The Idiap Research Institute introduces

What about a system able not only to recognize who is speaking during a meeting but to understand the conversation and propose relevant documents like past reports, web pages as well as summaries and analyses of past meetings?

Artificial Intelligence and Speech Recognition specialists at IDIAP are developing that smart automated assistant.

BY HERVÉ BOURLARD, DIRECTOR OF THE IDIAP RESEARCH INSTITUTE AND ANDREI POPESCU-BELIS, SENIOR RESEARCHER

The automatic recognition and transcription of conversational human speech is a long-term goal at the convergence of several disciplines such as signal processing, human language technology, and artificial intelligence. In the nearly twenty years since its foundation in Martigny, the Idiap Research Institute has contributed pioneering research to the challenge of Automatic Speech Recognition (ASR), and continues to be at the forefront of scientific and technical advancement in this area, which has a potentially very large number of applications.

More and more scientific and technological advances are making ASR systems increasingly user-friendly and intuitive depending on the types of applications. High performance ASR systems, even for unconstrained conversational speech, are now within reach, opening up new opportunities for applications.

The promising field of ASR presents a range of challenges which Idiap and its partners are addressing through large EU projects, namely AMI (Augmented Multi-party Interaction), and AMIDA, (AMI with Distance Access). These significant efforts aiming to draw user-friendly commercial applications using ASR have been ongoing for over 8 years and benefited from over 25 million Euros. Some immediately promising advanced applications using real-time ASR include a “virtual secretary” that suggests relevant reference documents or Web pages during meetings.

"IDIAP's virtual secretary listens to a meeting and searches quietly in the background for the most relevant documents."

Andrei Popescu-Belis

CHALLENGES FOR SPEECH RECOGNITION: INPUT SIGNALS

In optimal conditions, namely with a single speaker using a high-quality microphone in a noiseless environment, the performance of ASR systems has reached levels only slightly below human performance. This is especially the case when using an ASR system that has “learned the user’s voice”, as in the personal dictation systems already developed in the 1990s. However, performance degrades quite rapidly when one or more of the above conditions are not fulfilled: typically, for conversations involving several people (hence possible overlaps in speech), using far-field microphones, and in the presence of non-speech noise. Solving such challenges in the context of multi-party meetings, by developing a functional (and preferably real-time) ASR system, was the
the "virtual secretary"

“High performance Speech Recognition System, even for unconstrained conversation, are now within reach.”

long term goal of speech technology in the European AMI Consortium.

The AMI system for large vocabulary conversational speech recognition was developed and tested specifically for the meeting environment, with several possible types of input signals: from individual head microphones, or from multiple microphones on the meeting table, in a microphone array whose geometric configuration is known. Microphone arrays enhance speech signals thanks to beamforming, which is a technique that filters and combines individual microphone signals in order to enhance the audio coming from a particular location in the meeting room. In addition to an improved speech signal, microphone arrays also help finding which participant is speaking at a given moment, an important task known as diarization.

ARCHITECTURE AND RESULTS OF THE ASR SYSTEM

The AMI ASR system makes use of advanced statistical ASR technology, based on significant exploitation and enhancements of the so-called “hidden Markov models” for acoustic modeling of the pronunciation variability of the lexicon words. The system also uses sophisticated statistical language models referred to as “N-gram” models, which predict the probability of a specific word being pronounced given the N previous ones. For conversational speech recognition in meetings, the number of lexicon words can be as high as 100 000 and N can be as high as 5.

The complete system as used in competitive evaluations operates in no less than ten passes over the input data, exploiting more and more detailed models. The initial pass only serves to obtain a rough transcript to provide input for adapting acoustic models, while the following passes generate bigram word lattices which are expanded using 4-gram language models and rescored using models that are differently trained, for example on varying training data.

Each pass normally outputs both a first-best result and a word-graph: the latter is used to constrain the search space for subsequent stages, and allows for output combination of several complementary models. Depending on the constraints in terms of processing time, the system complexity is usually increased by the number of passes, with gains in later passes that tend however to decrease with the number of passes.

Recently, a major achievement has been the design of a real-time version of the ASR system, keeping up with a speaker’s production rate. The performance of the complete non-real-time system reaches about 25% word error rate on signals from individual head-mounted microphones, an extremely competitive performance assessed in international evaluation campaigns. Accuracy and speed are especially important for commercial applications such as those targeted by Koemei, a young Idiap spin-off.

In complete applications though (besides pure dictation systems), the ASR system is in reality only the first processing step of content extraction from spoken conversations, which includes other stages such as diarization, named entity recognition, syntactic chunking, dialogue act segmentation and classification, topic segmentation, and summarization. All these aspects of content extraction contribute to facilitate search in multimedia recordings of events such as conferences, a capacity that is put to work in another Idiap spin-off called Klevel. To improve the utility of ASR for these analyses, future objectives include...
improving the robustness, speed, and accuracy of the system, as well as dealing with larger or more flexible vocabularies of recognizable words. The addition of new languages, in particular Swiss national ones, is also under work.

