Voice Report 2013

Executive summary and state of the art in speech to text

HERIOT
WATT
UNIVERSITY

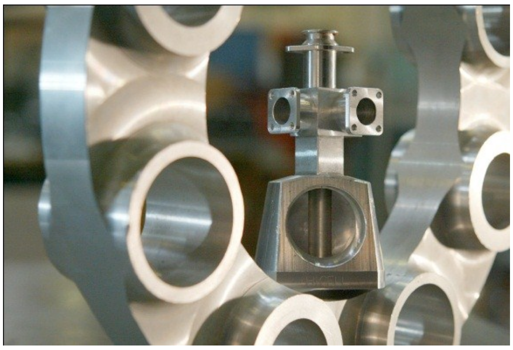
Interface demo executive summary

Decisions found quickly and easily

The demonstration interface aims to present a solution to a problem faced by many companies that undertake extensive design and planning phases. These phases often consist of many meetings over a period of months. Minutes from these meetings record actions, but rarely the rational or motivation behind these decisions. The solution this demonstration puts forward aims to show proof of concept that meeting data can be recorded, organised and searched through using modern computations techniques, the reasoning behind important decisions can be found quickly and easily.

Powerful search method

The search mechanism allows the meetings to be searched using topics found through Topic Modelling. This allows large text data sets to be broken down into their topical structures automatically, and can be used to create a more powerful search method than just traditional “key word” search. The search mechanism will then retrieve meetings, and rank them according to their relevance to the topic searched for.



Rapidly browse meetings

In order to facilitate quick and efficient browsing of the search results, the way in which a user is able to index into specific meetings points must be easy and intuitive. A graphical method was developed for doing this. The graph shows occurrences of topical words over the time of the meetings, a peak indicates a point in the meeting where the topic searched for has a high probability of being discussed, clicking on this peak will take the playback to the corresponding point in the meeting. This enables the user to rapidly brows meetings to find the relevant discussion.

- **Uncover decision rational from meetings**
- **Topic Modelling speeds search**
- **Train new staff easily**

- **Rapidly browse meetings**
- **Intuitive graphical interface**
- **Demo uses AMI corpus**
- **http://bit.ly/voice-project**

State of speech-to-text executive summary

Widely used in mobile devices

Current speech-to-text applications are widely used in mobile device interfaces and for dictation applications. Although relatively high accuracy and speed can be achieved with systems such as Nuance's Dragon, speech recognition is still far from being a primary input method, as most users will still opt for standard text input methods for most tasks. The ideal of user-independent continual speech-to-text, where a device is constantly listening and can understand all our requests and commands is not yet realised, but it is perhaps not so far off.

Important advances in the field

There have been a number of important advances in the field in the past two years. Microsoft demonstrated a real-time voice translation system in October 2012. This used very accurate, live, speech to text transcription, performing with an accuracy rate of approaching 90%. Apples Siri interface was also another big step in automatic speech recognition, with many believing that Nuance technology powers the speech to text portion of the system. Siri computes the speech to text on Apples dedicated servers, giving Apples access to large amounts of data (Big Data) that can be used for training of speech recognition algorithms. Big Data is also a term used by Google when talking about advancing speech to text technology, the idea being that having a large data set to train your model is more important than the complexity of the model.

The mobile market is a major driving factor

One of the main challenges that remains with text to speech beyond general accuracy, is known as the "cocktail party" problem. Speech to text systems currently find it very hard to deal with multi-party speech (differentiating many speakers at once) and background noise. These two problems are currently a focus of much research. The mobile market is also a major driving factor behind speech to text technology, but companies are also looking to introduce interfaces into other areas. Nuance are applying their ASR technologies to car and television interfaces, with Apple being rumoured to also be making a TV controlled by Siri.

- **Microsoft, Apple, Google, Nuance.**
- **Nuance is the technology leader.**
- **Nuance claims up to 99% accuracy.**
- **Cocktail party environment is a problem**

- **Big data improves accuracy solving accents, multiple speakers**
- **Currently huge amount of research.**
- **Limitations will be solved probably in 2-5 years**

Voice Project Report

Motivation

The motivation behind this project is the need for improved meeting review capabilities in industry. Companies that are involved in extensive planning and design phases often find themselves spending large amounts of time trying to uncover the rationale behind decisions made in past meetings. Meeting minutes often only record actions, and lack necessary information behind these actions.

