Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-Text Evaluation System

Andreas Stolcke[1,2], Xavier Anguera[1,3], Kofi Boakye[1], Özgür Çetin[1], František Grézl[1,4], Adam Janin[1], Arindam Mandal[5], Barbara Peskin[1], Chuck Wooters[1], and Jing Zheng[2]

[1] International Computer Science Institute, Berkeley, CA, USA
[2] SRI International, Menlo Park, CA, USA
[3] Technical University of Catalonia, Barcelona, Spain
[4] Brno University of Technology, Czech Republic
[5] University of Washington, Seattle, WA, USA
stolcke@icsi.berkeley.edu

Abstract. We describe the development of our speech recognition system for the National Institute of Standards and Technology (NIST) Spring 2005 Meeting Rich Transcription (RT-05S) evaluation, highlighting improvements made since last year [1]. The system is based on the SRI-ICSI-UW RT-04F conversational telephone speech (CTS) recognition system, with meeting-adapted models and various audio preprocessing steps. This year's system features better delay-sum processing of distant microphone channels and energy-based crosstalk suppression for close-talking microphones. Acoustic modeling is improved by virtue of various enhancements to the background (CTS) models, including added training data, decision-tree based state tying, and the inclusion of discriminatively trained phone posterior features estimated by multilayer perceptrons. In particular, we make use of adaptation of both acoustic models and MLP features to the meeting domain. For distant microphone recognition we obtained considerable gains by combining and cross-adapting narrow-band (telephone) acoustic models with broadband (broadcast news) models. Language models (LMs) were improved with the inclusion of new meeting and web data. In spite of a lack of training data, we created effective LMs for the CHIL lecture domain. Results are reported on RT-04S and RT-05S meeting data. Measured on RT-04S conference data, we achieved an overall improvement of 17% relative in both MDM and IHM conditions compared to last year's evaluation system. Results on lecture data are comparable to the best reported results for that task.

1 Introduction

Meeting recognition continues to be a challenging task for speech technology for several reasons. Unrestricted speech, recognition from distant microphones, varying noise conditions, and multiple and overlapping speakers pose problems not found in other widely used benchmark tests. Furthermore, meetings pose the interesting problem of designing *portable* recognition systems, in two regards. First, because of the relative novelty of the task, and limited size of in-domain training corpora, it is advantageous to try to leverage methods and data that have been developed for other genres of speech,

S. Renals and S. Bengio (Eds.): MLMI 2005, LNCS 3869, pp. 463–475, 2006.

such as conversational telephone speech (CTS) and broadcast news (BN), for which one can draw on a longer development history and an order of magnitude more data. The second motivation for portability is that the meeting domain itself is varied, with different collection sites, acoustic conditions, and conversational styles and topics.

As for last year's meeting evaluation (RT-04S), our development strategy for RT-05S was to start with an existing CTS system[1] and adapt it to the meeting domain. This allowed us to leverage research between the corresponding CTS evaluations, from the Fall of 2003 (RT-03F) to the Fall of 2004 (RT-04F), and was crucial to developing a meeting system in the short period available. Acoustic models were adapted to the available conference room data (some of it new for this year), and language models were rebuilt for the conference and lecture room domains (no special acoustic models were created for the lecture domain). A new aspect in our acoustic modeling this year was the use of discriminatively trained Tandem/HATS features, and the fact that features were adapted to the new task, in addition to the more standard model adaptation. The acoustic preprocessing for meetings was also improved significantly, for both distant and individual microphone conditions.

The evaluation task and data are described in Section 2. Section 3 gives the system description, focusing on new developments relative to the 2004 system [1]. Results and discussion appear in Section 4, followed by conclusions and future work in Section 5.

2 Task and Data

2.1 Test Data

Evaluation data. The RT-05S conference room evaluation data (`eval05`) consisted of two meetings from each of the recording sites AMI (Augmented Multi-party Interaction project), CMU (Carnegie Mellon University Interactive Systems Laboratory), ICSI, NIST, and VT (Virginia Tech). Systems were required to recognize a specific 12-minute segment from each meeting; however, data from the entire meeting was allowed for processing.[2] Separate evaluations were conducted in three conditions:

MDM Multiple distant microphones (primary)
IHM Individual headset microphones (required contrast)
SDM Single distant microphone (optional)

The lecture room data consisted of 120 minutes of seminars recorded by the Computers In the Human Interaction Loop (CHIL) consortium. In addition to the above conditions, lecture data provided the following recording conditions:

MSLA Multiple source-localization arrays (optional)
MM3A Multiple Mark III microphone arrays (optional). The MM3A condition has not yet been delivered for the evaluation set, and could be evaluated only on development data, using a single array.

[1] As explained later, we also made use of acoustic models developed for BN.
[2] We did not find significant gains from adapting on entire meetings, and, except in the acoustic preprocessing, used only the designated meeting excerpts.