is used for final decoding with a dynamic graph expansion decoder.

In the far field, where multiple ASR systems have been developed, ROVER over these systems is applied for obtaining the SDM condition output. For the MDM condition, ROVER is first applied over all available tabletop microphones, followed by ROVER over the available four systems.

6.5 Experimental Results

The work described in Section 6.4 has resulted in significant progress over the duration of the CHIL project. For example, in the far-field condition, the initial IBM approach yielded 64.5% and 70.8% WER in the 2004 and 2005 CHIL-internal evaluations, respectively. In contrast, performance improved significantly in RT06s and RT07, reaching 50.1% and 44.3% WER, respectively, for the IBM MDM systems. In the close-talking task, under manual segmentation, system improvement has been less dramatic: In 2004, a 35.1% WER was achieved by the IBM system, whereas in 2005, a 36.9% WER was recorded. The new system development for the RT evaluation runs improved performance to 27.1% in 2006 and 31.7% in 2007. The latter represented a slight degradation, due to the more challenging nature of the 2007 data, and the lack of time for retraining the close-talking acoustic model in the IBM system.

In addition to IBM, ASR systems developed by LIMSI and UKA-ISL also achieved significant milestones over the duration of the CHIL project. For example, LIMSI achieved the lowest WERs in the 2005 CHIL-internal evaluation for both close-talking and far-field conditions, whether the UKA-ISL ASR system yielded a 26.7% WER in the IHM (close-talking) condition at the RT06 evaluation.

Overall, as mentioned at the beginning of this chapter, the CHIL Consortium consistently improved ASR technology over the duration of the project. This is clearly depicted in Fig. 6.1, where the lowest WER of all developed and evaluated far-field ASR systems is depicted over the four technology evaluations. This progress is especially noteworthy due to the fact that the ASR task has become increasingly more challenging over time. In particular, between 2005 and 2006, the number of recording sites increased from one to five, and the task modified to cover ASR for all seminar participants. Furthermore, between 2006 and 2007, the seminar interactivity increased significantly, with more focus placed on meeting-like, interactive seminars.

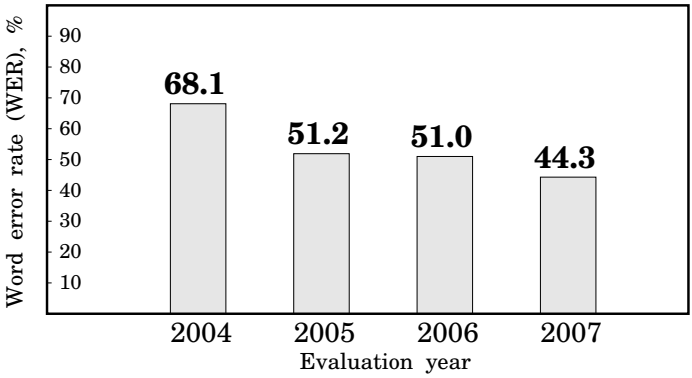

Fig. 6.1. Far-field ASR performance over the duration of the CHIL project.

6.6 Conclusions and Discussion

This chapter has presented an overview of the progress achieved in the automatic transcription of CHIL meetings over the past three and half years. Over this period, ASR technology developed by IBM, LIMSI, and UKA-ISL was evaluated four times, twice in internal consortium benchmarks, and twice in the Rich Transcription international campaigns, overseen by NIST. The latter also attracted significant interest by external parties. In these evaluations, the CHIL partner sites demonstrated significant improvements in ASR system accuracy over time and competitive performance compared to non-CHIL site systems.

Nevertheless, ASR word error rates remain high, particularly in the far-field task for the CHIL scenarios. The continued accuracy improvements indicate that further improvements are to be expected, driven by better acoustic and language modeling as well as further data availability. Future research will also address the modeling of disfluencies in spontaneous speech and pronunciation modeling for nonnative speech. In particular, better addressing the channel combination problem and concentration on advanced noise-removal techniques should benefit system performance.

Future challenges will focus on blind dereverberation, which is still a very difficult task, and the development of systems able to separate target speech from interference speech [26], the so-called cocktail party effect. This describes the ability of humans to listen to a single talker among a mixture of conversations.

It is also expected that in the future, visual speech information could be robustly extracted from the participants in CHIL interactive seminars and lectures, employing appropriately managed, active pan-tilt-zoom cameras in the CHIL smart rooms. Such information can then be fused with acoustic speech input to better address the CHIL ASR problem and its components, including speech activity detection, speaker diarization, and source separation. CHIL partners IBM and UKA-ISL have already expended a significant effort in this area and have focused on two problems of particular interest in the CHIL scenarios: the issue of visual feature extraction from

nonfrontal views [33, 28] and the problem of robust audiovisual speech integration [37, 27].