This can lead to confusion and misunderstanding, important decisions may be overturned and if the planning and design phases are long enough that new staff come on board during their course it can also be hard for these new employees to come up to speed with the motivation behind design decisions.

- Companies can find it hard to keep track of decisions made in meetings
- Minutes often only record actions, and not the reasoning behind these actions
- It is time consuming to try and uncover decision rational later in the life of a project
- New staff may find it hard to get to grips with the ins and outs a particular design process
- This can lead to the overturning of important decisions, with costs incurred

Solution

This project envisions a solution where multimodal data (video, audio, transcripts, slides) from meetings could be recorded. The data collected could then be organised digitally and made available for search and retrieval.

Automatic Speech Recognition (ASR) could be used to transcribe meeting audio, and computational methods such a Topic Modelling would enhance the mechanism by which the transcriptions could be browsed. Topic Modelling allows the thematic and topical structures of large text corpora to be uncovered automatically, Relevant meeting data could then be quickly retrieved, with an intuitive system allowing a user to index into meeting playback, the appropriate meeting information can then be found efficiently.

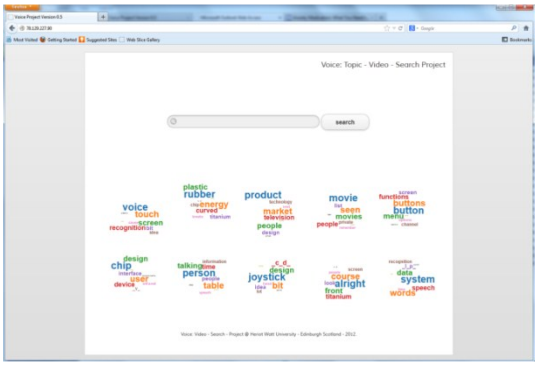
This will allow decisions to be understood and recalled in later meetings, avoid wasted time searching for decision rational and in overturned decisions, and also limit time spent discussing matters already covered in past meetings. New employees and team members can be quickly brought up to speed on aspect of a particular project. With rapid browsing discussions and data from past meeting can even be accessed during a current meeting.

- Meeting digitally recorded and audio transcribed
- Topic Modelling used to enhance search
- Fast search mechanism and intuitive interface
- Allow user to rapidly index into relevant points in meeting playback

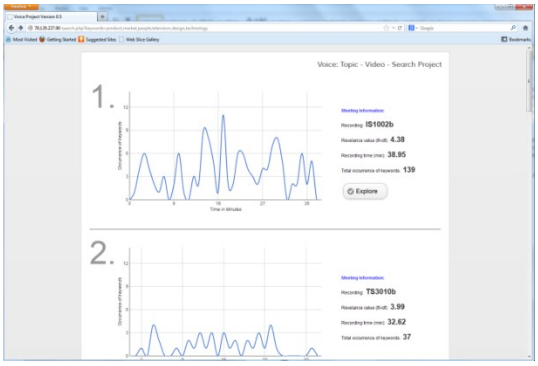
Demo Interface

As proof of concept a demo search and playback interface has been developed using the AMI meeting corpus. This corpus contains over 100 hours of transcribed and annotated multimodal meeting data, which translates to a text corpus of around 1.9 million terms. The Demo shows that search and retrieval can be conducted quickly, using the topics uncovered by the Topic Modelling. The meetings are ranked according to their relevance to the search and the likely points of relevance in the meeting are represented graphically. This graph shows at which points in the meeting relevant discussion is taking place. By clicking on this graph, the user can index rapidly into specific meeting point, and decide if it is relevant, providing an efficiant browsing method.

- Topic based search



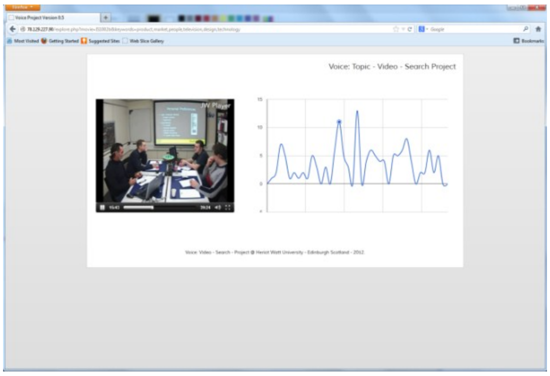
- Meeting retrieval and ranking



Meeting Information:
Recording: IS1002b
Relevance value (tf-idf): 4.38
Recording time (min): 38.95
Total occurrence of keywords: 139

