Médias numériques

Open AI's Whisper and Radio-Canada: From POC to Production

29 May 2024

Table of contents

Context

Speech to text at Radio-Canada

- **Context**
  - Audio/video content needs to be transcribed to apply certain AI techniques
  - Audio/video content is more and more popular and big tech companies invest a lot in it

- **Objectives**
  - Accessibility
  - Application of natural language processing models
    - Summarization
    - Topic classification
    - Named entity recognition (examples: people, dates, locations, organizations)

Open AI's Whisper

- State-of-the-art in STT for many languages, including French:
    - About 90% of words transcribed correctly in the evaluation datasets
- Code and pre-trained models are ***open source***:
    - Possibility of fine-tuning a Radio-Canada version of Whisper
- Bug fixes and improvements made by the community:
    - ***Faster Whisper*** (recommended by *Radio France*)
        - Faster processing times
        - Filters out parts of the audio without speech
            - VAD = Voice activity detector
- Existing API versions: Open AI's API and *Microsoft Azure Open AI Service*

Proof of concept

Evaluation framework: Metric

- ***Word error rate*** (***WER***) → Ratio of errors in a transcript to the total words spoken

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

- N = Total number of words
- S = Number of substitutions
- D = Number of deletions
- I = Number of insertions

ici jacques beauchamp aujourd'hui l'histoire la longue bataille contre la tuberculose la tuberculose s'attrapait par la respiration par tout simplement parler avec quelqu'un or ma mère était une grande malade et personne ne venait nous voir je n'amenais pas de petits(petite) amis(amie) chez moi c'était la maison de la tuberculose chez nous alors les gens rentraient(entraient) dans la porte pour pas ils pensaient que ça sautait sur le monde et j'ai(t'étais) remarqué(marqué) sa mère fait de la tuberculose il n'y avait plus personne il restait un enfant dans une famille de huit personnes il y en avait sept qui mourraient(mouraient) de tuberculose quand je suis rentrée au sanatorium les gens mouraient autour de moi je voyais les cadavres sortir tu te dis c'est mon tour l'animatrice jeannette(janette) bertrand se rappelle dans cet extrait de l'émission christiane charette en direct diffusé(diffusée) en 2001 de cette époque de sa jeunesse où la tuberculose faisait effectivement de véritables ravages au début des années 1940 la moitié des décès attribuables à la tuberculose chez les jeunes de 15 à 19 ans au pays touchaient(touchait) le québec en 1867 déjà la tuberculose était la première cause de décès au canada où vas-tu maman où vas-tu maman oui c'est toi mais on va te réchapper on va te guérir on va te faire un

Evaluation framework: Data

Type	TV show or podcast	Average length	Number of episodes	Particularities
TV	L'épicerie	22 min	5	Interview, regional accents, background noise
TV	Les grands reportages	45 min	2	Interview, scripted, not very good recording quality
TV	Oniva	15 min	18	Several speakers, regional accents
TV	Bonsoir bonsoir!	45 min	7	Several speakers, interruptions, music, regional accents
TV	On va se le dire	65 min	6	Several speakers, interruptions
OHdio	Aujourd'hui l'histoire	23 min	3	Regional accents, interview
OHdio	Les faits d'abord	2 heures	1.5	Regional accents, interview

Evaluation framework: Models

- **Available pre-trained Whisper models**

Size	Parameters	Required VRAM	Relative speed
tiny	39 M	~1 GB	~32x
base	74 M	~1 GB	~16x
small	244 M	~2 GB	~6x
medium	769 M	~5 GB	~2x
large	1550 M	~10 GB	1x

Results: TV shows

- **Résultats**
  - **Percentage of words that were correctly transcribed = *100% - WER***

	small	medium	large-v1	large-v2
L'épicerie	73.9%	82.7%	82.8%	85.5%
Les grands reportages	74.1%	83.1%	80.5%	86.4%
Oniva	70.9%	76.1%	77.2%	77.5%
Bonsoir bonsoir!	53.7%	56.5%	64.4%	61.7%
On va se le dire	52.5%	61.1%	55.7%	62.1%
Average	65.0%	71.9%	72.1%	74.6%

