Partner-driven Innovation in Automated Speech Translation

George Bara

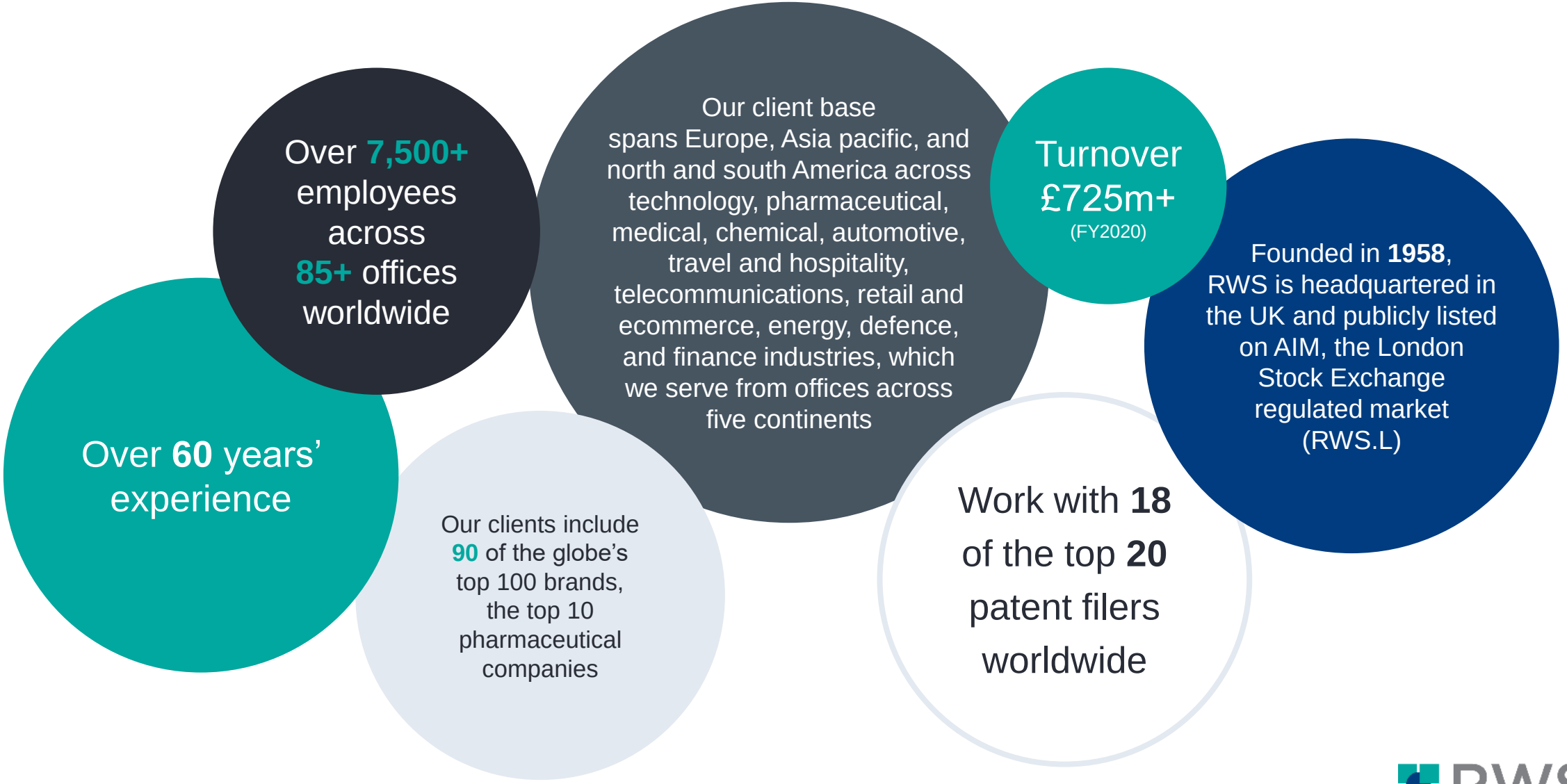
VP, Strategic Partnerships

24 March 2022

Agenda

- About RWS
- European Parliament live speech translation for plenary sessions – project.
- Challenges in combining ASR and NMT.
- Evaluation metrics for automated speech translation.
- Future development areas.
- Q&A.

