From lab to real world: the FlyScribe system.

Gianni Lazzari, Alessandro Tescari, Brian Martin and Fabio Brugnara

FONDAZIONE BRUNO KESSLER (FBK), Trento, Italia
and
PERVOICE SPA

1. Introduction

Accurate transcription of spontaneous speech that is speaker independent and which functions in a wide variety of environments is beyond the capacity of state of the art systems. To reach this objective, robust technologies must be developed that can handle: a variety of subjects, multiple speakers, many styles of speaking, and a wide range of transmission channels and environments. And finally, these technologies need to be able to adapt transparently to any new condition.

It is well known that recognition performance is highly dependent on the kind of application, characteristics of the speaker, style of speaking, as well as on environmental conditions. There can be enormous differences in the rate of error: 1–2% for the best speakers to more than 30% for the worst.

The job of helping professionals with automatic transcription of human speech is no longer a dream. The product created by PerVoice with technology developed at Fondazione Bruno Kessler Research and Technology Center (FBK-irst, formerly ITC-irst) shows that it is possible to offer high quality transcription services on a nation-wide scale.

This article presents the results of FBK-irst research in this field in European Project TC-STAR(1), and how they have been industrialized in FlyScribe, an automated transcription service.

2. Automatic Speech Trascription: technology overview

Automated transcription of an audio signal aims at generating both a word-level transcription as well as meta data (e.g. the gender of the speaker, identification of named entities, punctuation, etc.) which can be used for a variety of ends such as automated indexing, automated reports, automated translation, and others.

Transcription systems are based on leading-edge technology that use large dictionaries. They are used in a variety of domains such as news broadcasts, audio-video files (like historical films and documentaries), meetings, classroom lectures, speeches delivered to conferences and legislatures, phone calls, etc.

The automated transcription of an audio signal is typically conducted in multiple steps. First, the audio signal is divided into acoustically homogeneous segments and the portions of the signal that contain speech are identified and grouped according to their spectral similarity. Next, the speech segments are transcribed using two or more recognition passes alternated with adaptations based on the acoustic models.

Currently, the most widespread approaches to speech recognition for large vocabularies are based on static signal recognition models. The main problems are acoustic-phonetic modeling, language models, and adaptation of the models. Further development is required to make the technology more robust to changes in topic, speaker, speaking style, transmission channel, and environment. At the same time, the technology will need to be able to adapt to new conditions.

Transcription system performance is usually expressed in terms of average word error rate (WER), that is, the average number of word recognition errors for every hundred words. As an example, a state of the art transcription system for “unrestricted” radio news has a WER of about 20%. Transcription performance can vary a great deal depending on input data. For American English, the average WER for radio announcers is about 10% using the DARPA Hub 4 benchmark. Performance drops significantly for spontaneous speech (15% WER). Similar performance drops occur with worsened audio quality: for example, telephone speech (20% WER) or non-native speech (more that 18% WER).

Experiments on recognition (using manual separation of speaker turns) made in the TC-STAR project (www.tc-star.org) on English and Spanish transcription of European Parliament sessions gave a WER less than 10%.

The FBK-irst transcription system [4,5] handles all the elaboration steps needed to transcribe an unknown audio signal. The system is able to identify non-speech elements and classify the speech portions into categories: male/female voice, wide/narrow band speech, noisy/quiet background, etc.

With respect to the performance cited above, the FBK system has a WER of 10.6% for radio announcers using the DARPA test and 9.7% for European Parliament sessions.

In 2006, the transcription system developed by FBK was released for use at RAI (the Italian public radio and television broadcaster), which began using it to archive audiovisual materials to allow information searching.

In 2007, the FBK transcription system was used to transcribe sessions of the Italian Parliament, demonstrating an error rate of about 11% in several experiments.