- **Observations**
  - The **large-v2** model has the best performance overall
  - The average performance of the **medium** model is less than 3 percentage points lower

Results: Hallucinations

- Present in some of the transcriptions obtained with the **large** models but not in those obtained with the **medium** model

dominique qui admire anne-élisabeth mais tout le monde l'aime voyons mon dieu quelle actrice ça a pas de bon sens je te regarde à la télé puis des fois je m'arrache les cheveux je me dis comment est-ce qu'elle fait elle est donc bien bonne ça a pas de bon sens bien là mon dieu seigneur ah non je te trouve tellement bonne merci c'est le fun c'en est déplaisant c'en est déplaisant elle est tellement bonne c'en est déplaisant pantoute elle me plaît tout le temps vrai(bien) c'est ça puis ça adonne qu'on a patrice l'écuyer pour la première fois avec nous puis j'ai l'impression que ça te ferait une bonne équipe de silence on joue vrai(oui) c'est sûr bien oui on dirait qu'on est à silence on joue hé que j'aimerais ça le pop corn dans la face puis tout ça puis anne-élisabeth on a vu ça passer les simone qui vont être présentées en france en suisse en fait en suisse de quessé que se passe-t-il j'aimerais vraiment ça te donner des détails j'ai repartagé l'information avec joie t'en sais pas plus mais je peux pas vous dire exactement comment ça va se passer mais quelle belle nouvelle mais ils le tournent vrai(ou) c'est(s'ils) vrai(passent) c'est(votre) vrai(série) c'est(ils) vrai(vont) c'est(la) vrai(passer) c'est(là-bas) vrai(ok) c'est votre vrai(série) c'est(doublée) vrai(je) c'est(sais) vrai(pas) c'est(patrice) vrai(sous-titrée) c'est(peut-être) vrai(il) c'est(faudrait) vrai(demander) c'est(à) vrai(kotv) c'est(mais) vrai(non) c'est(mais) vrai(je) c'est(suis) vrai(mais) c'est vrai(des) c'est(bonnes) vrai(nouvelles) c'est vrai(des) c'est(très) vrai(belles) c'est(nouvelles) vrai(peut-être) c'est(que) vrai(ça) c'est(va) vrai(être) c'est(juste) vrai(sous-titré) c'est(puis) vrai(ça) c'est(va) vrai(être) c'est(nous) vrai(puis) c'est(j'ai) vrai(tellement) c'est(aimé) vrai(ça) c'est(tourner) vrai(ça) c'est(que) vrai(toute) c'est(la) vrai(francophonie) c'est(que) vrai(toute) c'est(l'europe) vrai(diffuse) c'est(ça) vrai(que) c'est(le) vrai(monde) c'est(m'aime) vrai(le) c'est(monde) vrai(entier) c'est(non) vrai(mais) c'est(une) vrai(carrière) c'est(en) vrai(france) c'est(non) vrai(on) c'est(parlait)

Results: OHdio episodes

	Faster Whisper			Azure's API
	medium	large-v1	large-v2	
Aujourd'hui l'histoire	89.1%	90.7%	92.7%	92.5%
Les faits d'abord	90.4%	88.9%	91.1%	91.5%
Average	89.7%	90.4%	91.7%	92.2%

- **Observations**
  - The performance of Azure's API is very similar to that of the **large-v2** model. However, the API has a file size limit of 25MB
  - Hallucinations were not present in any of the transcriptions for these episodes
  - The average performance of the **medium** model is 2 percentage points lower than that of the **large-v2** model

Speaker identification

- **WhisperX : *Open source*** library based on ***Faster Whisper*** and ***Pyannote***


```
{
   "start": 45.765,
   "end": 51.729,
   "text": "Le Canada participe à la Biennale de Venise depuis 1952 et y a son propre pavillon...",
   "speaker": "SPEAKER_05",
   "words": [
        { "word": "Le", "score": 0.834, … }
        ...
   ]
}
```

OHdio podcast	% of sentences with correctly identified speakers
Aujourd'hui l'histoire	93.96%
Les faits d'abord	93.15%