The world’s largest Language Services and Technology group



Partnership- driven innovation: RWS and Cedat85

RWS + CEDAT 85

Global leader in translation services and technology and **Neural Machine Translation** for **Public Sector** & **Enterprise**.

Language Weaver

SDL* | LANGUAGE WEAVER

Award-winning innovative technology company developing and implementing the world's most powerful and innovative **speech-to-text technology solutions** for private and public sector clients.

Mentioned 4 years in a row in Gartner's reports for innovation in Speech-to-text applications, chosen by 500 Local and central government organizations such as the Italian Chamber of Deputies, the Italian Ministry of Interiors, the British Library, plus hundreds of customers in different sectors Finance, Telco, utilities, education.

RWS

RWS and Cedat85 solution for the European Parliament

European Parliament plenary sessions are transcribed and translated by **RWS** together with **Cedat85** in 24 EU languages, in **real-time**, to allow people with impaired hearing to participate in the sessions and provide an inclusive experience.

Most prominent Linguistic AI project ever tendered by a European Union public institution.

- Consortium ranked 1st and awarded phase 1 of project in Oct 2020, outranking Microsoft and Translated.com.
- 1st phase of the project (10 languages) finalized June 15th 2021.
- **2nd & 3rd phases:** 19 languages delivered by June 2022, with 5 other languages to follow.

Winning Recipe

Customization of ASR engines using EP datasets at phonetic & language model

Adaptation of MT engines to specific EU terminology (IATE)

Deep integration between ASR & MT

Solution Architecture

100% Automated
No human intervention

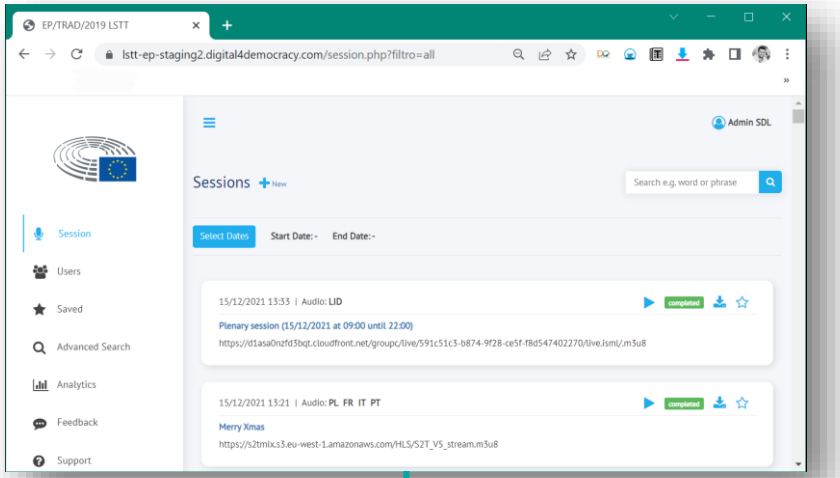
HLS – HTTP Live Streaming

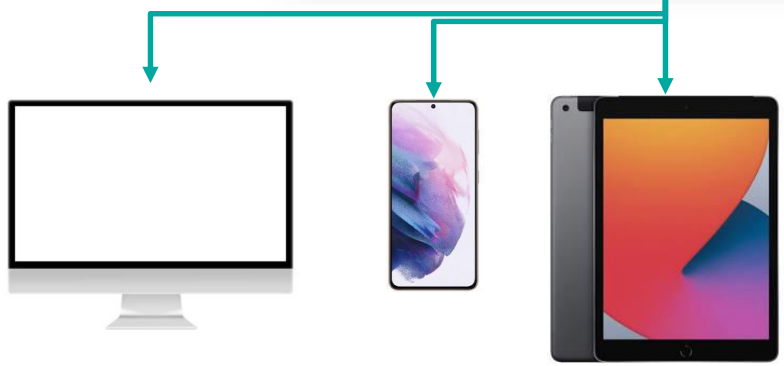
Audio language identification

Speech Transcription

Automated Translation

- Session Management
- User Management
- Advanced word or phrase search
- Analytics
- Feedback





Challenges in combining ASR and NMT

4 Key challenges when combining ASR & MT

1. Speech Translation = Automatic Speech Recognition **+** Automatic Translation **API pipeline integration**.

2. Maintaining Fluency and Meaning

ASR = Word by word transcription of audio.