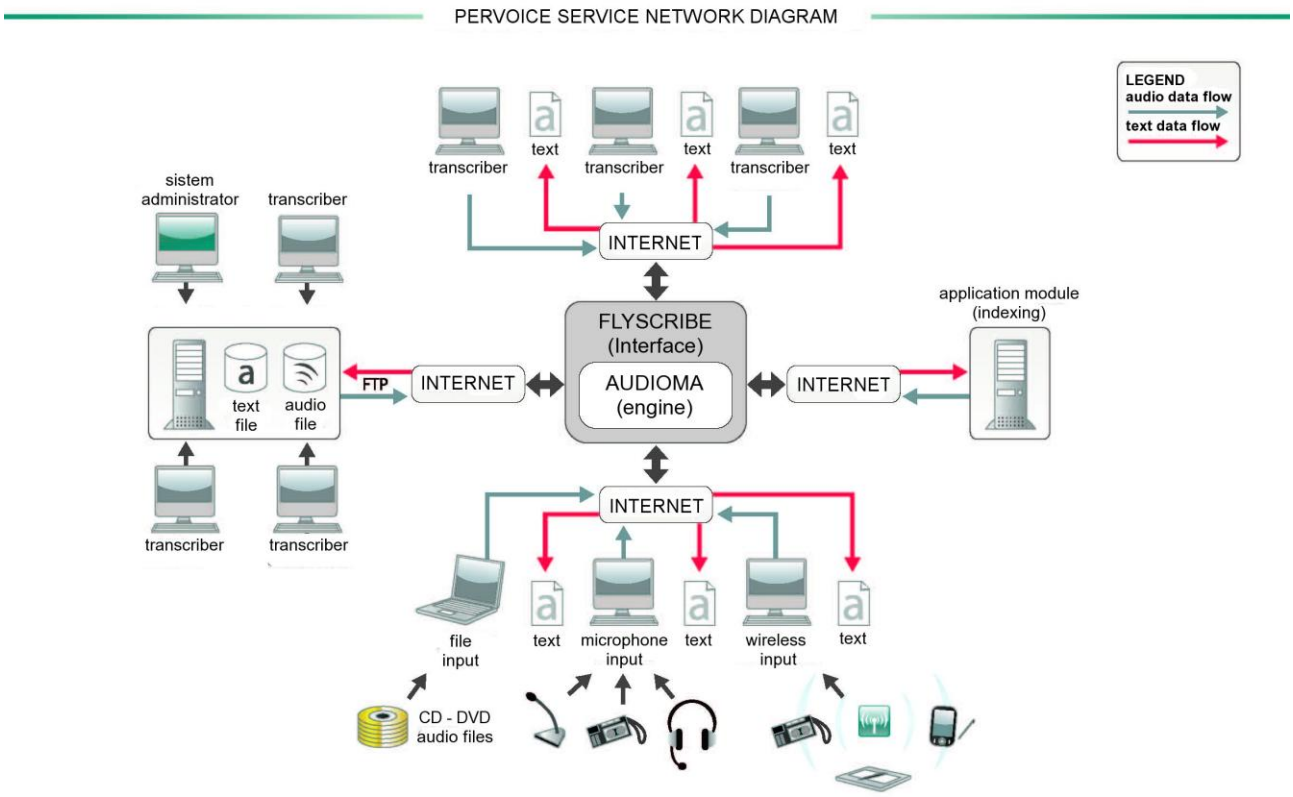
3. The FlyScribe service

Following improvements to the technology and interest on the part of the marketplace, FBK and investors decided to create a company to bring the transcription system to market. And in March 2007, PerVoice SpA was born (www.pervoice.it).

