IDIAP RESEARCH REPORT

AMIDA/KLEWEL MINI-PROJECT

Petr Motlicek Philip N. Garner Maël Guillemot
Vincent Bozzo

Idiap-RR-03-2010

JANUARY 2010

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

AMIDA/Klewel Mini-Project

Petr Motlicek[*]
Philip N. Garner[*]
Vincent Bozzo[+]
Mael Guillemot[+]

[*]**Idiap Research Institute, Martigny, Switzerland**
[+]**Klewel, Martigny, Switzerland**

January 12, 2010

Contents

Abstract

The goal of the AMIDA mini-project is to transfer some of the technologies developed within the AMIDA project to be used by a Klewel retrieval system. More specifically, the main focus is to develop a speech-to-text application based on the AMIDA Automatic Speech Recognition (ASR) system which could be potentially implemented by Klewel in their conference webcasting system.

First, this document describes experimental setup and results achieved in the project devoted to the automatic processing of real lecture recordings provided by Klewel. Then, a demonstrator — an application created for demonstrating Automatic Speech Recognition (ASR) results — is described.

1 Klewel

Klewel is a spin-off of the Idiap Research Institute created in November 2007. Klewel provides leading edge solutions for capturing, indexing and distributing online the full content of conferences. The main objective of Klewel is to develop and maintain solutions which can ensure the continued existence and visibility of various types of events (congress, conferences, symposiums, workshops, meetings). Klewel developed a multimedia capture station that can be plugged into any existing meeting room in order to record events. The content of the presentation talks (including sound, video, projected slides and animations) are recorded into digital files. All these media files are then processed and distributed on the web (and on any media, intra or extra net, DVD, etc.). Klewel provides its customer a way to navigate through the content of a particular event thanks to a web-based interface. Such a tool allows to access to the full event at anytime from any place. Klewel has included a search engine based on an optical character recognition (OCR). Each captured slide is scanned and then processed in order to localize and extract the text from the images. Klewel's Flash based conference video player allows to browse online and search through the content of the events. By entering a key-word, the user can retrieve specific slides which contain the given key-word. All audio, video and slides are fully synchronized. Therefore, users are able to play back the talk from the exact time of interest by clicking on a retrieved slide. The objective of this AMIDA mini-project is to evaluate and test the use of speech-to-text indexing for lecture retrieval purpose.

2 LVCSR and STD processing on Klewel lecture recordings

2.1 Introduction

The goal of this Section is to describe experiments and summarize the achieved results performed on lecture recordings provided by Klewel (http://www.klewel.com). The data consists of several sessions of lectures recorded by Klewel on various international conferences. These lectures are supplemented with corresponding video and slide presentations. From slide presentations of some sessions we extracted a text later used in spoken term detection experiments.

Achieved results provide important insights for potential implementation of a system for automatic speech recognition and spoken term detection on Klewel lecture recordings.

2.2 Audio recordings

In total, four sessions (lectures/audio recordings) were processed. In each session, a speaker gives a lecture in English at a conference (workshop), which was automatically recorded by Klewel. The audio data were recorded using close-talk microphones and stored in 16 kHz, 1 channel, WAV format. For speaker adaptation techniques, we assumed that each lecture was given by one speaker (no speaker change over a session). All the speakers were native English speakers (US, UK, Canada). For recordings, Seinnheiser lapel microphone (or equivalent) were used (except the last one recorded with a standing microphone).

Although the quality of processed audio recordings was not objectively measured, our informal examination indicates a large variation over different lectures. The main difficulties were due to the saturation (dynamic range) of the signal amplitudes.

2.2.1 Description

Here, more detailed description of audio recordings obtained from various lectures is given:

- session 0001: lecture - ACM China, 2008, closing keynote

  - Speaker: Bill Buxton (Toronto, Canada)
  - Talk title: Being Human in a digital world
  - Related link: http://www.klewel.com/acmdemo/
  - Recorded in: Florence, Italy
  - Date: April 10, 2008
  - Duration: 49 minutes 56 seconds
  - Microphone: lapel microphone
  - Note: Applause at the beginning, large auditorium
  - Length: 1018 sec. (~ 17 minutes) of audio processed

- session 0003: General assembly
  - Talk title: Company keynote
  - Speaker: Male, American
  - Recorded in: Montreux, Switzerland
  - Duration: 15 minutes 32 seconds
  - Microphone: 2 formal conference desk microphones
  - Note: large international congress auditorium
  - Transcribed and processed length: 932 sec. (~ 16 minutes)

- session 0004: lecture - SSPnet, talk 1
  - Talk title: Co-constructing Social Signals
  - Speaker: Jeffrey Cohn (University of Pittsburgh/Carnegie Mellon University, USA)
  - Related link: http://www.sspnet.eu/recordings/ieeessp09/talk1.html
  - Recorded in: Amsterdam, NL
  - Date: September 13, 2009
  - Duration: 51 minutes 07 seconds
  - Microphone: lapel microphone
  - Note: small room equivalent to a classroom environment

  - Transcribed and processed length: 1127 sec. (~ 20 minutes)

- session 0005: lecture - SSPnet, talk 7
  - Talk title: Tracking the Second Channel of Information in Speech
  - Speaker: Nick Campbell (British, Trinity College Dublin, Ireland)
  - Related link: http://www.sspnet.eu/recordings/ieeessp09/talk7.html
  - Recorded in Amsterdam, NL
  - Date: September 13, 2009
  - Duration: 47 minutes 46 seconds
  - Microphone: lapel microphone
  - Note: saturated signal during the 1st minute, classroom environment
  - Transcribed and processed length: 912 sec. (~ 15 minutes)

In total, about 70 minutes of audio was used in our experiments. Such an amount should be sufficient to provide reliable statistics of achieved results.