MT = “Decoding” of source language text into target language text. Not word by word, output can have multiple output variants.

3. Evaluation Metrics

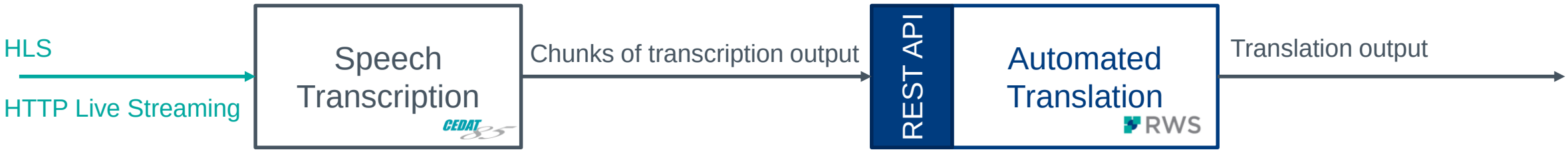
ASR & MT have **different evaluation metrics**:

- Word Error Rate (WER) from ASR.
- BLEU, NIST-score, TER and Usability Likert Scale.

4. User Experience

- Latency.
- Readability of live transcription.

API Integration of ASR & MT

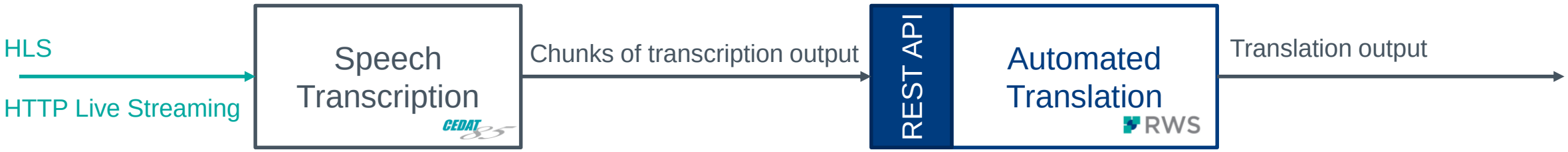


- Live audio stream is split into process-able “chunks”.
- Input audio is segmented by speech & non-speech classes (speech vs silence).
- Segmentation on temporal context, to ensure output fluency and latency requirements.

- Output from ASR is sent to automated translation into all target languages (23).
- Using blocking (synchronous) parallel API calls for the input text.
- Optimized for quality and performance.

RWS

Maintaining Fluency & Meaning



Meine Fraktion	My Group
Meine Fraktion ist bereits	My group is already
Meine Fraktion ist bereits eine Sondersitzung	My group is already a special session
Meine Fraktion ist bereits eine Sondersitzung des Parlaments zu	My group is already in a special session of Parliament
Meine Fraktion ist bereits eine Sondersitzung des Parlaments zu ermöglichen, um einen	My group is already allowing for a special session of Parliament to get one
Meine Fraktion ist bereits eine Sondersitzung des Parlaments zu ermöglichen, um einen harten	My group is already allowing a special session of Parliament to make a tough one
Meine Fraktion ist bereits eine Sondersitzung des Parlaments zu ermöglichen, um einen harten ökonomischen Brexit zum Jahresende zu	My group is already in a position to allow a special session of Parliament in order to achieve a tough economic Brexit by the end of the year
Meine Fraktion ist bereits eine Sondersitzung des Parlaments zu ermöglichen, um einen harten ökonomischen Brexit zum Jahresende zu vermeiden.	My group is already in a position to allow a special session of Parliament to be held in order to avoid a tough economic Brexit at the end of the year.

