FAST

FULLY AUTOMATED SPEECH TO TEXT

A human evaluation approach

What we are talking about today

1. Why is human evaluation a crucial part of developing and improving speech to text technology?

2. How is accuracy measured in FAST?

3. What are the challenges associated with human evaluation?

THE PROJECT

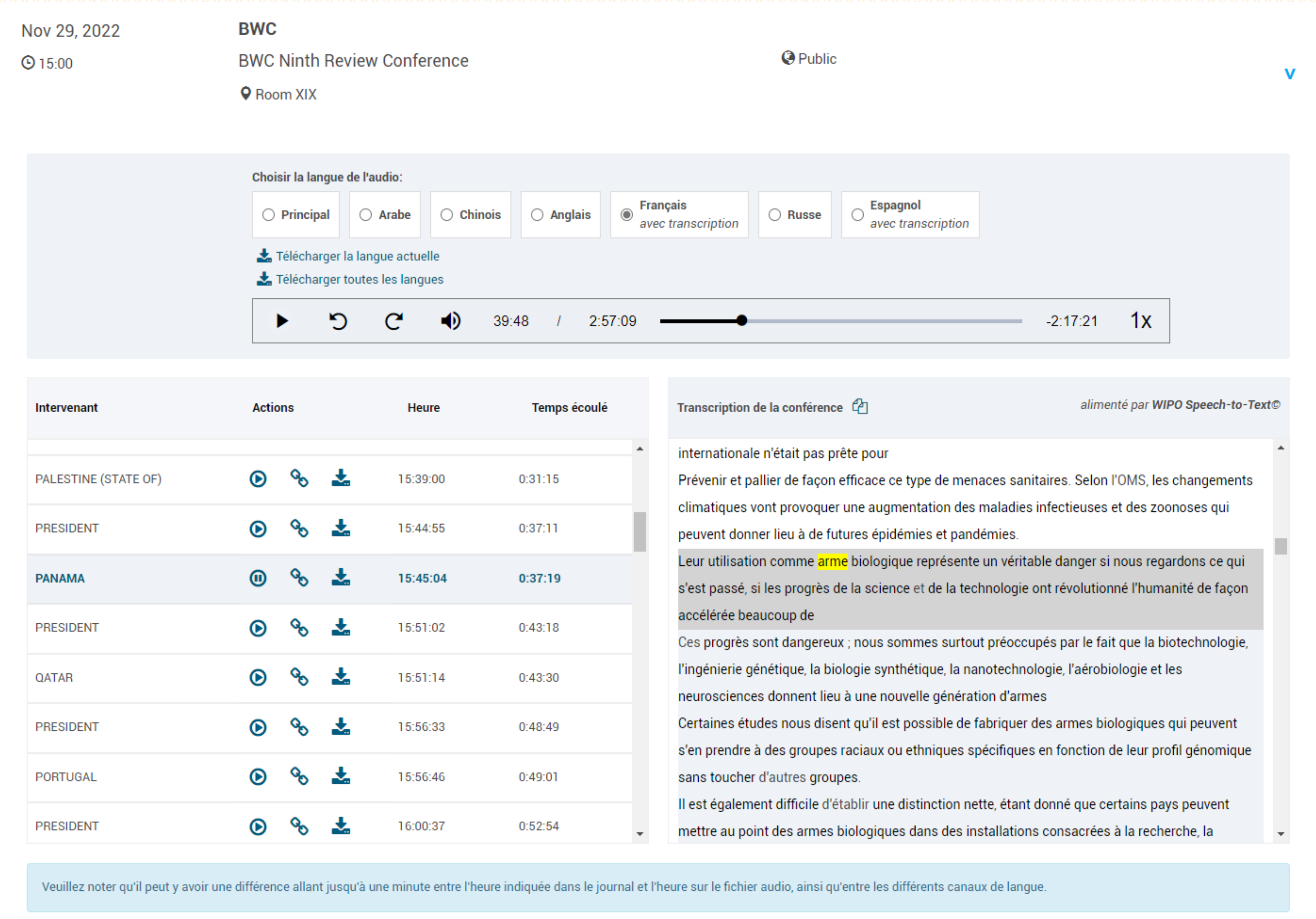
- Thanks to the World Intellectual Property Organization's Artificial Intelligence expertise, the FAST Project (Fully Automated Speech-to-Text) assists secretariats, delegates and internal UNOG staff by providing automated transcriptions for public meetings.

- The team collects UNOG audio data and matching transcripts to retrain the tool.

- FAST presently provides transcriptions in English, French and Spanish. Arabic, Mandarin and Russian language versions will be released in late 2023.

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

DIGITAL RECORDINGS PORTAL (DRP)



Nov 29, 2022

15:00

BWC

BWC Ninth Review Conference

Room XIX

Public

Choisir la langue de l'audio:

○ Principal ○ Arabe ○ Chinois ○ Anglais ● Français *avec transcription* ○ Russe ○ Espagnol *avec transcription*

Télécharger la langue actuelle

Télécharger toutes les langues

Intervenant	Actions	Heure	Temps écoulé
PALESTINE (STATE OF)		15:39:00	0:31:15
PRESIDENT		15:44:55	0:37:11
PANAMA		**15:45:04**	**0:37:19**
PRESIDENT		15:51:02	0:43:18
QATAR		15:51:14	0:43:30
PRESIDENT		15:56:33	0:48:49
PORTUGAL		15:56:46	0:49:01
PRESIDENT		16:00:37	0:52:54

Transcription de la conférence

*alimenté par **WIPO Speech-to-Text©***

internationale n'était pas prête pour

Prévenir et pallier de façon efficace ce type de menaces sanitaires. Selon l'OMS, les changements climatiques vont provoquer une augmentation des maladies infectieuses et des zoonoses qui peuvent donner lieu à de futures épidémies et pandémies.

