New Approaches to Improving the Quality of Media Accessibility Services on HbbTV

Figure 1: NERstar Logo

Andreas Jesina
VerbaVoice GmbH
Munich, 81677, Germany
a.jesina@verbavoice.de

Juan Martínez Pérez
TransMedia Catalonia
Bern, 3007, Switzerland
juan.martinez@speedchill.com

Robin Nachtrab-Ribback
VerbaVoice GmbH
Munich, 81677, Germany
r.nachtrab@verbavoice.de

Marko Nalis
VerbaVoice GmbH
Munich, 81677, Germany
m.nalis@verbavoice.de

Gion Linder
SWISS TXT
Biel, 2501, Switzerland
gion.linder@swisstxt.ch

Pablo Romero-Fresco
University of Roehampton
London, SW15 5PU, UK
p.romero-fresco@roehampton.ac.uk

Abstract
This paper describes the efforts carried out by the authors to improve the quality of currently existing media accessibility services within the Interpreter Telepresence System by VerbaVoice, focusing on subtitle quality using the NERstar Editor as well as accessibility of TV programming for hearing impaired people via the VerbaVoice HbbTV App.

Specifically, it describes algorithms currently used to improve subtitle readability and synchronicity, as well as issues and solution approaches towards synchronizing source material, sign language video and subtitles within a live broadcasting context for HbbTV.

Author Keywords
NERstar; NER model; HbbTV; inclusion; accessibility services, live subtitling, respeaking, speech recognition, video

ACM Classification Keywords
H.5.m [Information Interfaces and Presentation (e.g., HCI)]: Miscellaneous

Introduction

Creation of media accessibility services is becoming a more and more important topic. With countries pushing to fulfill the requirements set by the EU convention on the rights of persons with disabilities ([1]), providing additional accessibility content – mostly subtitles – for live and prerecorded broadcasts is slowly being prioritized by broadcasters around Europe. Especially in a live context, it can be challenging to provide captioning in an adequate quality and with enough synchronicity to the original content that the understanding of the broadcast does not suffer. There are a couple of different approaches to this, for example:

- velotyping: someone is directly typing what is broadcast via the original audio into subtitles;
- respeaking: someone is listening to the broadcast and dictating with speaker-dependent speech recognition software in order to create subtitles in real-time;
- automatic speech recognition (ASR) directly using the source material audio.

In a TV environment, respeaking is the most widely used method. It has become an economically viable solution for broadcasters wishing to increase their production of subtitles. ASR has not yeen been proven mature enough for being applied in real-time on a large scale.

Within VerbaVoice's system, subtitles are mostly being created using the first and second method and then distributed via the Interpreter Telepresence System, of which the next sections contain a short overview.

Interpreter Telepresence System

The Interpreter Telepresence System (ITS) is a technical solution developed by VerbaVoice which is widely used throughout Europe. In 2014, for example, it was used to deliver roughly 300,000 hours of live transcription, particularly in the field of education, but also for live events, parliamentary activities, international soccer matches and television programming. Both the NERstar Editor and the VerbaVoice HbbTV App are integral components of the ITS.

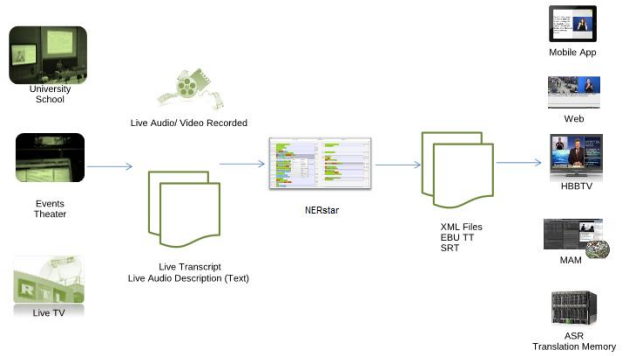

Figure 2: High level diagram

ITS System Architecture

The core system is designed as a "true" cloud system. It is set up across multiple data centers, each of which features a switched connection to the internet and is composed of three system groups:

1. Workflow system: This system is used to book and coordinate all assignments. Communication with the participants is based on different job steps defined for a

particular area. Moreover, it deals with the management of the language profiles.