APPLICATION OF ASR TO A “VIRTUAL SECRETARY”
A very large number of more user-oriented applications have been considered for automatic speech recognition, from dictation-based interfaces replacing keyboards to search-and-retrieval from spoken archives and to human-computer voice-based dialogue. In current work at Idiap, we are also paying particular attention to the integration of content extraction modules into several types of meeting assistants, which are systems that can help meeting participants with various tasks, in close-to-real-time (in some cases, delays of several seconds or even minutes may be acceptable). In particular, Idiap has been working on an application of real-time ASR to design a speech-based document retrieval system called the “Automatic Content Linking Device”, but often referred to as a “virtual secretary”.

This prototype answers the well-known need for information access as a secondary activity, for instance at a time when users are involved in a principal activity that does not allow them to use a traditional search interface (with a keyboard, mouse, and display), and does not even allow them to concentrate fully on initiating a search. Such a particular need for secondary search arises during meetings. Often people need further information during a meeting (e.g., previous meeting minutes, Google search results, etc.), but they cannot lay their hands on it, at least not during the meeting itself, because searching would require an interruption of the discussion. And yet, producing the right piece of information at the right time can change the course of a meeting.

A CAREFUL LISTENER
The Automatic Content Linking Device answers this need by listening to a meeting and searching quietly in the background for the most relevant documents and past meeting segments from a multimedia database, or from the Web. The past meeting segments are made available thanks to offline speech recognition, and the documents include past reports, emails, or presentation slides. The system performs searches at regular intervals over the multimedia databases, with a search criterion that is constructed based on the words that it recognizes automatically from the ongoing discussion using real-time ASR.

The system keeps up-to-date search results ready for whenever someone in the meeting feels the need to consult them, and is also able to indicate which of the recognized words have enabled the retrieval of each document. Participants in the discussion thus only need to decide if they want to explore any further, and possibly introduce in their subsequent discussions, the documents or past meeting fragments retrieved automatically for them. The system can be used by each participant, but another approach is to have it used jointly by all participants, on a dedicated projection screen. And, it can also be used to enrich a past recording with documents. Search on demand at a given moment, as opposed to regular intervals, is also possible.

While other query-free systems for just-in-time retrieval have been proposed in the past, the Idiap system is the first one that is implemented in the context of human conversations, based on ASR and keyword spotting. Moreover, it is also the first system to give access to processed multimedia recordings, documents and websites at the same time, in a fully autonomous way.

The Automatic Content Linking Device is a joint achievement that was coordinated by Idiap within the European AMI Consortium and the ongoing Swiss National Centre of Competence in Research (NCCR) on "Interactive Multimodal Information Management" (IM2). The system is composed of several modules, which were completed partly at Idiap and partly at collaborating institutions. The first prototype was designed in 2008, and since then several versions have been demonstrated at academic or user-oriented events.

The prototypes have received positive verbal evaluation from potential industrial and academic partners, who proposed additional application scenarios that were of interest to them, and provided useful feedback and suggestions for future work. The most recent version of the system is being installed in one of the collaborative spaces of the EPFL Rolex Learning Center, in collaboration with the CRAFT team. The goal is to assist and stimulate discussions in these rooms, in an education-oriented perspective, while integrating the speech capture and information suggestion functions into the own specific architecture of the rooms.

IDIAP’S SPINOFF
THE VIRTUAL VIDEO
EDITOR OF KLEWEL
Founded in 2008 as a spin-off of IDiap, Klewel provides solutions for capturing, searching and sharing the information contained in multimedia digital recordings of conferences, its system can handle multiple cameras, one or multiple audio channels (e.g., the original speech and its interpretation in several foreign languages) and the projected slides. The system is completely non intrusive as data is captured directly from the sources and synchronized. The speakers do not need to provide Klewel with any original slides. Once the capture step is performed, all the data is uploaded to servers to be processed. This solution automatically references the full content of presentations. The content of the slides is automatically indexed. Multimedia files are encoded into a format suitable for the web. These presentations are quickly published and then fully accessible from an Intranet or a website. Any interested parties can immediately access and retrieve specific information without having the need to play back the full presentation. Search using keywords will retrieve relevant information from all archived presentations, across multiple conferences if needed.

THE VIRTUAL CLERK OF KOEMEI
Thanks to the cutting-edge research in ASR conducted by IDiap and its partners, Koemei, a spin-off of the Martigny-based institute incorporated in 2010, provides an advanced solution that automatically transcribes conversational speech into text – thereby opening up a wide range of new computer applications based on spoken human language.

The cloud-based speech recognition solution of the young company is specifically designed for multiparty conversations. Koemei targets particularly the transcription markets for meeting recording, lecture capture, videoconferencing, telepresence, multimedia indexing, speech mining and analytics, search engine optimization and voice-mail-to-text. Its technology enables global businesses, government agencies, educational institutions, telecom operators, professional service providers and multimedia organizations to use speech to power multiple mission-critical applications and services.

Moreover, third-party application developers can access this speech recognition platform to develop specific solutions.