2.2.2 Automatic segmentation

Audio recordings were automatically segmented using a Speech/Non-Speech (SNS) detector and then later decoded using various LVCSR systems. SNS detector - a state-of-the-art Multi-Layer Perceptron (MLP) based approach [2] uses 12 MF-PLP coefficients along with their first and second derivatives. To these, the following auxiliary features were added: normalized energy from all channels, signal kurtosis, mean cross-correlation and maximum normalized cross-correlation. The MLP was trained on 98 hours of training data recorded using Individual Headset Microphones (IHM) with a hyperbolic tangent hidden activation function and soft-max output activation function.

The process of segmentation was not evaluated. In order to increase the quality of such the automatic segmentation, the prior probabilities of speech/non-speech classes were manually selected based on preliminary overview of segmented data. These prior probabilities differ over audio recordings (sessions). Overall, the minimum allowed speech segment length was 4 sec. The maximum length was not restricted. A histogram of lengths of created segments from Klewel audio recordings is shown in Figure 1.

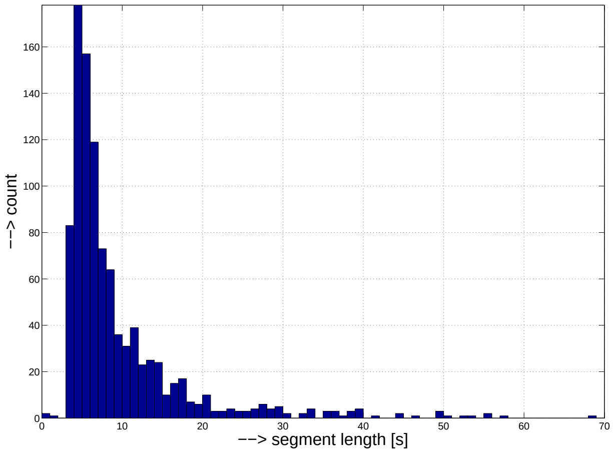


Figure 1: Histogram showing the segment lengths created from Klewel audio recordings.

2.2.3 Transcription of audio data

70 minutes (4 sessions) of Klewel audio recordings were manually annotated. We used a transcription tool available from sourceforge[1], which is distributed as a free software under GNU General Public License. Audio segments generated by SNS segmentation tool were annotated at the utterance level. Obtained annotations were cross-checked by a native English speaker.

2.2.4 OOVs

Annotations (stored as "trs" files created by the transcriber) were then transformed into Master Label File (MLF) to be used for scoring. At this

[1] http://trans.sourceforge.net

time, Out-Of-Vocabulary (OOV) words did not have to be carefully considered. More specifically, OOVs were tagged as “UNDEF” words. Total proportion of OOVs (with respect to all annotations) was about 1.1%.

2.2.5 Forced alignment of audio recordings with respect to manual annotations

Klewel audio recordings were force aligned in order to obtain a ground-truth (time boundaries) for Spoken Term Detection (STD) experiments. For achieving precise time boundaries of the annotated words, the dictionary was first enriched by OOV words (see next Section 2.2.6).

Then, forced alignment was performed using the SVite decoder (see STK[2]). Models used came from AMIDA 16k IHM LVCSR system (pass 3), described later.

2.2.6 GTP

OOV words were automatically added to the dictionary using a Grapheme-To-Phoneme (GTP) converter. GTP exploits linguistically based and automatically trained rules. A dictionary trained on English texts contains approximately 132k items.

2.3 LVCSR systems

The main goal of these experiments is to measure performance of AMIDA LVCSR when applied on real lecture recordings (recorded in real environments). In first experiments, we tested the state-of-the-art multi-pass LVCSR systems, which obviously do not run in real-time. More specifically, 3-pass systems were employed, based on various acoustic models trained on different audio data. Further, “close-to” real-time LVCSR systems were used for decoding.

No speaker enrollment was necessary; for the the multi-pass systems this is built into the first passes in the form of MLLR.

LVCSR systems used in the experiments:

[2] http://speech.fit.vutbr.cz/en/software/hmm-toolkit-stk

1. AMI CTS, 8kHz, 3-pass system: the system used in the experiments is based on the Conversational Telephone Speech (CTS) system, partially described in [3], derived from AMIDA LVCSR [5]. 250 hours of Switchboard data is used for training Hidden Markov Models (HMMs). The decoding is done in three passes, always with a simple bigram Katz backoff LM. In the first pass, PLP features (accompanied with delta coefficients) are used and processed by Heteroscedastic Linear Discriminant Analysis (HLDA) (to perform a robust data-driven dimension reduction). HMMs are trained using a Minimum Phone Error (MPE) procedure. In the second pass, Vocal Tract Length Normalization (VTLN) is employed on similar features from pass 1. In addition to HLDA, MPE and Speaker Adaptive Training (SAT) are applied. Finally, the third pass is similar to the second pass, except input PLP features are replaced by posterior-based features estimated using a Neural Network (NN) system. The NN processes 300 ms long temporal trajectories of mel filter-bank energies. The NN is represented by a Multi-Layer Perceptron (MLP) with 1 hidden layer (500 neurons).

   The LVCSR system reaches Word Error Rate (WER) of 2.9% on Wall Street Journal (WSJ1) Hub2 test set composed from November 92 (2.5 hours, with 5k dictionary and a trigram LM).

   For decoding Klewel audio recordings, a 50k dictionary was used together with a 3-gram Language Model (LM).