2. Telepresence system: This system provides the speech-to-text reporter (STTR) with real-time feeds for audio, phone calls, video, PowerPoint presentations and other background information. It also handles the distribution of all real-time transcripts that are created. Another big advantage of ITS lies in the fact that STTRs located around the world can be assigned specific tasks, making it no longer necessary for them to be on-site. Another benefit: The system enables transparent interpreter switching, i.e., different STTRs can rotate during a broadcast.
3. Respeaking assistance systems: These systems provide functionalities that simplify the respeaker's job. As in the example of driving a car, users must still operate the system themselves, but the assistance systems in place increase responsiveness and quality. The diverse range of support systems includes glossaries for specific subject areas and the availability of automatic transcriptions. These features lead to an improvement in quality and also make the service more affordable for customers.

Furthermore, numerous adapter systems with access to the centralized system allow for the system to be used in various situations; enabling both unidirectional and bidirectional access. Some examples of adapter systems are fully accessible web players, mobile second-screen apps, live subtitling inserters, HbbTV apps, smart TV apps,, subtitling editors etc. One of these adapter systems is the NERstar Editor, a software for subtitle quality control.

NERstar Editor

The NERstar Editor is a quality measurement tool for subtitles, applying the NER formula (see Fig. 4, [2]) and providing many other statistical data to supply the user with measurable and comparable figures to evaluate and compare subtitles. It also contains functionalities to automatically adapt and improve existing subtitles in terms of, among other things, readability (speed) and synchronicity (delay).

The project has been in development for over a year and is still actively being developed further. Almost all running code is written in Scala, which has proved to be suitable for both small and large scale projects over the past years.

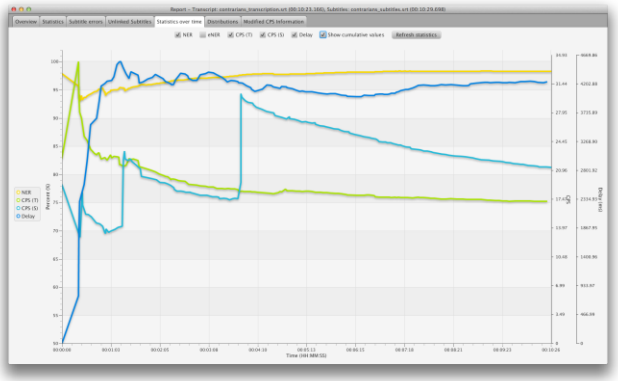

Figure 3: Some of the statistical data generated by NERstar for a specific project. Different measurements such as the NER value, reading speed and delay are plotted against the running time of the project to be able to directly link drops in quality with specific occurrences of other values, such as very fast reading speed.

$$\text{Accuracy} = \frac{N - E - R}{N} \times 100$$

Figure 4: Description of the NER formula

N: Number of words in the produced text, including commands (punctuation marks, speaker identification, etc.)

E: Edition errors, usually caused by the strategies applied by the STTR. They are calculated by comparing the produced text and the original text and may be classified as serious, normal or minor, scoring 0.25, 0.5 and 1 respectively.

R: Recognition errors, which are usually misrecognitions caused by mispronunciations or mishearing. They may be insertions, deletions or substitutions, and are calculated by comparing the produced text and the original text. They may be classified as serious, normal or minor, scoring 0.25, 0.5 and 1 respectively.

The integration of specific qualitative differences based on viewers' preferences when defining the standardisation of types of errors and their respective quantitative weightings mentioned above is the reason why the NER model delivers scores that are internationally comparable, auditable and relatively consistent, with an average discrepancy around 0.1% ([4]).