Proper segmentation and punctuation is very important or ASR output is critical to MT output quality.

Punctuation on ASR output is challenging for live streams.

ASR language model & post-segmentation can change the final output of a phrase

RWS

12

Evaluation Metrics

Speech Transcription

ASR commonly uses **WER – word error rate as an evaluation metric**.
To compute the WER two files must be given as input:
a) a text file containing the correct transcription of an audio ‘x’; this is the reference text or the so called ‘golden text’, it should be produced manually from a native speaker;
b) a text file containing the automatic transcription of the same audio ‘x’.

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

S – substitutions
D – deletions
I – insertions
N – number of words

Automated Translation

MT commonly uses **BLEU (Bilingual Evaluation Understudy) as an evaluation metric**.
But BLEU is best to be used to compare different MT engines, while the golden evaluation standard is **Human Evaluation based on Usability**:

- **5** - Sentence is perfectly clear and intelligible. Not necessarily a perfect translation, but grammatically correct, with all information accurately transferred.
- **4** - Sentence is generally clear and intelligible. Acceptable; not perfect, but understandable and captures most of the source meaning.
- **3** - Sentence contains grammatical errors and/or poor word choices; with effort, some, although not all, meaning is able to be gleaned from the source.
- **2** - Contains a few key words, but little of the source meaning is expressed in the translation.
- **1** - Unacceptable; incomprehensible with little or no information transferred accurately. None of the meaning expressed in the source is expressed in the translation.

13

Evaluation Metrics - Combined

The European Parliament opted for a combination of Automatic and Manual Evaluation process consisting of:

ASR: **Word Error Rate** (WER < 20%) and **Latency** (2s/words, 8s/segment) + **Human Evaluation**.
MT: **Human Evaluation**.

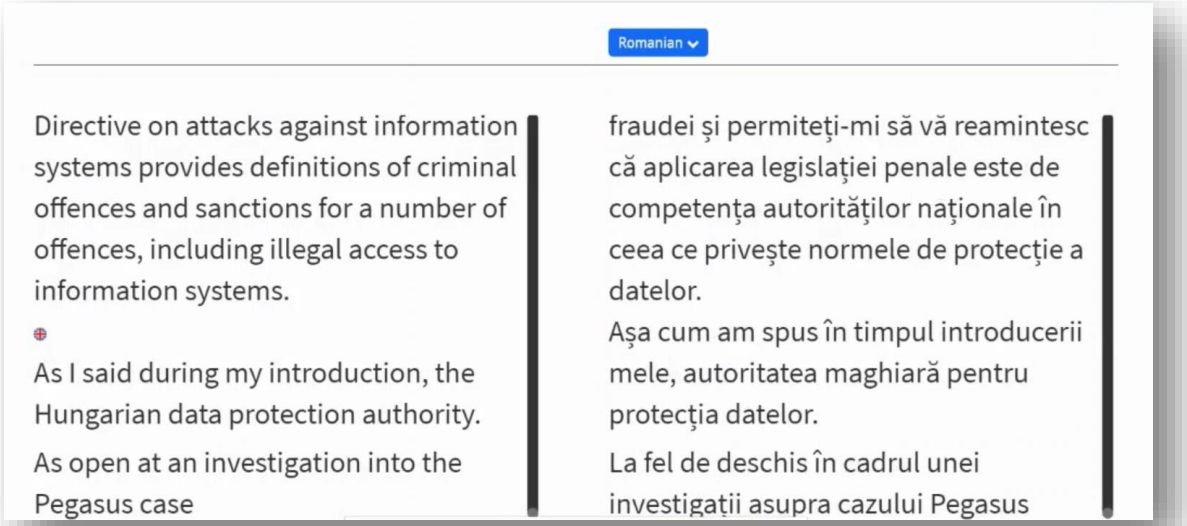
The Evaluation will produce two scores per language, one for ASR and another for MT :
1=Good
2=Usable
3=Unusable

1. Coherence: The transcript makes sense and a message is identifiable and understandable in it.
2. Omission: The transcript lacks part(s) of the original speech, which leads to changes in or deprivation of the intended meaning.
3. Substitution: The transcript contains mistranscribed words or phrases, which leads to changes in the intended meaning or inappropriate comical effects.
4. Clarity: The transcript is intelligible in terms of terminology (inclusion of unknown words, proper nouns, etc.)
5. Readability: The transcript is comfortable to read in terms of grammatical consistency, punctuation, and removal of disfluencies.

