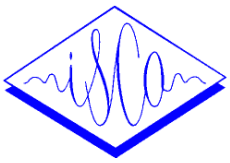

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13–16, 2003

Advances in ISL's Lecture and Meeting Trackers

Alex Waibel, Ivica Rogina
Interactive Systems Labs
Universitaet Karlsruhe
E-mail: {waibel|rogina}@ira.uka.de

Abstract

Most speech applications to-date have attempted to provide a more natural interface for human-computer interaction or human-computer data-input. Only recently, a whole new class of applications is coming to the fore: computer enhanced human-human interaction. In these applications the computer is no longer addressed directly, but must observe, process and understand the interaction between humans in a room. In this paper we discuss two such applications: a meeting browser, that observes and tracks meetings for later review and summarization, and a lecture tracker, that provides not only summarization, but also implicit services during a presentation, such as control of AV equipment and selection of the most suitable slides. Processing human-human conversational speech under unpredictable recording conditions and vocabularies presents new challenges for speech and language processing. We describe techniques designed to overcome these difficulties and report speech recognition as well as overall system performance results.

1. Introduction

The standard way of presenting a lecture or a talk is by using powerful tools that help in designing a multimedia presentation which itself is presented in a room equipped with many supporting audio and video devices.

We expect a reduction of the workload on the speakers in lectures and participants in meetings by letting the intelligent meeting room track the presentation and discussion and offer context aware services. One precondition for being context aware is the capability to align the speech of a lecturer with the presentation documents.

Standard speech recognition systems are trained on huge amounts of data from many different domains, often about news. Such recognizers are not very well suited for scientific lectures and disussions. Therefore we propose to build an adaptable baseline that allows easy and fast adaptation to a lecture by looking at the presentation documents (usually slides, but possibly also notes, or video/audio files). Presentation slides do not contain enough text to use it for adapting the language model or vocabulary. Instead, in our system, we extract important words from the slides and retrieve greater amounts of text from related internet pages. Again, those pages are not used to extract additional language model training data. They are only used to decide into which class of the class based baseline language model every important word should be added.

2. The FAME Room

Rather than an intelligent living room, the FAME room is more of an intelligent meeting room. In addition to the tasks performed by our previously presented meeting tracker systems [1][2], the FAME project foresees activities of the room during a meeting or lecture, namely to act as an information butler in the background. Meetings and lectures should be held as usual, only in cases where the participants explicitly or implicitly require additional information, the information butler becomes active.

The most important differences for the system between tracking a meeting and tracking a lecture are the quality of the speech acoustics, the availability of prepared documents (the presentation documents), and the speaking style. We can expect much more planned speech in a lecture than in a meeting.

The speech acoustics are easier for the recognizer for two reasons: first, the speech is more planned than in a discussion, and secondly, it's

more reasonable to expect a lecture speaker to wear a head-mounted microphone or at least a lapel microphone. Although we have experienced rather significant differences in the audio quality of speech recorded with lapel microphones due to effects like acoustic shadow made by the speaker's chin or like rubbing the microphone with parts of the body or the clothing, the recognition accuracy is still better than with distant speaking table-mounted microphones. Although experiments [5] have shown, that using microphone arrays can improve the recognition of distantly recorded speech, in the FAME room, we will first focus on using microphone arrays only for localization of sound sources.

The two major tasks in assisting the lecturer consist of controlling the audio and video devices in the room and controlling the presentation. The former means turning on and off devices, dimming the lights, setting volumes of speakers etc., the latter means automatically selecting and displaying slides and optionally audio or video documents. Both services can be performed implicitly by the system "guessing" what is currently needed, or by having the speaker give explicit commands.

3. The Lecture Assistant

The tasks of the lecture assistant addressed in this paper are

- analysis of the presented documents
- related information retrieval from the internet
- adaptation of the vocabulary and language model and pronunciation lexicon
- speech recognition and lecture tracking

We will now describe these tasks in greater detail (see Fig. 1).

3.1 Analysis of the Presentation

The analysis of the presentation documents has two goals. One is to extract the important content words, the other is to retrieve an ASCII representation of the document which can be used to correlate it with the recognizer output such that the system always knows which part of the presentation the speaker is talking about.