Results: Limitations and how to work around them

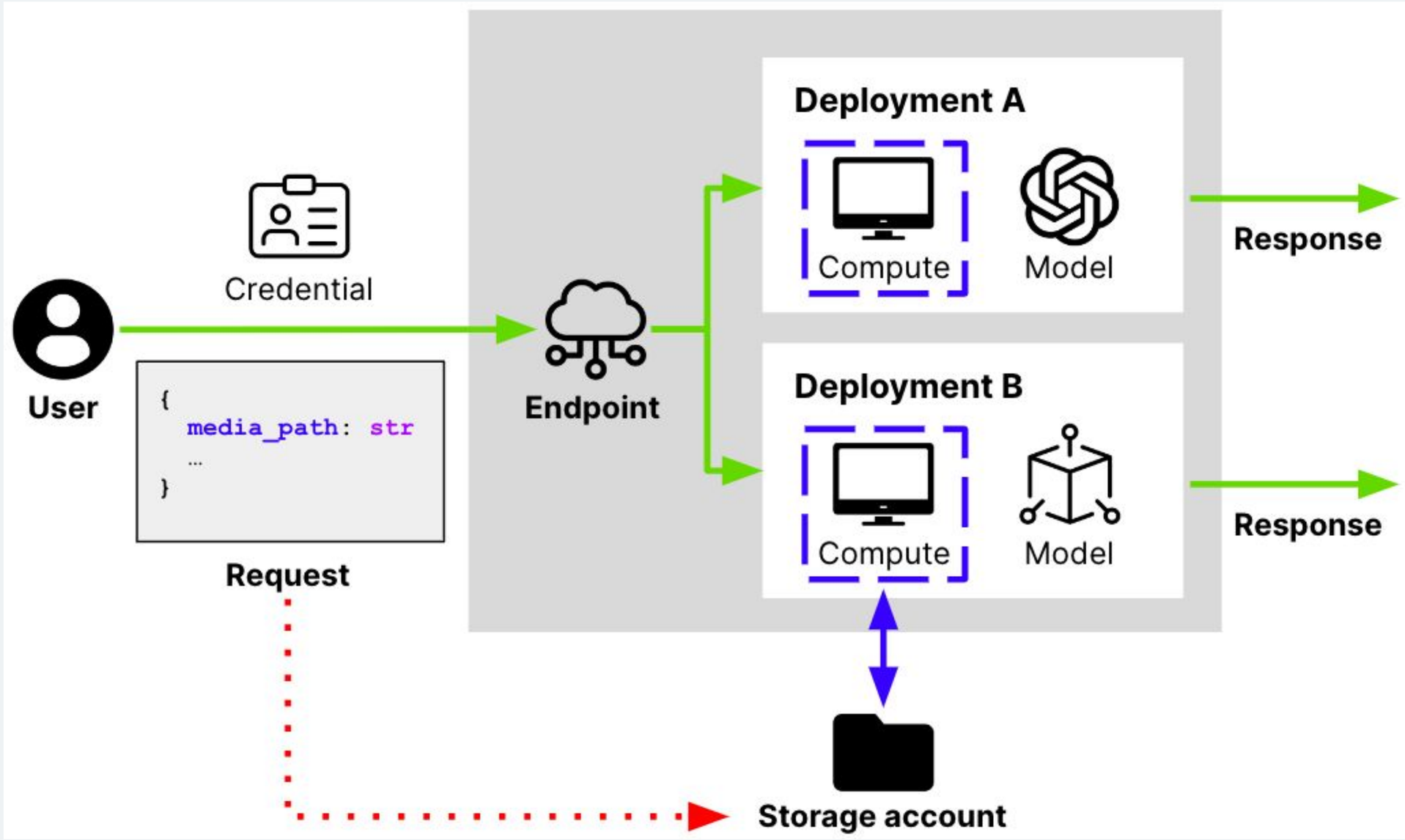
- **Hallucinations**
- **Song lyrics** are sometimes included in the transcription
    - The **duration** of a segment could be an indicator of its nature
        - Very long segments have a higher chance of being song lyrics
- **Interruptions** sometimes lead to transcription or speaker identification errors
- Portions of audio of **substandard quality** (*interviews on the phone*, *old audio recordings*) can also lead to errors
    - Low **confidence scores** can be indicative of this type of errors