1. Coherence: The MT output makes sense and a message is identifiable and understandable in it.
2. Clarity: The MT output is intelligible in terms of terminology (inclusion of unknown words, proper nouns, etc.).
3. Readability: The MT output is comfortable to read in terms of grammatical consistency, punctuation, spelling and removal of disfluencies.
4. Accuracy: The amount of meaning expressed in MT output appropriately mirrors the amount of content expressed in the reference (the human interpretation as delivered in the relevant plenary session).

User Experience

VS



- Movie subtitles are not generated real-time.
- Movies subtitles are translated from a written script.
- The subtitle is displayed at the beginning of a video sequence, and do not change.

- Automated transcript generated real-time, with appended output built on the screen.
- ASR + MT incur delays from source audio stream.
- Translation output changes as the ASR output gets to its “final” segmented & punctuated form.

Fragmented Conversations

Speeches read from scripts	Plenary sessions, scripted presentations, political speeches.
Single- speaker presentations	Product release presentations, educational & training sessions.
Disciplined multi- speaker sessions	Webinars, quarterly business reviews, workshops.
Fragmented Conversations	Dynamic multi-speaker sessions with no scripts: real day-to-day web calls. Overlapped conversations, bad audio quality.

16

Future development areas

Expand use of the speech translation platform within the EP: online multilingual events and non-plenary sessions.

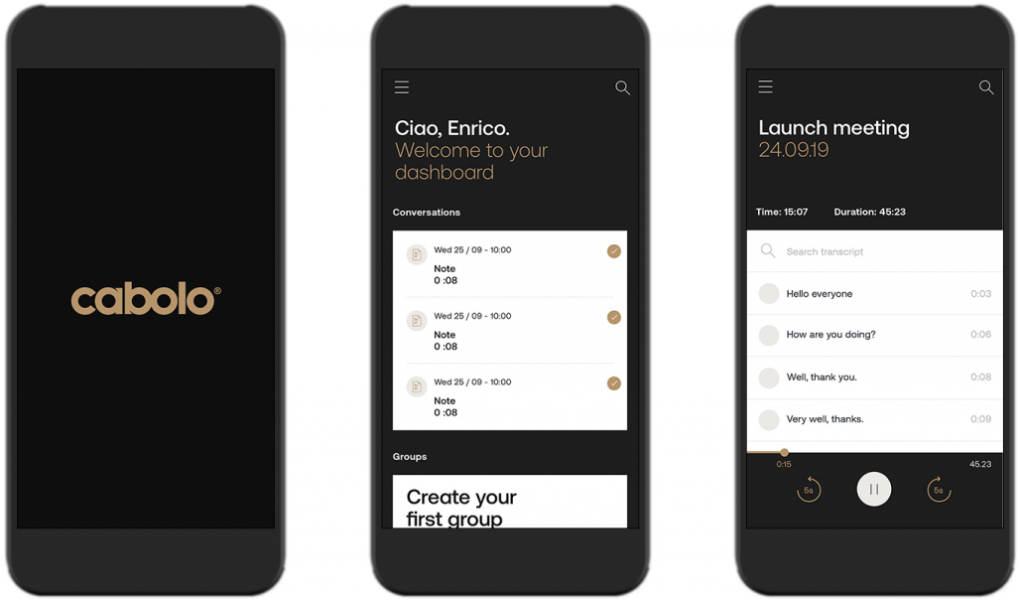
Replicate across other commercial use-cases: remote learning, webinars & online events, web conferencing.

Enhance the speech translation capability with Text Analytics for automated insights extraction.

Edge cloud & device deployment capability.

Introducing Cabolo

A portable standalone device that works without any internet connection, provides utmost data security, records and automatically transcribes & translates any meeting and interview content.



www.rws.com