Meeting Information:
Recording: TS3010b
Relevance value (tf-idf): 3.99
Recording time (min): 32.62
Total occurrence of keywords: 37

- Index into meeting playback


Voice Project Version 0.5
Voice: Topic - Video - Search Project
JW Player
15
10
5
0
Voice: Video - Search - Project @ Heriot Watt University - Edinburgh Scotland - 2012

Topic Modelling

Today almost all text created is stored digitally, from academic libraries to news archives, all the data is available to be searched and browsed using computers. The most common form of search is “key word” searches, where a user inputs words or a phrase, and the search engine retrieves documents where these words occur, and ranks them by some method deemed fit. Topic Modelling is a computation method that assumes a probabilistic model of creation for documents in a large corpus, and uses statistical methods to decompose the corpus and its documents into “topics”, where a *topic* is a probability distribution over words. This allows us to see the themes and topics that make up an entire corpus and its constituent documents, which in turn creates a new way for us to search and brows any large text data set. Rather than merely using the occurrence of *key words* or phrases, a user can search for the occurrence of a theme or topic, which is potentially a much more powerful search method.

- Large text corpora are usually searched using key word search
- Topic modelling uncovers the thematic structure of large text corpora

- Allows for more powerful search methods using topics

Challenges

The demo interface shows that the meeting search and browsing portion of the solution is currently achievable, the challenges lie in the meeting recording and transcribing. This would require reliable automatic speech recognition software, with word error rate low enough as to allow accurate meeting search and retrieval.

Summary

- **Uses topic modelling**
- **Meetings can be easily retrieved**
- **Intuitive graphical interface**
- **Automatic speech recognition**
- **Action can be embedded**
- **Index documents too**
- **Interface can be tune to requirements**
- **Search large corpus**

State of the art in speech-to-text

Automatic Speech recognition (ASR) is currently an currently an area of great interest in the tech industry. Daily we see ASR used in simple user interfaces, dictation applications and automatic information phone services. There are widely used, off the shelf systems, such as Nuance's Dragon that allow for fast and accurate speech to text transcriptions, however for most users, this is not the preferred method of input. The ideal of user-independent continual speech-to-text, where a device is constantly listening and can understand all our requests and commands is not yet realised, but it is perhaps not so far off. The field of ASR has seen a number of significant steps-forward in recent years with mobile technology acting as a major driving force.

- Widely used for dictation and human computer interfaces
- Ideal system not yet achievable
- Number of significant steps taken in recent years

Recent Advances

In October Microsoft unveiled a new real-time translation system in a live demonstration. The demonstration was made by Rick Rashid, chief research officer, and consisted of live transcription of his speech, showing the running average accuracy, the dictation was then translated in real-time in to Chinese, and output using a speech synthesiser. The live speech to text portion of the demo achieved a high accuracy of around 90%. The System was also demoed in the context of a phone call, where an English speaker and German speaker had a conversation in real-time, speaking their own language. The phone transcript was also saved, allowing for fully searchable phone call records.

Link to video and article covering Microsoft live translation demo:
http://www.technologyreview.com/news/507181/microsoft-brings-star-treks-voice-translator-to-life/

This shows that in controlled situations, with the appropriate microphones, and training, systems can readily achieve useable word accuracy rates.

Microsoft has also published research into the use of neural networks for large vocabulary speech recognition models. The highest performing models up until now have been of the Hidden Markov Model type, however by using Deep Neural Networks the research showed that for certain tasks the relative error rate could be dropped significantly, stating an improvement of over 30%. This is a significant improvement for automatic speech recognition.

Link to article covering Microsoft research into Neural Networks:
http://research.microsoft.com/en-us/news/features/speechrecognition-082911.aspx

Apple's high profile ASR interface was introduced to the iPhone 4S. Siri allows a user to query a system which then returns results, also using speech synthesis. Interestingly it appears that Siri quietly uses Nuance technology to do the speech recognition tasks, though this has not been official confirmed.

Though Siri performs with reasonable accuracy, users have been slow to embrace the system, and it has been received as mainly a gimmick, the real advantage to apple may lie in the amount of training data that can be gathered from users. As each query is handled by Apples serves, this means that as ASR becomes more commonly used by iPhone owners, Apple will hold a large database of speech-to-text actions, with which to train their systems. The importance of these large data bases (Big Data) is becoming more apparent to those developing ASR algorithms.