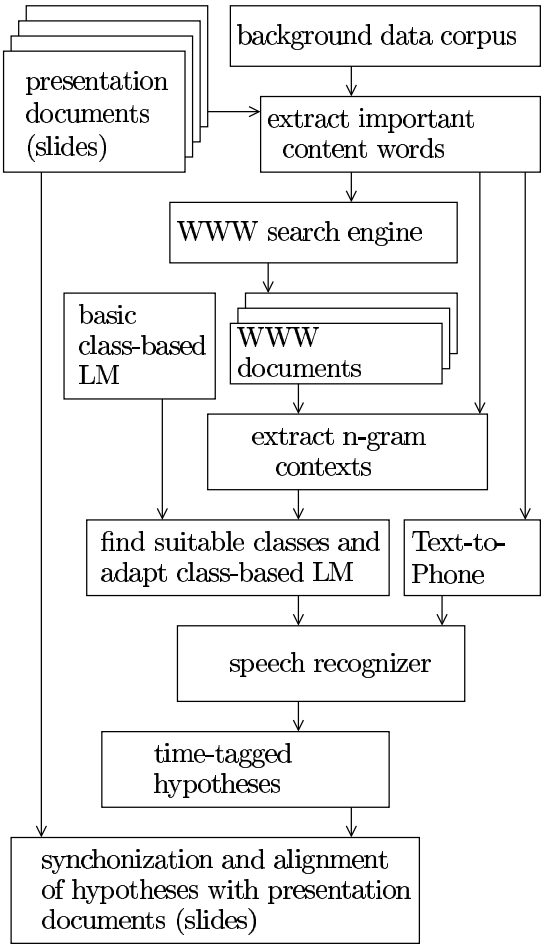

Figure 1: Overview of the lecture assistant

The extracted words are compared with the recognizer's background vocabulary. A $tfidf$-value is computed for every content word. All out-of-vocabulary words (OOV) and the words with higher-than-average $tfidf$-values are considered to be important.

3.2 Adapting the Language Model

The primary goal of the language model design was to allow easy adding of new words. Therefore a class-based design was chosen. The basis was a trigram model on a 40k vocabulary trained with the HUB-4 (broadcast news) training corpus. This language model did not contain any classes. To define classes that would be suitable to accept new OOV words, 20k more words from the HUB-4 vocabulary were taken. All sentences from the

corpus which contained one of the additional 20k words were fed into a Kneser-Ney bigram clustering algorithm. The result of the clustering process was a set of 72 classes – representing the OOV-words from the point of view of the base 40k-language model. Every important word found in the presentation documents was then put into the class which best fit the extracted contexts from web pages containing the important word. A detailed description of the adaptation process can be found in [6].

3.3 Recognition and Tracking

A tracking system for a presentation or a lecture uses the same basic technology that most systems would use which have to monitor people in action and relate their actions (esp. their spoken utterances) with corresponding documents or parts of documents. A good lecture tracker can be used as a basis for a good meeting tracker in which several people talk about and work with a set of possibly shared documents.

The tasks performed by the lecture tracker during the lecture consist of recognizing the lecturer's speech, possibly switching slides and displaying documents (WWW, video, audio) when appropriate, and, in hopefully rare cases, interacting with the lecturer to resolve problems. While interaction is a problem that is part of the FAME project, we have not addressed it in this paper.

After the lecture, the recorded audio and the presented documents have to be aligned, indexed and stored, such that it will be possible to retrieve the documents and the audio recording of the speaker for later browsing.

3.4 Automatic Slide Switching

The system can switch slides on commands like "next slide" or "go to slide number seven", but also implicitely by tracking the lecturers speech. The decision to change the displayed slide is made by aligning the speech hypotheses with the texts on the slides and by using heuristics about trigger words. Trigger words are important words that occur on slide n but not on slide $n-1$.

3.5 Experiments

We have conducted experiments on three self-recorded lectures. The speech recognizer used in our experiments is based on the JANUS speech recognition toolkit [3]. It was trained on the HUB-4 broadcast news training corpus and was used in other systems like [1][4].

Lecture	3	4	5
baseline	33.5%	43.7%	31.0%
system 1	28.1%	39.7%	29.8%
system 2	26.8%	37.8%	27.6%

Table 1: improvements by LM-adaptation

Lecture	3	4	5
baseline	34.7%	31.8%	32.1%
adapted	28.3%	31.0%	27.8%
with trigger words	8.0%	8.8%	4.7%

Table 2: Improvements in tracking error rates

We trained two systems, system 1 used only OOV words for adapting the language model and the vocabulary, system 2 also used important content words and treated them as if they were OOV. After adapting the language model, the word error rates were as shown in table 1.