References

1. AMI – Augmented Multiparty Interaction, `http://www.amiproject.org`.
2. T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul. A compact model for speaker adaptation training. In *International Conference on Spoken Language Processing (ICSLP)*, pages 1137–1140, Philadelphia, PA, 1996.
3. A. Andreou, T. Kamm, and J. Cohen. Experiments in vocal tract normalisation. In *Proceedings of the CAIP Workshop: Frontiers in Speech Recognition II*, 1994.
4. X. Anguera, C. Wooters, and J. Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(7):2011–2022, 2007.
5. C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain. Multi-stage speaker diarization of Broadcast News. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1505–1512, 2006.
6. R. E. Bellman. *Adaptive Control Processes*. Princeton University Press, Princeton, NJ, 1961.
7. S. Burger, V. McLaren, and H. Yu. The ISL meeting corpus: The impact of meeting type on speech style. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO, 2002.
8. S. M. Chu, E. Marcheret, and G. Potamianos. Automatic speech recognition and speech activity detection in the CHIL seminar room. In *Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, pages 332–343, Edinburgh, United Kingdom, 2005.
9. S. Davis and P. Mermelstein. Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.
10. A. P. Dempster, M. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B (Methodological)*, 39:1–38, 1977.
11. F. Faubel and M. Wölfel. Coupling particle filters with automatic speech recognition for speech feature enhancement. In *Proceedings of Interspeech*, 2006.
12. J. G. Fiscus. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER). In *Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 347–352, Santa Barbara, CA, 1997.
13. J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo. The Rich Transcription 2006 Spring meeting recognition evaluation. In S. Renals, S. Bengio, and J. G. Fiscus, editors, *Machine Learning for Multimodal Interaction*, LNCS 4299, pages 309–322. 2006.
14. J. S. Garofolo, C. D. Laprun, M. Michel, V. M. Stanford, and E. Tabassi. The NIST meeting room pilot corpus. In *Proceedings of the Language Resources Evaluation Conference*, Lisbon, Portugal, May 2004.
15. J.-L. Gauvain and C. Lee. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Transactions on Speech and Audio Processing*, 2(2):291–298, Apr. 1994.
16. R. Gopinath. Maximum likelihood modeling with Gaussian distributions for classification. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 661–664, Seattle, WA, 1998.

17. R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 13–16, 1992.
18. H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustic Society of America*, 87(4):1738–1752, 1990.
19. R. Hincks. *Computer Support for Learners of Spoken English*. PhD thesis, KTH, Stockholm, Sweden, 2005.
20. J. Huang, E. Marcheret, and K. Visweswariah. Improving speaker diarization for CHIL lecture meetings. In *Proceedings of Interspeech*, pages 1865–1868, Antwerp, Belgium, 2007.
21. J. Huang, E. Marcheret, K. Visweswariah, V. Libal, and G. Potamianos. The IBM Rich Transcription 2007 speech-to-text systems for lecture meetings. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 429–441, Baltimore, MD, May 8-11 2007.
22. J. Huang, E. Marcheret, K. Visweswariah, and G. Potamianos. The IBM RT07 evaluation systems for speaker diarization on lecture meetings. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 497–508, Baltimore, MD, May 8-11 2007.
23. J. Huang, M. Westphal, S. Chen, O. Siohan, D. Povey, V. Libal, A. Soneiro, H. Schulz, T. Ross, and G. Potamianos. The IBM Rich Transcription Spring 2006 speech-to-text system for lecture meetings. In *Machine Learning for Multimodal Interaction*, pages 432–443. LNCS 4299, 2006.
24. 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 *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003.
25. N. S. Kim. Feature domain compensation of nonstationary noise for robust speech recognition. *Speech Communication*, 37:231–248, 2002.
26. K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wölfel. Adaptive beamforming with a minimum mutual information criterion. *IEEE Transactions on Audio, Speech, and Language Processing*, 15:2527–2541, 2007.
27. K. Kumatani, S. Nakamura, and R. Stiefelhagen. Asynchronous event modeling algorithm for bimodal speech recognition. *Speech Communication*, 2008. (submitted to).
28. K. Kumatani and R. Stiefelhagen. State-synchronous modeling on phone boundary for audio visual speech recognition and application to multi-view face images. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 417–420, Honolulu, HI, 2007.
29. L. Lamel, J. L. Gauvain, and G. Adda. Lightly supervised and unsupervised acoustic model training. *Computer, Speech and Language*, 16(1):115–229, 2002.
30. L. F. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillmann. The translanguage English database (TED). In *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, 1994.
31. The LDC Corpus Catalog. `http://www.ldc.upenn.edu/Catalog`.
32. C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2):171–185, 1995.
33. P. Lucey, G. Potamianos, and S. Sridharan. A unified approach to multi-pose audio-visual ASR. In *Interspeech*, Antwerp, Belgium, 2007.