2. AMIDA IHM, 16kHz, 3-pass system: here, a previous CTS recognition system forms the starting point for this LVCSR system developed to process meeting data. More specifically, CTS data was used for bootstrapping [4]. Then, Individual Headset Microphone (IHM) recordings from the following databases were used in the training: ICSI meeting corpus, NIST and ISL meetings, and AMI meetings. Furthermore, discriminative training in 3-pass decoding is employed, similar to the previous AMI CTS system.

3. AMIDA MDM, 16kHz, 3-pass system: this ASR system is similar to the previous one. Instead of IHM data, Multiple Distant Microphone (MDM) recordings were used to train acoustic models.

4. STK based “real-time” LVCSR system: for comparison purposes, we

employed a relatively simple LVCSR 8kHz system. Although, this system currently runs in "off-line" mode, it can be simply turned into "on-line" mode to process the speech (with the decoding speed around real-time). As features, 13 PLPs with deltas and double-deltas were used. Off-line mean and variance normalization was then applied. The 30k dictionary was used with simple bi-gram language model. SVite - WFST decoder was used for decoding[3].

5. Juicer/tracter based "real-time" LVCSR system: this 16kHz LVCSR system uses Juicer for decoding [7]. "Tracter-basic" features (MFCCs) were derived with application of "on-line" cepstral mean normalization. The 50k RT07 dictionary was used with simple bi-gram (RT07) language model. The lack of cepstral variance normalisation (which cannot be easily implemented on-line) is likely to cause worse results.

2.4 LVCSR - experimental results

Experimental results of automatic recognition of Klewel audio recordings are represented in Word Error Rates (WERs). The recordings were decoded using all 5 hitherto described LVCSR systems. In case of AMIDA systems (1-3), the recognition results are provided also for the first pass, where the decoding takes place, as well (due to Maximum Likelihood based VTLN). Achieved results are given in Table 1 and Figure 2.

Overall, the best performace is provided by AMIDA IHM system (pass 3). AMIDA MDM and AMI CTS systems perform slightly worse for sessions 0001, 0003 and 0004. However, for session 0005, WERs are significantly worse. Achieved performances of STK based LVCSR are about 10% below multi-pass systems (for session 0005, WERs are much worse). WER performance obtained with "real-time" tracter-basic features is about 5% worse than those achieved with STK system. The main reason is almost certainly the employment of "on-line" normalization and lack of cepstral variance normalisation.

The performance of the systems is governed by the feature extraction and the complexity of the language models. In general, the decoders are interchangeable; we used the decoders that were most familiar to the system builders. The more recent RT09 system is able to achieve similar perfor-

[3] http://speech.fit.vutbr.cz/en/software/hmm-toolkit-stk

	Sessions			
	0001	0003	0004	0005
CTS 8k P1	37.1	32.3	35.3	48.0
CTS 8k P3	36.1	32.8	29.5	41.7
IHM 16k P1	38.8	34.8	31.8	32.5
IHM 16k P3	**33.5**	**29.4**	**25.9**	**26.9**
MDM 16k P1	43.6	35.4	37.5	36.5
MDM 16k P3	36.6	31.7	30.7	32.5
STK 8k	44.8	39.7	39.2	52.1
Tracter Basic 16k	50.6	46.4	46.6	47.8

Table 1: WER [%] performances achieved on Klewel lecture recordings for different LVCSR systems.

mance in fewer passes by using a high-order language model and WFST based decoder from the first pass.

2.5 Spoken Term Detection

Spoken Term Detection (STD) detects a word or phrase in unconstrained speech. STD is usually used in searching large archives of recorded speech such as in meeting data. Therefore, STD could be an interesting application for Klewel to search spoken terms in archived conference lectures. Compared to the traditional key-word spotting, an often used technique in speech processing, STD is a more complex application. In case of STD, a searched term is not known a-priori thus a representative model can not be built before the search.

State-of-the-art STD systems employ LVCSR to obtain an index represented by generated word lattices. Then, the index is searched in order to return the location of the determined term.

Experimental results presented in [10] showed that another modality (slide information) used to enhance speech transcription of meetings does not eventually improve ASR performance. In our experiments we decided to apply this modality in different way, to improve acoustic detection of spoken terms.

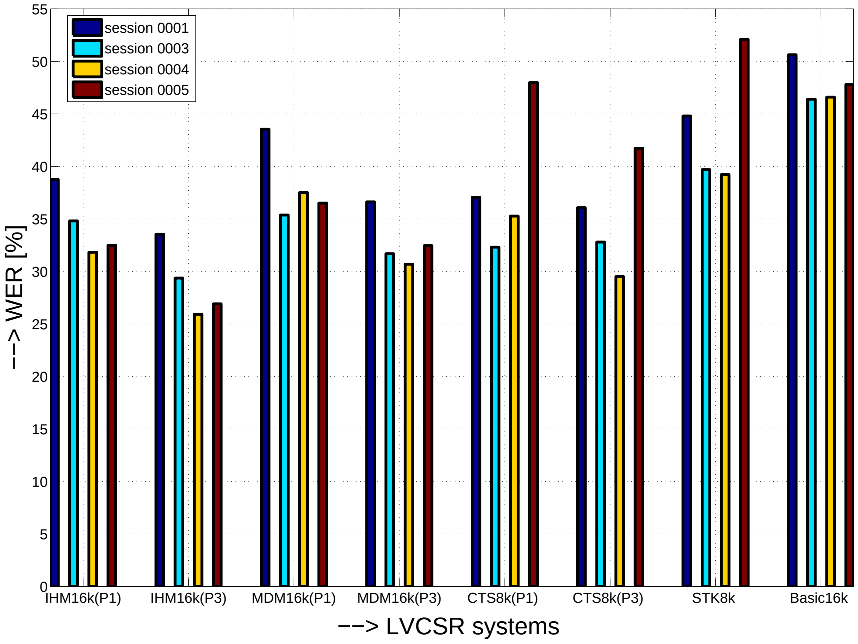


Figure 2: Graphical results of WER performaces for different LVCSR systems across different Klewel audio recordings.