PerVoice is developing a number of applications to improve the use of FBK voice recognition and automated transcription:

- Legislative assemblies (senate, house, ministries, local and regional government, etc.).
    - You can create, with little lag time, draft transcriptions that are similar to documents produced by stenographers.
    - You can create a draft transcription that noticeably improves the productivity and work quality of transcribers. Calculations show that the time it takes to transcribe is cut in half if speech is pre-transcribed with automated recognition.
    - You can make audio/video archives accessible by indexing their automated transcriptions. At this point it is possible to do keyword searches on audio/video files using engines like Google. Using speech, you can correlate and temporally align transcriptions with AV files.
- Court reporting.
    - You can use uncorrected transcriptions to navigate audio recordings of court proceedings. (This opens up the possibility of saving millions of Euros per year.)
    - You can rapidly transcribe interrogations and do searches by keyword.
- Conventions.
    - You can subtitle convention presentations to make them more usable, in particular for non-native speakers, people who have hearing problems, or simply because the hall is noisy.
    - You can use subtitling to do text searches to find specific concepts in a presentation.
- Broadcasting.
    - You can subtitle. This is currently post-production, but within a year it will be “live”.
    - You can automatically index content. Research engines allow keyword searches and semantics-based indexing.
    - You can search for keywords to help with film editing.
- Call centers.
    - You can classify phone calls using a “business intelligence” solution. For example, it might be interesting to know what problems occur most often, what clients are asking and what they say about your competitors, and if they have seen your advertisements. To get this result, you take the complete transcription and create statistics using a semantics engine.
- Universities and education.
    - You can transcribe and classify university courses to allow searches for high level scientific content.

The FlyScribe portal is the heart of the application. Anyone with an Internet connection and computer terminal can use their credentials to access FlyScribe, upload audio files, and download transcription files. Here is a diagram of the FlyScribe network.



The main kinds of remote connections are:

- A transcription professional submits files directly to the portal (see figure, top).
- An application installed on a client machine submits files for indexing, for example, automated transcriptions of television programs (see figure, right).
- A client application organizes audio files for upload and automates download of the transcription. This is appropriate for transcription services companies with a high volume of traffic in need of an FTP/FTPS connection (see figure, left).
- A user who wants to upload an audio file recorded with a microphone or dictaphone (see figure, bottom).

The upload page of the portal provides a number of options:

- Language model. PerVoice currently has more than 20 models to choose from.
- Acoustic model. There are 3 models which handle: general microphone, telephone, and radio-television.
- Localization. You can use words specific to a region where an event takes place. It includes things like common family names, place names, and toponomy. To transcribe the meetings of Venice city council, you want to include words specific to the region around Venice.

- Personalization. Along with the audio file, you can upload a list of words that were pronounced during the recording. For example, a list could include the names of speakers at a convention or participants at a court proceeding that are likely to be missing from the general language model
- Punctuation. You can ask FlyScribe to insert punctuation automatically.
- Execution priority.

The result of the automated transcription process is a file formatted in XML that includes automatically identified information about the speakers and changes of speaker, but most importantly the recognized text and the moment – in milliseconds – in which it was pronounced. This means that synchronization of text and spoken words is native to the platform, allowing for subtitling of audio/video content and assisted editing for transcribers.

In fact, FlyScribe has an editing tool that integrates with Microsoft® Word: Users navigate the audio file with a footswitch while the cursor moves along to highlight the text of the word being pronounced. In addition, the user can control the speed of audio playback, automatically insert the changes in speaker, and use other functions that speed up the revision process.

The advent of FlyScribe is changing the transcription process. New courses are emerging to train people in the areas of audio recording and document revision. On the one hand, there is a need to train technicians to record high quality audio so as to facilitate automatic transcription. On the other hand, there is the need to train people in the efficient use of synchronized audio-text revision tools. The goal is to create specific competences in transcription revision while reducing focus on typing speed. In the near future, a role will emerge for people to personalize language models to optimize the results of automatic transcriptions.

One area of application that is particularly advanced is for Italian Courts. In this case, a multichannel recording system present in each courtroom makes it possible not only to precisely identify the speaker (judge, witness, accused, prosecutor, attorney, etc.) , but also to reconstruct in the correct sequence the phrases spoken by each speaker even if two or more persons talk at the same time.

