Information discovery using distant speech recognition

Florian Metze

Deutsche Telekom Laboratories, Berlin, Deutschland, Email: florian.metze@telekom.de

Introduction

Continuous incremental improvements lead to the deployment of automatic speech recognition (ASR) applications such as dictation or speech dialog systems several years ago. The business case however is usually based on efficiency, instead of an added value for the user such as new services or an improved experience, for which high accuracy systems are needed.

Still, even moderately accurate speech recognition could be very useful in "walk in" situations and for "topic spotting" applications: here, ASR using distant microphones represents a non-intrusive technology which can be used as a barrier-free modality to initiate interaction of users with multi-modal devices, for example information kiosks. An ASR enabled information kiosk could detect conversations matching its domain of knowledge in the vicinity and pro actively invite users to explore its contents by non-disturbing methods, for example by flashing pictures of products or exhibits on the screen. Once the initial "discovery of information" is achieved by ASR, further interaction could happen via touch or keyboard for robustness and speed.

In this paper we therefore describe work focusing on distant speech recognition as an enabler for multi-modal interaction with devices, allowing easy discovery of devices, capabilities, and content by non-expert users. We will present the "SAT/ CAT" approach to training a recognizer using distant speech and present results of the resulting recognizer in terms of word error rate (WER) and topic detection error rate.

The "Meeting" Scenario and Data

The system described in this paper is based on ISL's RT-04S evaluation system [7, 6]. The wide-band training and testing "Meeting" data used in this work mainly consists of group meetings in a professional or research environment, where participants were usually seated around a meeting table. As the meetings occurred naturally, they contain spontaneous effects and sloppy speech, although the amount varies among the four collection sites CMU, ICSI, LDC, and NIST. Speech was recorded through head-mounted microphones as well as through a variable number of omni-directional table-top microphones placed on the table, see Table 1. Pointers to these corpora as well as descriptions of their properties are available on the RT-04S evaluation web-site [7], the data itself is available through LDC.

Development test data for RT-04S consisted of 10-minute excerpts of eight meetings, two per site, with manual segmentation. Three to ten people participated in each meeting, while the number of distant channels varied between one (CMU) and ten (some LDC meetings). No training data was available for LDC.

Table 1: All "Meeting" training data sets contain recordings of individual speakers with personal microphones in addition to the above number of distant microphone recordings. "BN" data was added from the BN'96 and BN'97 training corpora for robustness.

Corpus	Duration	Speakers	Dist. Mics
CMU	11h	93	-
ICSI	72h	455	4
NIST	13h	77	7
BN	180h	4236	-

To test the performance of the system for information access, where word error rate (WER) may not be the main criterion, we additionally generated a database containing wide-band distant microphone speech from several speakers (without discarding overlapping speech), which we transcribed and segmented into 70 topics, for which we had determined about 200 keywords on the CUIMPB corpus [5]. This database was recorded under similar conditions as the training data.

Our experiments use interpolated n-gram language models trained on BN, Meeting, and TED [4] corpora, which covers topics similar to the ones used in our experiments. Language model training is described in [6] for the experiments on the "Meeting" data and [5] for the topic detection experiments. The dictionary contained ca. 50k entries collected from various sources for the LVCSR task.

SAT/ CAT Acoustic Model Training

The recognizers used in these experiments were based on the Janus recognizer and the Ibis single-pass decoder [8]. We computed a 42-dimensional feature space based on MFCCs with Cepstral Mean Subtraction (CMS) and Cepstral Variance Normalization (CVN) applied on a per-utterance basis. Additionally, the mean of the LDA features was also normalized to zero mean. We use a ± 7 frames context window before applying separate LDA and global STC transforms [2]. No specific noise-filtering has been employed. Initial experiments were run with a 2k codebooks, 6k distribution, 100k diagonal Gaussians system using a phonetic context of ± 2 for the clustering tree.

A system trained on close-talking data only reaches a WER of 67.2% on the RT-04S development test set. Two extra iterations of Viterbi training on all four high-quality channels of the ICSI distant microphone data

resulted in a WER of 62.5%, an improvement of 4.7% absolute. Employing feature space adaptation (FSA, akaċonstrained MLLR) [1] and VTLN during testing only reaches 58.6%. As a next step, we performed a combination of channel-adaptive (CAT) and speaker-adaptive (SAT) training also using constrained MLLR [3], by estimating a separate normalization matrix for *every speaker* and *every recording channel*. Input vectors $\mathbf{o}(\tau)$ are therefore transformed according to:

$$\hat{\mathbf{o}}^{(s+c)}(\tau) \leftarrow \mathbf{A}^{(s+c)}\mathbf{o}(\tau) + \mathbf{b}^{(s+c)} \quad (1)$$

where the transformation parameters $\mathbf{A}$ and $\mathbf{b}$ depend on speaker and channel ("$s + c$"). This resulted in a WER of 54.5%, which is an 8.0% absolute (13% relative) gain over the unadapted system. Performing SAT alone (using $\mathbf{A}^{(s)}$ and $\mathbf{b}^{(s)}$ transformation parameters) on the close-talking data did not decrease word error rate.