2.5.1 System description

The AMIDA IHM, 16kHz, 3-pass system was used to generate word lattices. More specifically, word lattices were generated in the third pass using HTK[4] (HDecode) with bi-gram language model. Word posterior probabilities were estimated and the index was created using HTK, as well.

2.5.2 List of searched spoken terms

The list of spoken terms consists of 312 items. The terms were selected manually from available annotations (in a random fashion over all sessions) based only on a potential interest of Klewel end-users. The list of key-words was then transformed into a format following NIST STD evalu-

[4] http://htk.eng.cam.ac.uk

ations[5].

In addition, experiments were focused on testing potential detection improvement by employing a prior source of information available from accompanying text slides. For this reason, the list of searched terms contains (among 312 terms) 42 items which were chosen from the text slides (actually available only for the session 0004). The list of terms is at appendix B.

2.6 STD - experimental results

Usually, false alarm probabilities and miss probabilities in the STD task are evaluated. Then, performance is shown using a standard Detection Error Trade-off (DET) curves [6]. In addition, we also present Equal Error Rates (EERs), a one-number metric, mainly used to optimize the system performance. Figure 3 (red dashed curve) shows the performance of STD built on AMIDA IHM system (pass 3). Overall achieved EER is about 7.4%.

The set of keywords used is at appendix A.

2.6.1 Exploiting prior information from corresponding slides

In the following experiments, additional available modality (text slides in form of PowerPoint presentations) was exploited to define a prior probability for a given searched term. As mentioned in Section 2.5.2, 42 terms were randomly selected from text slides. Although slides for all lectures are available, only those from lecture 0004 were used in this study.

In general, word posterior probabilities were modified using a prior which represents a relevance of a key-word (spoken term) to the topic (given by corresponding text slides). More specifically, in our experiments the posterior probabilities P_{old} were updated by a multiplicative varying constant c:

$$\begin{aligned} P_{new} &= cP_{old}, \quad \text{if} \quad c <= 1/P_{old}, \\ P_{new} &= 1, \qquad\quad \text{otherwise}. \end{aligned}$$

Since no time allocation of individual slides and their precise assignment with audio segments is available (only the general lecture number

[5] http://www.itl.nist.gov/iad/mig/tests/std

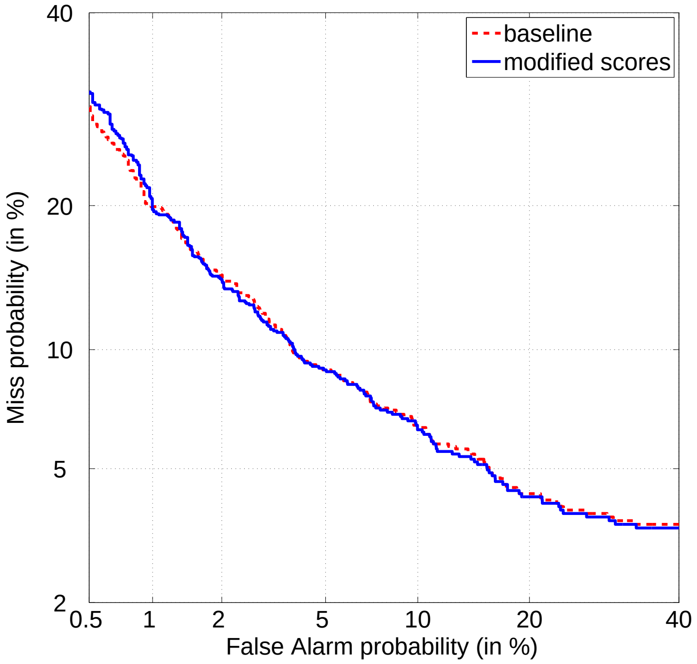


Figure 3: DET plot - Performance of STD system built on LVCSR lattices: a) Red (dashed) curve shows base-line performance, b) Blue curve represents performance when additional modality (text slides) was used as a prior.

assigning), no temporal information was employed. This means that c was not time dependent (i.e., applied only at given speech segments) and is unvarying to all 42 terms selected from text slides.

In the first experiment, posterior probabilities P_{old} related to the terms occurring in the list of 42 terms were updated (but only for session 0004). Then we estimated detection scores over all recordings (all 312 spoken terms). Corresponding DET (blue) curve is plotted in Figure 3. Figure 4 graphically shows EERs for different value of the multiplicative constant c. The performance is compared to the "baseline" system (red line, no prior information from text slides). In the following experiment, the detection scores were estimated only for lecture 0004. Corresponding EERs for varying c are also plotted in Figure 4. The baseline EER (black line) shows the detection performance computed only for recording 0004.

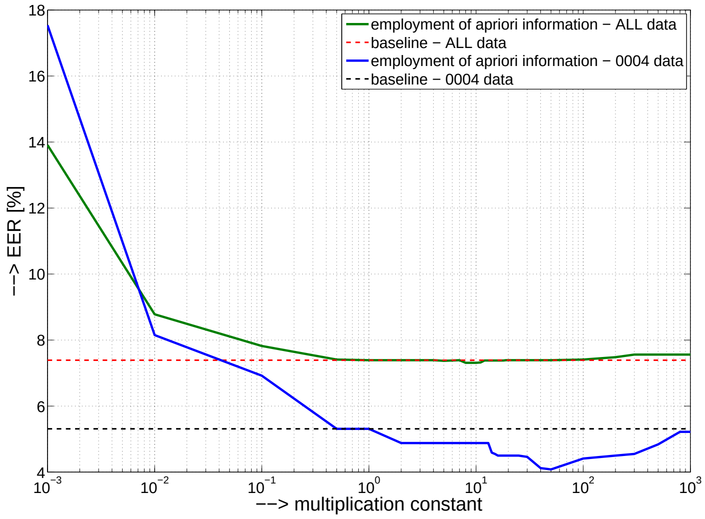


Figure 4: Overall Equal Error Rates (EERs) of STD when additional prior information is exploited: (a) Detection over all recordings (green curve and red line); (b) Detection only for recording 0004.