Link to article cover speculation on Siri using Nuance technology:

http://techcrunch.com/2011/10/05/apple-siri-nuance/

The importance of Big Data has also been highlighted by Google in their quest to improve speech-to-Text technology. The idea is that generally speaking, having a large data set to train your algorithm yields better results that trying to create better, more sophisticated algorithms, and this idea translates directly into the field of ASR.

As Big Data becomes a driving idea, the large tech companies (Apple, Microsoft and Google) are well placed to collect these large data sets. This explains why many text to speech applications are given away freely to users on devices, as the more people use, and become comfortable using these interfaces and applications, the more Big Data can be collected. The company that can collect the most data, may be the one to take the next big step.

Link to article covering Google's approach to Big Data:

http://gigaom.com/2012/10/31/google-explains-how-more-data-means-better-speech-recognition/

- Microsoft live speech translation
- Use of Deep Neural Networks
- Introduction of Apple's Siri
- Collection of "Big Data" used to train models

Challenges

Beyond improving the basic accuracy of speech-to-Text models, there are number of challenges that are the focus of current research. The problems of background-noise and multiple speakers pose the next hurdle to the advancement of the field. The problem is known as the cocktail party problem. In dealing with the situation of cocktail party, in order to hold a conversation with somebody, a person must be able to separate the sound of the person speaking from the background murmur of the party, and recognise that this sound is separate from sounds of people speaking nearby, and at the same time. Currently this is beyond a computational system to any degree of useable accuracy.

Link to "Automatic speech recognition in cocktail-party situations: A specific training for separated speech" by A. Marti et al (2011): **http://asadl.org/jasa/resource/1/jasman/v131/i2/p1529_s1**

- Cocktail party problem (background noise and multiple speakers)
- Challenge of separating unwanted signals and speakers
- Focus of research

Readiness for Meeting Capture

An ideal system for meeting capture would be minimal, easy to set up and unintrusive. The limitations of speech to text are such that an ideal system cannot currently be achieved, however a useable and useful set up is possible. Use of head-set microphones provides the audio fidelity required for good accuracy on individual users, provided that they don't pick up background noise and the user has trained the system, though this may be considered intrusive. As can be seen by the recent advances in the field, close microphones can achieve very high accuracy rates, and the accuracy of the shelf ASR products like Dragon should keep increasing in the near future.

- Ideal system is minimal, easily set up and unitrusive
- Current tech can achieve usable accuracy
- But may be intrusive

The Near Future

The rise in use of powerful mobile devices is acting as a driving force for progress at present. As these devices increase in computing power and are able to handle the demands of ASR technologies, they become well placed to use such applications. The desire for mobile devices to have more hands free potential is clear, but there are many other applications where hands free interfacing is seen to be the future. Nuance are going to be applying their ASR technologies to cars and television interfaces, with apple being rumoured to also me making a TV controlled by Siri. As mentioned above, one challenges of these environments as opposed to a mobile device that allows a user to speak directly into, there may be much more background noise in the context of a car or TV.

- Mobile devices are a strong driving force
- Companies looking to add ASR to cars and TVs
- This may lead to improvements in background noise cancelation

Speech to Text Summary

ASR is a field of study that is already in a state of great fruition, and also has large challenges to overcome in order to reach its vision. It is already widely available in many forms, such as mobile interfaces and in dictation applications, and in the field is rapidly improving as has been discussed above, with many significant advances being achieved in recent years. The challenges of multiparty speaking and background noise elimination are ones that once overcome could lead to a major increase in the use of speech-to-text technologies, and as these technologies become more widely used, the ability to use big data to improve train algorithms will improve accuracy greatly. This leads to the conclusion that in the context of meeting capture, the ideal system may not yet be achievable, it is not far off.

Summary

- **Cocktail environment still a problem**
- **Can be solve with individual microphones**
- **Main players: Microsoft, Google, Apple and Nuance**
- **Nuance leader in speech to text for dictation**
- **Big data very important**
- **Lots of research using big data to improve deficiencies**
- **Most problems will probably be solve in 2-5 years.**

Contacts

Professor Mike Chantler

School of Mathematical & Computer Sciences,
Heriot-Watt University, EH14 4AS, UK.
Telephone: +44 (0)131-451-3352
Email: **M.J.Chantler@hw.ac.uk**

Stefano Padilla

Email: **S.Padilla@hw.ac.uk**

Patrick Campbell

Email: **Pcc5@hw.ac.uk**