Functionality overview

The NERstar Editor can be used for:

a) applying the NER model to assess the accuracy of live speech-to-text services in general and live subtitling services on TV in particular, thus obtaining not only an overall idea of quality in terms of accuracy, but also of aspects which need to be improved.

b) generating statistical data on existing subtitles to assess different aspects such as readability, synchronicity with the original content, missing information when compared to the original content, etc.

c) running optimization algorithms to improve certain aspects of the produced subtitles. These algorithms are outlined further in the next section.

Optimization functionalities

In order to improve the quality of produced subtitles, the NERstar Editor contains multiple algorithms adapting certain aspects of any given subtitle file, for example:

- Timecode adaptations to enhance the readability of given subtitles, particularly of those whose speed exceeds critical thresholds that would pose problems in terms of readability and comprehension. This algorithm tries to elongate standing time of subtitles by moving start and end points of subtitles, adhering to certain maximum and minimum values. For example, according to research from the US, a subtitle's starting or end point can be moved forwards or backwards respectively by up to 600 milliseconds – or 15 frames in a 25 fps video – to improve standing time without losing perceived synchronicity with the original content ([3]). Furthermore, pauses between subtitles can also be used to increase standing time and thus improve readability further by making it easier to actually follow the content of the captioned program.

- Subtitle shifting to reduce delay. This algorithm uses metadata or external information on a part of the subtitles to reduce delay in relation to the original content and to improve synchronicity. Usually, live subtitles have a larger delay than preproduced subtitles which are simply cued manually. This information can be used to improve the synchronicity of live subtitles, for example for a re-run of the same broadcast. If no external information is available, this can also be done by hand on specific sections of a subtitle file.

While these algorithms are currently being used in a post-processing context, work is being done in order to be able to use them in a real-time context to improve the live delivery of subtitles in all front-end components of the ITS, such as the VerbaVoice HbbTV App, which the next sections will focus on.

HbbTV

The technology that enables Smart TVs is not only incorporated into television sets, but also into devices such as set-top boxes, Blu-ray players, game consoles, hotel television systems and other companion devices. These devices allow viewers to search and find videos, movies, photos and other content on the web, on a local cable or satellite TV channel, or on a local storage drive.

Thanks to HbbTV (Hybrid broadcast broadband TV), we can see a first level of convergence with the arrival of

interactive services, using broadband networks, on the TV display in the living room allowing the interaction between Broadcast TV channels and complementary data such as Electronic Program Guides, Teletext services, Interactive Quizzes, etc.

Hybrid TV services were deployed with the first HbbTV services launched by German broadcasters in December 2009. HbbTV was formally standardized in 2010. It uses a double connection, such as broadcast cable, satellite or terrestrial coupled with a broadband (Internet) connection. The broadcaster sends web links with pictures, and the end user can access catch-up services, enhanced program guides and other interactive content related to the broadcast program.

HbbTV puts a lot of constraints on the terminal and its remote control unit. Applications can run on connected TV sets, set top boxes or personal video recorders, which are all TV oriented components. Access to web applications is limited and the browser uses mainly the address provided in the stream.

HbbTV is not yet able to offer a means for a real time and effective interworking of different streams as they include no mechanism able to ensure the synchronization of content components sourced from servers anywhere in the cloud and delivered to the end user over networks using different transport protocols and exhibiting different end-to-end delays.

One of the challenges for European broadcasters today is the multi-platform delivery of TV content on broadcast, IP and hybrid delivery platforms (i.e. both big screen, hybrid and two-screen solutions).

The introduction of the HbbTV standard helps to overcome the fragmentation of the connected TV market. HbbTV provides a straightforward specification on how to combine broadcast and broadband content plus interactive applications. TV content can be enhanced with additional synchronized services in a customized manner.

Figure 5: The VerbaVoice HbbTV App. Sign Language interpreter and subtitles in different languages can be added to provide an accessible TV experience

For access services, HbbTV has opened an entirely new opportunity for users who may choose an access service delivered via their IP connection, which then seamlessly integrates with the regular TV program. Elderly people and people with various disabilities rely on subtitles, audio description or sign language. In addition, Web-based and HbbTV-based solutions offer the potential for customizable services enabling the user to adapt these to his or her special needs and abilities or preferences.