2.7 Discussion of results

This document summarizes experimental results achieved on Klewel lecture recordings. Our goal was to objectively measure the performance of LVCSR systems currently available at Idiap. Five LVCSR systems were tested and achieved performances were compared in a task of ASR. Then, the best system was chosen and its ASR outputs were used to perform STD. In order to be able to measure ASR and STD performances, approximately 70 minutes of audio data was manually annotated.

Figure 2 shows ASR results (in terms of WER) for all 5 LVCSR systems over all processed recordings. In overall, AMIDA IHM, 16kHz, 3-pass system performed the best. Pass 3 gives approximately 6% absolute improvement over the pass 1. The LVCSR systems based on MDM and CTS models performed about 5% (absolutely) worse. We also tested (close to) ”real-time” systems, provided by Juicer and STK decoders. The overall degradation in WER is about 15-20% compared to the best IHM P3 system.

Figures 3 and 4 represent detection performances (DET curve and EER plot, respectively) of the best LVCSR system (IHM P3) in the STD task.

Achieved EER estimated over all audio recordings (for 312 spoken terms) is equal to 7.4%. We also experimented with additional source of information available from associated text slides. 42 spoken terms, which also appear in text slides, were among those selected for STD. For these terms, corresponding posterior probabilities obtained from the index (only for 0004 recording), were modified by a prior. Achieved experimental results showed that such an approach can increase the detection EERs by about 25% relatively (from 5.3% to 4.1%).

3 Demonstrator

3.1 Introduction

3.1.1 Background

Live and on-demand Internet broadcasting of lectures in the workplace, at conferences and in educational settings has attracted more and more interest due to improvements in network bandwidth, computer performance, and compression technologies. Many corporations make seminars and training sessions available for employees who cannot attend a live presentation [11].

Some of the main online examples are:

- Google Tech talks: http://research.google.com/video.html
- iTunesU: http://www.apple.com/education/mobile-learning/
- MIT World: http://mitworld.mit.edu/
- VideoLectures.net: http://videolectures.net/
- ePresence: http://epresence.tv
- Omnisio: http://www.omnisio.com/.

Regarding speech recognition for such applications, Google announced recently automatic caption for YouTube videos. More and more webcasting systems [8, 1, 9] are using speech recognition systems to index the content of lectures.

3.1.2 Goals

The overall idea behind speech-to-text systems is to make use of it for a specific task hopefully useful for humans. Klewel has developed tools to browse, search and visualize audio-visually captured conferences. We are going to focus on the searching and browsing aspect of visually captured conferences. The goal is to create a tool to allow easy and fast access to the information based on speech information. This part of the project requires to focus on man-machine interaction by providing an easy and intuitive graphical interface.

This demonstrator is using the ASR output in two different ways. Firstly, it allows to display the ASR text along with the video and audio while maintaining the synchronization with all the streams. Secondly, it should let the end user to make search into the ASR data, and therefore act as data retrieval system. For later integration purposes, the demonstrator is based on the existing Klewel conference player.

3.1.3 About the Klewel conference player

The current Klewel conference player is coded in ActionScript 3.0 and his compiled with Adobe Flex Builder 3 in Adobe Flash format. It is able to display synchronously the orators video stream with the slides images. The player can also handle videos stream in slides. As a backend technology, it uses a Flash Streaming Server. Because of the Flash technology, the player can easily be embedded into any existing webpage, it is also platform independent.

4 Detailing some of the mechanisms

It seemed important to make the point on some technical aspects of the web-based graphical interface. Of course, detailing every part of it would not be very interesting and would be out of the scope of this report. Instead, we do prefer to focus on aspects that we found interesting.

4.1 Application's flow and overview

Figure 5 depicts the normal flow of the Player application. We are here showing a basic scenario where the application is embedded into a web page. When the web server receives the query from the client's browser it is redirected to the web server that hosts the application (1,2). The file is then sent to the client (3) and executed locally. The first action of the application, after being initialized, is to ask for the data (whether from an XML file or from a database using in this case a web service (not shown on the figure)) (4). Once the data is downloaded, the application establish a connection with the streaming server and downloads all the slides thumbnails from the web server (5).