Leur utilisation comme arme biologique représente un véritable danger si nous regardons ce qui s'est passé, si les progrès de la science et de la technologie ont révolutionné l'humanité de façon accélérée beaucoup de

Ces progrès sont dangereux ; nous sommes surtout préoccupés par le fait que la biotechnologie, l'ingénierie génétique, la biologie synthétique, la nanotechnologie, l'aérobiologie et les neurosciences donnent lieu à une nouvelle génération d'armes

Certaines études nous disent qu'il est possible de fabriquer des armes biologiques qui peuvent s'en prendre à des groupes raciaux ou ethniques spécifiques en fonction de leur profil génomique sans toucher d'autres groupes.

Il est également difficile d'établir une distinction nette, étant donné que certains pays peuvent mettre au point des armes biologiques dans des installations consacrées à la recherche, la

Veuillez noter qu'il peut y avoir une différence allant jusqu'à une minute entre l'heure indiquée dans le journal et l'heure sur le fichier audio, ainsi qu'entre les différents canaux de langue.

METRICS

- **Word Error Rate** (**WER**): Widely used metric that evaluates speech recognition system performance. It calculates the percentage of incorrectly recognized words (deletions, insertions, substitutions).

+ Quantative measure of accuracy
- Does not account for context or meaning.

Factors

- Accent
- Background noise
- Microphone / Signal quality
- Speech rate

- **Manual error analysis**: Testers help identify errors and limitations of the system.

+ Feedback for improvement
“User in the loop” approach
- Time-consuming and costly

GOALS

- Focus on quality evaluation as opposed to quantity.
- Evaluate the transcription in terms of readability and comprehension.
- Take into account variables that can’t be measured through automation.
- Identify the most critical errors and their impact on the user experience.

METHODOLOGY

- Listening to meeting recordings and comparing transcribed text with the original recordings.

- Identifying errors, such as deletions, insertions, substitutions, including incorrect capitalization and significant punctuation errors.

- Defining two categories of errors: disruptive and non-disruptive.

Output: *[…] par les **états**-Unis d’Amérique […]*
Reference: *[…] par les **États**-Unis d’Amérique […]*

SEVERITY OF ERRORS

- A disruptive error is critical based on its potential impact on the user experience and comprehension of the utterance.

  Important type of words: nouns, verbs, proper names, time, places, numbers

- Deletion errors: Considered unfavorable as the tool is missing its primary goal to transcribe the text.

Which error is worse?

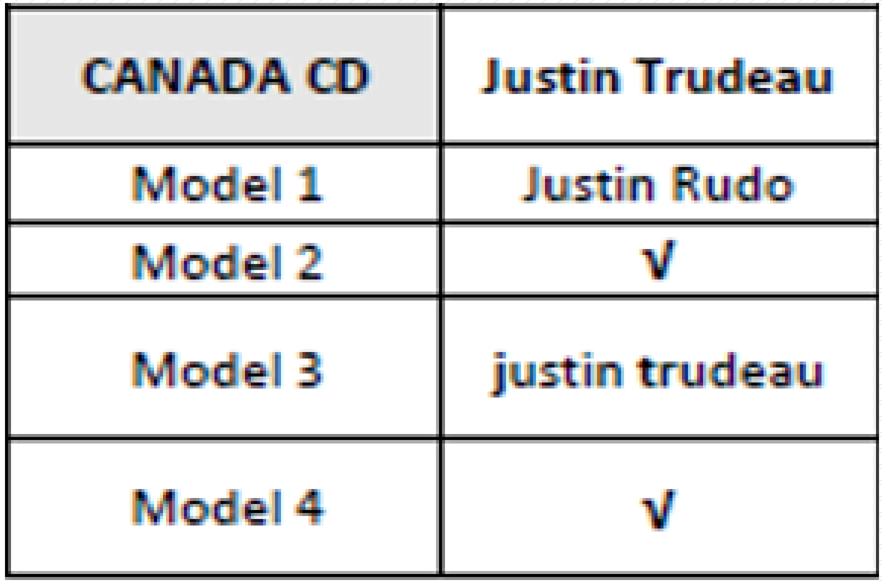
Reference
Le Royaume du Maroc qui a […] estime que notre sécurité collective demeure tributaire
Model 1
Le Royaume du Maroc qui a […] estime que notre sécurité collective demeure ?
Model 2
Le Royaume du Maroc qui a […] estime que notre sécurité collective des meurtres tributaire

Table: Comparison between a deletion and a critical substitution error

SEVERITY OF ERRORS

- Minor errors do not change the meaning of the sentence
  Pronouns, prepositions, adjectives, adverbs = function words

- Terminology



CANADA CD	Justin Trudeau
Model 1	Justin Rudo
Model 2	√
Model 3	justin trudeau
Model 4	√

Justin Rudo? Justin Trudeau?

CHALLENGES

- Native speakers
- Subjectivity
- Environmental complexity
- Diversity of accents
  - English: 58 countries
  - French: 35 countries
  - Spanish: 21 countries

"It may be wrong, but it's how I feel."

LANGUAGE SPECIFICITIES

- **French and Spanish**

Complex grammar, many rules including gendered nouns and adjectives, verb conjugation, and articles.

Presence of:

- Strong accents
- Regional pronunciation variations
- Different dialects
- Homophones

Example of a verb endings with the same pronunciation:
Aim-*er*, aim-*é*, aim-*ait*, aim-*ais*, aim-*aient*

Key Takeaways

1. Human evaluation is crucial for improving speech recognition accuracy, effectiveness and quality

2. While human evaluation has challenges, the benefits outweigh the costs.

3. Testers identify areas for innovation.