The VerbaVoice HbbTV App

In order to provide accessible television for people with impairments, VerbaVoice has taken on the challenge to develop the world's first HbbTV app for deaf and hard of hearing people (see fig.5). The main benefit of the application is that it offers a picture in picture sign language interpreter video on top of any broadcast or ip-stream as well as subtitles. Thus it becomes possible to add customizable accessibility services such as sign language interpreters and subtitles to live broadcasts independently of the TV Station's workflow.

The interpreter can be hidden or shown according to the user's preference. The first version is currently being tested and works on a set top box with two video decoders and DVB-C support provided by TARA Systems. The service offers customizable subtitles and automatic translation of subtitles into the English language. There are two main use cases. The VerbaVoice app automatically detects the running broadcast channel and displays the appropriate accessibility services. In the current implementation the broadcast is manually transcribed by a speech-to-text reporter. In the near future we can also expect the increased use of (semi-)automatic speech recognition. The audience can either watch a program with accessibility services (sign language interpreter and/or subtitles) or you can watch a live stream from the Internet with the same accessibility options. Furthermore the services can be customized. You can change the size and position of the sign language interpreter as well as the position, color and background of the subtitle fonts. Like that, when watching a football game and the score is placed in the top right corner of the TV image, the video of the interpreter can be moved to the left corner. If the user prefers to use one of the accessibility options only, e.g. the subtitles, the sign language interpreter can easily be removed. When watching a news channel users can decide to move the subtitle to the top of the screen in order to read the news ticker on the bottom. However, the second video is only available on set top boxes and Smart TVs with two video decoders. Since the service is meant to run independently from the TV Stations, some interesting challenges arise.

The main challenging use case is for live television. The accessibility services cannot be prepared in advance so it is necessary to reduce latency in every aspect with the goal of achieving the best synchronization. Unfortunately, a perfect synchronization is impossible at the moment.

On the one hand, subtitle synchronization is not a huge issue since the latency is within the microsecond range. The only limiting factor is the speed of the human speech-to-text reporter. On the other hand, major challenges need to be resolved for the sign language interpreter internet stream. The sign language interpreter video is delivered through broadband but the TV program itself is delivered through DVB-C. The process lies outside of the TV station's workflow and, of course, we are looking at live events here. The interpreter receives the broadcast at the same time as the audience. He can thus only start interpreting once the broadcast is received by all viewers and everybody has already seen the actual footage. With technical latencies and the time it takes to create and re-stream the interpreted footage this would result in a delay of a few seconds. Due to technical restrictions and the HbbTV standard it is obligatory to use MPEG-TS HTTP with chunked encoding for streaming. This results in a

minimum latency of 2-3 seconds for the sign language video. The users would receive the interpretation a few seconds after the actual content.

There are various possible solutions. In the future, it might be allowed to use rtsp which would significantly reduce latency and improve the synchronization. Another solution would be to use the difference in latency between DVB-C and DVB-S to our advantage. According to "heise.de", during the FIFA World Cup 2014 [5] DVB-C had a 6 second delay compared to DVB-S in standard quality. High Definition is a little (1.5s) slower.

In order to achieve good synchronization the sign language interpreter would need to watch the live television event via DVB-S. He would then have 6 seconds to translate the broadcast and stream it to the users. With an approximate time to stream the footage to the users of 2-3 seconds he would effectively have 3 seconds to translate the spoken content into sign language. According to [5](see fig.6) 46.3 % of households in Germany are using cable and 46.2% of households are using satellite. With the approach described above, latency-free sign language interpretation could be provided to all of the DVB-C users in Germany.