The OOV-rate of the test set lectures is approximately 5%. So the adding of OOV words found in the presented documents to the recognizer's vocabulary is not sufficient to explain the significant gain in word accuracy. The other major contribution to the improvement comes from increasing the unigram probabilities for the important content words.

In the synchronization experiments, every hypothesis word was automatically assigned a slide of the presentation and compared to the actual slide that was displayed at the corresponding time. To find a temporal alignment of the slides, we use a standard dynamic time warping algorithm. The tracking error rate improved as shown in table 2.

Of course, the tracking accuracy highly depends on the amount of information found on the slides. Presentations containing only headlines and images are much harder to track than presentations with lots of text.

4. Meeting Tracking

4.1 Acoustic Models and Training Procedure

Since there is no sufficient training data for the meeting domain currently available the acoustic

models were trained on a large telephony speech corpus of around 300h of data. The Switchboard corpus offers a variety of speaking styles and is therefore well suited to capture colloquial speech as it occurs in meetings.

The front-end features are based on cepstral coefficients derived from a Mel filterbank analysis. Cepstral mean and variance compensation are used to reduce channel variations. In the final preprocessing step, 11 frames are concatenated to form a $11 * 13 = 143$ feature vector, followed by a dimension reduction to 42 by sorting the eigenvectors of the LDA matrix.

An initial set of $2.4m$ polyphones, induced by a phonetic context of +/- 3 and word boundaries, is transformed into a set of $50k$ context dependent HMM states using an divisive clustering procedure based on an entropy criterion. In order to optimize the parameter sharing we use a 2-stage decision tree. The first tree is used to create a set of $10k$ codebooks, while the second tree defines a set of $50k$ mixture weights on top of the codebooks.

The means and covariances were trained using an incremental growing of gaussians algorithm. Starting with one component per state, the training procedure split and merges gaussians iteratively. This procedure generates approximately $280k$ gaussians. In the next step, semi-tied full covariances are created to optimize the feature space according to the mixture densities. We apply this transform on the feature space, e.g. the resulting transformation is combined with the LDA matrix.

Speaking adaptive training is also applied on the feature level. Given the feature space obtained by the LDA/STC matrices, we apply feature space adaptation for each conversation and train the models on that normalized features. In summary, the full training procedure for the context dependent models consists of 4 steps.

1. estimate means/covariances via merge&split
2. train semi-tied full covariances
3. estimate mixture weights
4. speaker adaptive training on the feature space

4.2 Experiments

With the described approach, we were able to improve the word accuracy of our meeting tracking system significantly. The baseline word error rate of 38.7% as reported in [7] was reduced to 36.4%.

5. Conclusion and Further Plans

We have shown, that it is possible to improve the speech recognizer's word accuracy significantly if we can used data from documents that a speaker plans to present during a lecture. We have defined and evaluated the tracking accuracy and have shown that this too can profit from prior exploitation of the presentation documents.

In the future, we plan to allow the speaker to refer to specific slides either by naming them or their number or by addressing their contents.

Eventually the meeting tracker and the lecture tracker will be blend into a single system that will support both lectures and meetings and will allow later browsing with the same user interface.

Acknowledgements

Part of this work was carried out within the FAME project and has been funded by the European Union as IST project No. IST-2000-28323.

6. References

[1] Alex Waibel, Michael Bett, Michael Finke, Rainer Stiefelhagen: "*Meeting Browser: Tracking and Summarizing Meetings*", Proceedings of the Human Technology Conference, San Diego 2001

[2] Alex Waibel, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, Klaus Zechner: "*Advances in Automatic Meeting Record Creation and Access*", ICASSP 2001, Salt Lake City

[3] Ivica Rogina and Alex Waibel: "*The JANUS Recognizer*", ARPA Workshop on Spoken Language Technology, 1995

[4] Thomas Schaaf: "*Detection of OOV Words Using Generalized Word Models and a Semantic Class Language Model*", Proceedings of the Eurospeech 2001, Aalborg, September 2001

[5] Michael L. Seltzer, Bhiksha Raj, and Richard M. Stern: "*Speech Recognizer-Based Microphone Array Processing for Robust Hands-Free Speech Recognition* ",Proceedings of the ICASSP 2002

[6] Ivica Rogina and Thomas Schaaf: "*Lecture and Presentation Tracking in an Intelligent Meeting Room* ",Proceedings of the ICMI 2001

[7] Alex Waibel, Hua Yue, Hagen Soltau, Tanja Schultz, Thomas Schaaf, Yue Pan, Florian Metze, and Michael Bett: "*Advances in Meeting Recognition*",Proceedings of the HLT 2001