The most important technological aspect is the transcription of overlapping speech, a prohibitive task for any recognition system. Today, more than 5% of criminal trials in Italian courts are automatically transcribed with this system. These transcriptions are later revised by experts with a 50% time savings with respect to manual transcription, as well as improved quality.

Use of FlyScribe is expanding rapidly in a number of areas. At the time of writing (May 2009), FlyScribe transcribes more than 1,000 hours of speech each month after just one year in operation. The goal is to transcriber over 10,000 hours per month within 12 months.

FlyScribe stands out from other automated transcription services on the market by adapting to the field of application. Some practical examples are:

- Highly personalizable

- Wide choice of language models
- Strong security implementation
- Redundant hardware for continuous uptime

From the point of view of transcription service evolution, the results are:

- Reduced human work in general, and use of less skilled workers
- Rapid turn-around time for automated transcriptions
- 24x7 service availability
- No limits on quantity
- Automated punctuation
- Innovative tools for editing and reviewing transcriptions

The FlyScribe solution is also available as a stand-alone software installed on customer hardware. This means having control where privacy is a necessity.

On the other hand, use of a service-based transcription software is a novel approach that FlyScribe brings to the international voice recognition market.

4. Conclusion and future work

Speech recognition technology has taken many years to reach an operational stage. Regarding FBK technology, the most important milestones are:

- 1993 Dictation for medical records.
- 1995 First presentation at Intersteno of the FBK voice transcription project.
- 2002 First version of Video Indexing for RAI.
- 2005 Second presentation at Intersteno Vienna: transcription of European Council debates.
- 2007 Founding of PerVoice.
- 2009 Online transcription of criminal court proceedings. Transcription and indexing of conference materials. Quality monitoring applications for Call Centers.

It is now possible to automatically transcribe human speech with high precision at low cost. Planned developments include:

- Live subtitling: Streaming transcription will be available for subtitling of television programs, conventions, university lectures, and other events in real time. This advancement will improve accessibility of content for persons with hearing difficulties as well as non-native speakers.
- Business intelligence applications for call centers: This is a well established discipline in the USA. It provides data about customers' questions about a product or service. It functions by semantic interpretation of transcriptions of phone calls.
- Media monitoring: This is an automated system for checking on whether specific words are used in radio and television broadcasts for reporting and research purposes.

5. References.

F. Brugnara, M. Cettolo, M. Federico and D. Giuliani. Advances in Automatic Transcription of Italian Broadcast News. *In Proceedings of ICSLP*, Beijing, China, October, 2000, pp. 660-663.

N. Bertoldi, F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani. From broadcast news to spontaneous dialogue transcription: Portability issues. *In Proceedings of ICASSP*, Salt Lake City, UT, 2001, vol. 1, pp. 37-40.

F. Brugnara, M. Cettolo M. Federico and D. Giuliani. Issues in Automatic Transcription of Historical Audio Data*. In Proceedings of ICSLP*, Denver, CO, September, 2002, pp. 1441-1444.

D. Giuliani and M. Gerosa and F. Brugnara. Speaker Normalization through Constrained MLLR Based Transforms. *In Proceedings of ICSLP,* Jeju Island, Korea, 2004, pp. 2893-2897.

N. Bertoldi, F. Brugnara, M.Cettolo, M. Federico, D. Giuliani E. Leeuwis and V. Sandrini. The ITC-irst News on Demand Platform. *In Proceedings of the European Conference on Information Retrieval Research*, Pisa, Italy, April 2003.

6. Contacts.

Authors:

- Gianni Lazzari: lazzari@fbk.eu
- Alessandro Tescari: alessandro.tescari@pervoice.it
- Brian Martin: brian.martin@pervoice.it
- Fabio Brugnara brugnara@fbk.eu

Web Sites:

www.pervoice.it/en
www.fbk.eu
www.flyscribe.it