Re-training the context decision tree on the combined data sets, increasing the model complexity to 6k codebooks, 24k distributions, ∼300k Gaussians assigned by Merge-and-Split training while also re-training STC achieved an additional reduction in WER for the SAT/ CAT procedure and reduced the error rate by an extra 3.5% absolute.

Topic Spotting using Room Microphones

A real-time recognizer derived from the above systems listening on a single microphone placed in the middle of a meeting room table was now used to detect topics from the conversation between people discussing the scientific talks presented at the Barcelona "Forum of Cultures" [5]. These talks had manually been segmented into 70 topics, for which 200 keywords had been defined. These keywords were then being spotted by a normal n-gram based speech recognizer.

To improve the performance of the distant speech recognition system, data was collected over the Internet to generate a feasible language model (LM), as only the headlines of the talks and the presenters of the seminar were available. A 3-gram LM was calculated and interpolated with a background LM consisting of the BN corpus. Furthermore the LM was tuned for topic spotting by a higher weight to the keywords, so that they appear more often. Furthermore, words with a low confidence value, which are more likely to be wrong, were deleted from the hypos to not disturb the topic spotting process.

The feature space adaptation matrix needed for decoding was seeded with a small amount of supervised adaptation data and then updated continuously in an unsupervised manner. The initial system using a BN LM interpolated with data collected from the Internet had a topic detection error rate of 41% at a WER of 83%. The topic detection error rate is defined as the error rate of the observed topic sequence string when compared against the reference topic string, when every topic counts as one "word". Re-weighting the keywords for topic spotting and taking confidence scores into account reduced the topic spotting error rate to 38%.

Conclusion and Future Work

Our experiments show how normalization techniques such as speaker-adaptive training can be employed to improve performance of an ASR system on data from distant microphones. By normalizing for both speaker and channel characteristics during training, multi-channel distant microphone audio data can be used effectively for recognizer development. The gains reached were higher and complementary to gains reached using other techniques such as VTLN and FSA during decoding only. Initial experiments showed that microphones placed on the table or embedded in the table perform better than microphones suspended from the ceiling in this setting, while microphones put on window boards picking up indirect audio from one direction mainly also reached good performance.

We plan to extend this approach to more corpora and evaluate integration with single-channel de-noising and de-reverberation techniques as well as discriminative training techniques for acoustic models.

References

[1] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical report, Cambridge University, Cambridge; UK, May 1997. CUED/F-INFENG/TR 291.

[2] M. J. F. Gales. Semi-Tied Covariance Matrices for Hidden Markov Models. *IEEE Transactions on Speech and Audio Processing*, Vol. 2, May 1999.

[3] H. Jin, S. Matsoukas, R. Schwartz, and F. Kubala. Fast Robust Inverse Transform SAT and Multi-stage Adaptation. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA; USA, 1998.

[4] L. Lamel, F. Schiel, A. Fourcin, J. Mariani, and H. Tillmann. The translanguage english database (ted). In *Proc. ICSLP1994*, pages 1795 – 1798, Yokohama; Japan, 1994. ISCA.

[5] F. Metze et al. The 'FAME' Interactive Space. In *Proc. 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms (MLMI2005)*, Edinburgh; UK, July 2005. Springer.

[6] F. Metze, C. Fügen, Y. Pan, and A. Waibel. Automatically Transcribing Meetings Using Distant Microphones. In *Proc. ICASSP 2005*, Philadelphia, PA; USA, Mar. 2005. IEEE.

[7] NIST. Proceedings NIST ICASSP 2004 Meeting Recognition Workshop. http://www.nist.gov/speech/test_beds/mr_proj/icassp_program.html, May 2004.

[8] H. Soltau, F. Metze, C. Fügen, and A. Waibel. A One-pass Decoder based on Polymorphic Linguistic Context Assignment. In *Proc. ASRU 2001*, Madonna di Campiglio, Italy, Dec. 2001. IEEE.