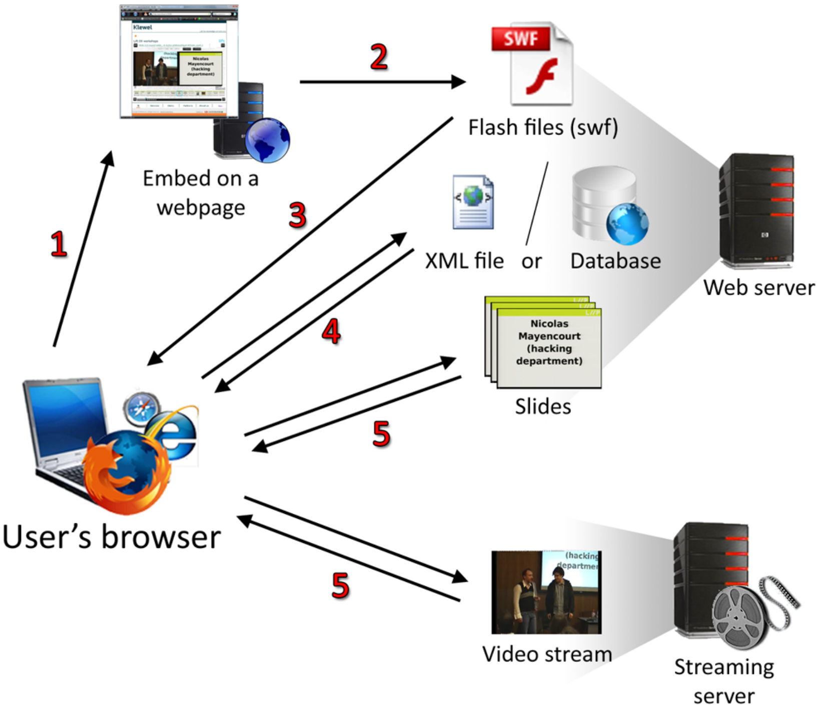


Figure 5: Player's flow: (1) Browsing to a page where the application is embedded (2) The server redirects to where the application is hosted (3) The application's file is sent to the client (4) The application runs and query for data (either for XML or database) (5) The application downloads the slides and connect to the streaming server for playing the video(s).

Despite it seems to be a relatively long path with back and forth queries, this process is quite fast and takes only a few seconds to complete with modern connexions.

4.1.1 Automatic captioning display

Having the ASR data available allows us to enhance the player, adding an automatic subtitling feature. Figure 6 shows the automatic captioning in action. This feature displays subtitles generated from ASR data and synchronized with the audio signal from the orator's video stream. Each subtitle is a suite of words with a start and end time computed automatically. The player simply display or hide them accordingly to the speaker's video

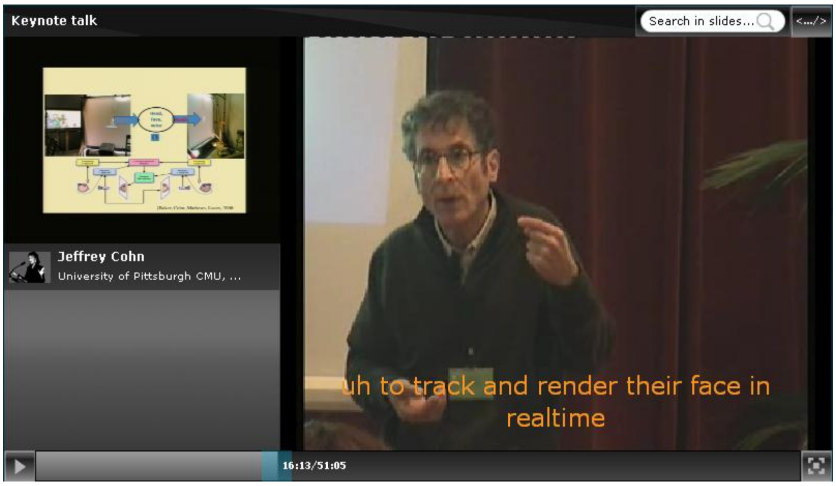

Keynote talk
Search in slides...
<.../>
Jeffrey Cohn
University of Pittsburgh CMU, ...
uh to track and render their face in realtime
16:13/51:05

Figure 6: AMIDA close-captioning demo

stream playhead time.

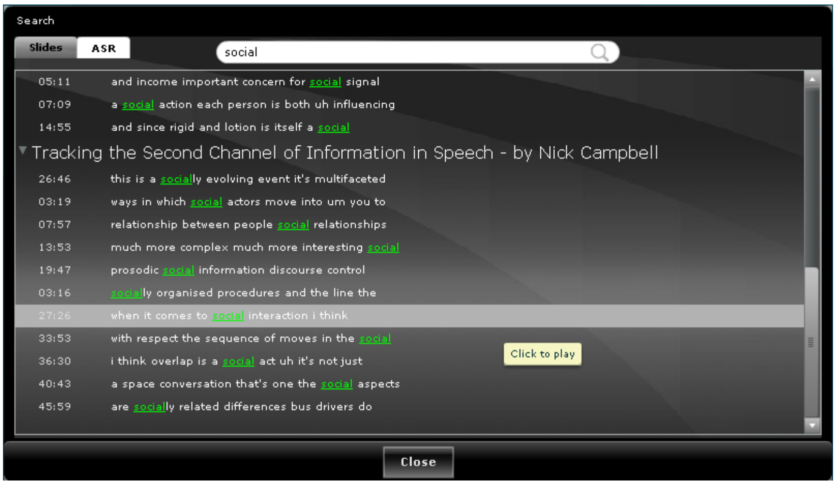
4.1.2 Limitations

As the ASR data does not contain any semantic information. There is no simple way to detect where a sentence may start or end. Furthermore, there is no punctuation nor uppercase lettering in the data. Finally, as the ASR audio recognition engine is not giving 100% results, some words may be erroneous. This may be confusing for the user who is watching if he or she is used to standard movies subtitles formatting, with clean and nice formatting. This limitation is not too dramatic, and the added value of ASR subtitling overcomes those minor drawbacks.

4.1.3 Data retrieval system

The existing player allows the users to search through the textual content of the slides. This content is generated by a special OCR (text recognition) algorithm that generate textual data from the captured slides images. Similarly we would like to allow the users to search into the ASR data. Figures 7 and 8 show the information retrieval feature of the demo in action.