In order to offer latency-free subtitles as well, the same approach could be valid. The speech-to-text reporters could watch the event via DVB-S. They subsequently produce the subtitles with their personal typing speed as the only limit. This process is much faster than the video sign language counterpart and takes no longer than 1-2 seconds. Adding 100ms for distribution, the subtitles can be produced before the actual content in

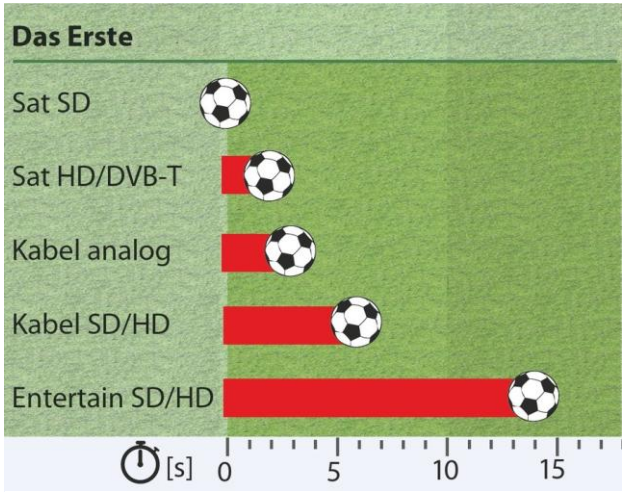

Figure 6: Differences in the delay for different distribution ways of the German TV Channel "Das Erste" [5]

the broadcast. This means that the subtitles have to be delayed. The remaining time can be calculated and the delay of the subtitles can be adjusted accordingly. For the previous example this means: If the speech-to-text reporter needs 2 seconds to produce subtitles, considering the 6 second difference between DVB-S and DVB-C and 100 ms to deliver the text information, distribution of subtitles would have to be delayed by 3.9 seconds.

Unfortunately this is only part of the solution. Due to different internet and cable providers and different technologies and services in use (subtitles vs video streaming), there are inconsistent latencies and delays to deal with – for each user individually. Furthermore, the above approach doesn't work for satellite users.

In order to automate and synchronize the different processes, a central time provider or server with high availability could be used. The subtitles and the video would be enriched with timestamps and an algorithm could calculate the delay per user based on a personal round trip delay time which could be calculated by sending a short audio signal from the set top box to the VerbaVoice servers. They can then be synchronized with the central clock and delivered with the appropriate delay.

The other main use case are IP live streams with added sign language. In this case the issues of the interpreter latency don't come up since both the main live stream and the picture in picture video have approximately the same latency. Being able to control the live stream, one could delay it for the users but deliver it with maximum speed to the interpreter. The problem that arises in this case is that the subtitles might be delivered before the actual content. This can be solved by adding an appropriate delay to the subtitles.

We can conclude that there are still many open questions about the best solution, regarding the Verbavoice HbbTV app synchronization but also some promising approaches and solutions.

Final Remarks

In general, efforts to synchronize live broadcast material with additional content being delivered over IP will always face issues since there are no internet connections always working as expected , and individual users might have different experiences. However, we are optimistic that the outlined approaches to improving quality as well as synchronicity of accessibility services will lead to a better experience for all users.

Acknowledgements

This research is supported by the Catalan Government funds 2014SGR027, and partially funded by VerbaVoice, SWISS TXT and the European project HBB4ALL #621014.

Sources

[1] : Persons with disabilites – Employment, Social Affairs & Inclusion (Retrieved April 23rd, 2015) http://ec.europa.eu/social/main.jsp?catId=1137&langId=en

[2]: Accuracy Rate in Live Subtitling – the NER model (Retrieved April 24th, 2015) https://roehampton.openrepository.com/roehampton/bitstream/10142/141892/1/NER-English.pdf

[3]: DCMP Captioning Key 2011 (Retrieved April 20th, 2015) http://www.dcmp.org/captioningkey/captioning-key.pdf Page 20 - Synchronization

[4]: OFCOM – Measuring live subtitling quality (Retrieved April 24th, 2015) http://stakeholders.ofcom.org.uk/binaries/consultations/subtitling/statement/sampling-report.pdf

[5]: Anpfiff – Technik für eine ungetrübte Fußball-WM (Retrieved April 24th, 2015) http://www.heise.de/ct/ausgabe/2014-13-Technik-fuer-eine-ungetruebte-Fussball-WM-2221818.html