34. J. Luque, X. Anguera, A. Temko, and J. Hernando. Speaker diarization for conference room: The UPC RT07s evaluation system. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 543–554, Baltimore, MD, May 8-11 2007.
35. D. Macho, C. Nadeu, and A. Temko. Robust speech activity detection in interactive smart-room environments. In *Machine Learning for Multimodal Interaction*, LNCS 4299, pages 236–247. 2006.
36. L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: Word error minimization and other applications of confusion networks. *Computer, Speech and Lanuage*, 14(4):373–400, 2000.
37. E. Marcheret, V. Libal, and G. Potamianos. Dynamic stream weight modeling for audio-visual speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 945–948, Honolulu, HI, 2007.
38. E. Marcheret, G. Potamianos, K. Visweswariah, and J. Huang. The IBM RT06s evaluation system for speech activity detection in CHIL seminars. In *Machine Learning for Multimodal Interaction*, LNCS 4299, pages 323–335. 2006.
39. J. M. Pardo, X. Anguera, and C. Wooters. Speaker diarization for multi-microphone meetings using only between-channel differences. In *Machine Learning for Multimodal Interaction*, pages 257–264. LNCS 4299, 2006.
40. D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig. fMPE: Discriminatively trained features for speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 961–964, Philadelphia, PA, 2005.
41. D. Povey and P. Woodland. Improved discriminative training techniques for large vocabulary continuous speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, 2001.
42. D. Povey and P. C. Woodland. Minimum phone error and I-smoothing for improved discriminative training. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 105–108, Orlando, FL, 2002.
43. E. Rentzeperis, A. Stergiou, C. Boukis, A. Pnevmatikakis, and L. C. Polymenakos. The 2006 Athens Information Technology speech activity detection and speaker diarization systems. In *Machine Learning for Multimodal Interaction*, pages 385–395. LNCS 4299, 2006.
44. C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2nd edition, 2004.
45. Rich Transcription 2007 Meeting Recognition Evaluation. `http://www.nist.gov/speech/tests/rt/rt2007`.
46. G. Saon, G. Zweig, and M. Padmanabhan. Linear feature space projections for speaker adaptation. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 325–328, Salt Lake City, UT, 2001.
47. H. Schwenk. Efficient training of large neural networks for language modeling. In *Proceedings of the International Joint Conference on Neural Networks*, pages 3059–3062, 2004.
48. R. Singh and B. Raj. Tracking noise via dynamical systems with a continuum of states. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, 2003.
49. O. Siohan, B. Ramabhadran, and B. Kingsbury. Constructing ensembles of ASR systems using randomized decision trees. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 197–200, Philadelphia, PA, 2005.

50. *The Translanguage English Database (TED) Transcripts (LDC catalog number LDC2002T03, ISBN 1-58563-202-3)*.
51. S. E. Tranter and D. A. Reynolds. An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(5):1557–1565, 2004.
52. M. Wölfel. Warped-twice minimum variance distortionless response spectral estimation. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2006.
53. M. Wölfel. Channel selection by class separability measures for automatic transcriptions on distant microphones. In *Proceedings of Interspeech*, 2007.
54. M. Wölfel. A joint particle filter and multi-step linear prediction framework to provide enhanced speech features prior to automatic recognition. In *Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, Trento, Italy, 2008.
55. M. Wölfel and J. McDonough. Combining multi-source far distance speech recognition strategies: Beamforming, blind channel and confusion network combination. In *Proceedings of Interspeech*, 2005.
56. M. Wölfel and J. W. McDonough. Minimum variance distortionless response spectral estimation, review and refinements. *IEEE Signal Processing Magazine*, 22(5):117–126, 2005.
57. M. Wölfel, K. Nickel, and J. McDonough. Microphone array driven speech recognition: influence of localization on the word error rate. *Proceedings of the Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI)*, 2005.
58. M. Wölfel, S. Stüker, and F. Kraft. The ISL RT-07 speech-to-text system. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 464–474, Baltimore, MD, May 8-11 2007.
59. X. Zhu, C. Barras, L. Lamel, and J. L. Gauvain. Speaker diarization: from Broadcast News to lectures. In *Machine Learning for Multimodal Interaction*, pages 396–406. LNCS 4299, 2006.
60. X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain. Multi-stage speaker diarization for conference and lecture meetings. In *Multimodal Technologies for Perception of Humans, Proceedings of the International Evaluation Workshops CLEAR 2007 and RT 2007*, LNCS 4625, pages 533–542, Baltimore, MD, May 8-11 2007.