The main challenge is to create an interface that can:

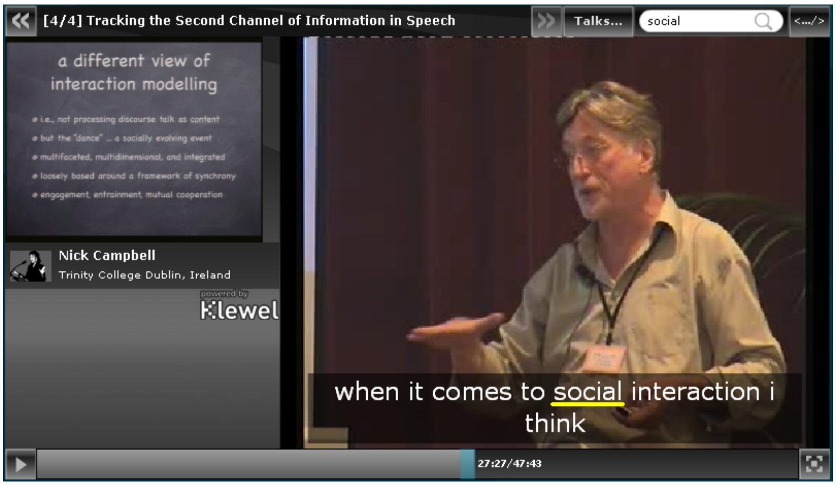


05:11 and income important concern for social signal
07:09 a social action each person is both uh influencing
14:55 and since rigid and lotion is itself a social

Tracking the Second Channel of Information in Speech - by Nick Campbell

26:46 this is a socially evolving event it's multifaceted
03:19 ways in which social actors move into um you to
07:57 relationship between people social relationships
13:53 much more complex much more interesting social
19:47 prosodic social information discourse control
03:16 socially organised procedures and the line the
27:26 when it comes to social interaction i think
33:53 with respect the sequence of moves in the social
36:30 i think overlap is a social act uh it's not just
40:43 a space conversation that's one the social aspects
45:59 are socially related differences bus drivers do

Figure 7: AMIDA retrieval demo



[4/4] Tracking the Second Channel of Information in Speech

a different view of interaction modelling

- i.e., not processing discourse talk as content
- but the "dance" ... a socially evolving event
- multifaceted, multidimensional, and integrated
- loosely based around a framework of synchrony
- engagement, entrainment, mutual cooperation

Nick Campbell
Trinity College Dublin, Ireland

when it comes to social interaction i think

Figure 8: AMIDA retrieval demo

- Fit into the existing player's interface (blend seamlessly with the existing search feature).

- Display efficiently the results so the user is not confused or distracted by a too large amount of potentially wrong information.

To achieve it, we decided to provide the end user with a way to filter the amount informations. Concretely, we have added on the interface a slider that, given a search, will filter the results based on the confidence probability of the ASR engine (this value is computed by the engine upon recognition process).

5 Conclusion

This project was one of the first attempts to apply and evaluate speech recognition in the context of lectures captured in real environments. By lecture recordings we do not only mean academic lectures in universities but also keynotes and talks presented in congresses, workshops or general assemblies. These Klewel talks were originally not specifically captured for speech recognition purposes. Intuitively we knew that the acoustic environment of a congress room is quieter than a busy meeting room. In this application, we do not have the problem of over-lapping speech and multiple speakers that we have in multi-party meetings. Better acoustic conditions lead to better speech recognition performances. The ideal case for speech recognition is a professional English native speakers talking in a radio-like studio with professional microphone adequately placed. Testing the AMIDA speech recognition tools from meeting to conference applications is a way to approach this ideal case.

The results achieved in this project are very promising. The results have not been evaluated qualitatively via a user study due to a lack of time at the end of the project. But the results show that the AMIDA speech recognizer is quite stable and robust through different speakers (gender, accent) and conditions (microphone type, conference room, background noise, gain level). The Klewel demonstrator allows to visualize those results. The demonstrator also shows that speech recognition can help in the context of lecture recordings: it can potentially help transcribers in their work, and it can help retrieving a video sequence of interest from a particular word that was pronounced at that time. At conferences captured by Klewel, it can happen that speakers (e.g., politician - here in French: http://www.idiap.ch/grandconseil) do not use slides. In that case, OCR indexing is useless. Such a system as developed in this mini-project could definitely help adding automatic indexing based on speech for later information retrieval.

References

[1] Keith Bain, Sara H. Basson, and Mike Wald. Speech recognition in university classrooms: liberated learning project. In *Assets '02: Proceedings of the fifth international ACM conference on Assistive technologies*, pages 192–196, New York, NY, USA, 2002. ACM.

[2] J. Dines, J. Vepa, and T. Hain. The segmentation of multi-channel meeting recordings for automatic speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1213–1216, Pittsburgh, USA, 2006.

[3] L. Burget et. al. Combination of strongly and weakly constrained recognizers for reliable detection of OOVs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4081–4084, Las Vegas, USA, 2008.

[4] T. Hain, M. Karafiat, J. Dines, I. McCowan, M. Lincoln, G. Garau, V. Wan, R. Ordelman, and S. Renals. The development of the AMI system for the transcription of speech in meetings. In *proc. of the 2nd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, page 12, Edinburgh, UK, 2005.

[5] Thomas Hain, Vincent Wan, Lukas Burget, Martin Karafiat, John Dines, Jithendra Vepa, Giulia Garau, and Mike Lincoln. The AMI system for the transcription of speech in meetings. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume IV, pages 357–360, Hononulu, US, 2007.

[6] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proceedings of EUROSPEECH*, pages 1895–1898, Rhodes, Greece, 1997.

[7] Darren Moore, John Dines, Mathew Magimai Doss, Jithendra Vepa, Octavian Cheng, and Thomas Hain. Juicer: A weighted finite-state transducer speech decoder. In *Proceedings of the 3rd Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, 2006.

[8] Cosmin Munteanu, Gerald Penn, Ron Baecker, and Yuecheng Zhang. Automatic speech recognition for webcasts: how good is good enough and what to do when it isn't. In *ICMI '06: Proceedings of the 8th international conference on Multimodal interfaces*, pages 39–42, New York, NY, USA, 2006. ACM.

[9] Miltiades Papadopoulos and Elaine Pearson. Accessible lectures: moving towards automatic speech recognition models based on human methods. In *Assets '08: Proceedings of the 10th international ACM SIGACCESS conference on Computers and accessibility*, pages 273–274, New York, NY, USA, 2008. ACM.

[10] Artem Peregoudov, Alessandro Vinciarelli, and Hervé Bourlard. Towards using slide information to enhance speech transcription of meetings. IDIAP-RR 06-01, Idiap, January 2006.

[11] Cha Zhang, Yong Rui, Jim Crawford, and Li-Wei He. An automated end-to-end lecture capture and broadcasting system. *ACM Trans. Multimedia Comput. Commun. Appl.*, 4(1):1–23, 2008.

A Final keyword set

0:abstract 1:abstraction 2:accelerate 3:accent 4:accomplish 5:achievement 6:activity 7:actor 8:adam 9:affect 10:age 11:aggressive 12:agreement 13:alternative 14:america 15:analysis 16:appearance 17:art 18:article 19:artist 20:asleep 21:assignment 22:audacious 23:audience 24:audio 25:australian 26:avatar 27:background 28:bar 29:barrel 30:best 31:beverage 32:bill 33:biography 34:biological 35:blind 36:bonnie 37:book 38:booklet 39:brand 40:breaking 41:brilliant 42:broadcaster 43:canada 44:capitalist 45:carbon 46:category 47:challenge 48:characteristic 49:chart 50:childhood 51:class 52:clinical 53:coincidence 54:colour 55:combination 56:communication 57:complex 58:computer 59:concern 60:confederate 61:construction 62:consumption 63:content 64:conversation 65:coordination 66:copy 67:cost 68:council 69:craftsman 70:culture 71:curiosity 72:decades 73:decision 74:deference 75:deliver 76:depressing 77:description 78:design 79:designer 80:device 81:dialogue 82:difference 83:digital 84:disambiguation 85:disease 86:dollars 87:dominate 88:down 89:dynamic 90:economic 91:education 92:effect 93:egomaniac 94:emotion 95:environment 96:episode 97:especially 98:essays 99:ethics 100:europe 101:evaluation 102:everyone 103:example 104:expectation 105:explorer 106:expression 107:face 108:facial 109:fashion 110:faster 111:feedback 112:feeling 113:female 114:film 115:florence 116:formation 117:formula 118:foundation 119:france 120:friend 121:fundamental 122:future 123:gender 124:goldsmith 125:graphics 126:harmful 127:head 128:health 129:history 130:home 131:human 132:humanity 133:hydration 134:impact 135:importance 136:improvise 137:incredible 138:individual 139:influence 140:intention 141:interaction 142:interest 143:interface 144:italy 145:japanese 146:juggler 147:kilos 148:lab 149:lack 150:latin 151:law 152:leader 153:learning 154:life 155:lifetime 156:lincoln 157:literature 158:luxury 159:machine 160:male 161:man 162:market 163:masterpiece 164:masters 165:material 166:maximum 167:median 168:michelangelo 169:mind 170:minimum 171:minute 172:mirroring 173:model 174:module 175:monitor 176:monkey 177:morphology 178:motion 179:mouse 180:mutually 181:myself 182:nature 183:notion 184:obesity 185:obscure 186:observation 187:oil 188:okay 189:online 190:ontario 191:opening 192:operation 193:opportunity 194:organisation 195:ottawa 196:overall 197:overlap 198:part 199:partially 200:participation 201:people 202:percent 203:perception 204:performance 205:period 206:perpetrate 207:person 208:perturbation 209:piece 210:place 211:plastic 212:player 213:plenary 214:point 215:poland 216:portfolio 217:position 218:presence 219:principle 220:probably 221:problem 222:process 223:productivity 224:profession 225:program 226:project 227:promoting 228:proof 229:prosody 230:publication 231:question 232:quote 233:ranting 234:reaction 235:reciprocity 236:recording 237:relax 238:renaissance 239:response 240:responsible 241:results 242:retire 243:right 244:role 245:rotation 246:satisfaction 247:science 248:sculptor 249:search 250:second 251:security 252:sense 253:sex 254:signal 255:silence 256:slide 257:smith 258:social 259:sort 260:soundtrack 261:source 262:speaker 263:speech 264:standard 265:standpoint 266:states 267:structure 268:student 269:study 270:stuff 271:sugar 272:support 273:surprise 274:swiss 275:symmetry 276:synthesis 277:system 278:talk 279:tap 280:tape 281:teamwork 282:technology 283:telephone 284:term 285:texture 286:thailand 287:thing 288:thousand 289:title 290:topic 291:toronto 292:town 293:track 294:united 295:utterance 296:value 297:velocity 298:video 299:vowel 300:watch 301:water 302:way 303:week 304:wellness 305:west 306:woman 307:workshop 308:world 309:xerox 310:year 311:yesterday

B Keywords from slides

Construction Social Signal Science Foundation Symmetry formation breaking Coordination Head expression prosody influence perception biological motion disambiguation conversation dynamic process observation perturbation system Model device face episode Actor male female Reciprocity confederate effect Velocity Symmetry blind assignment gender sex avatar study results