From proof of concept to production

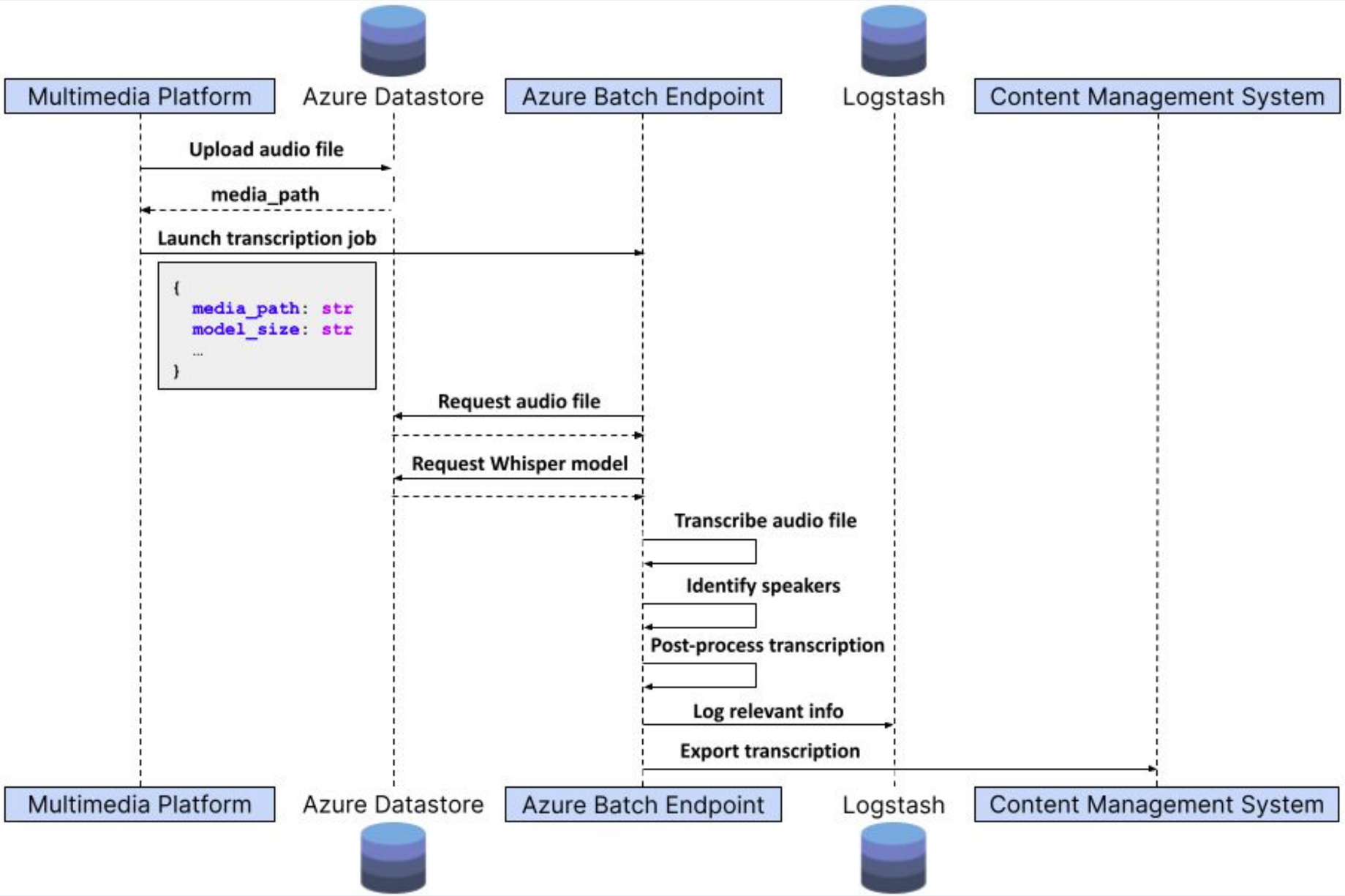
Requirements

- No low latency constraints for inference
- Flexibility
    - Whisper's params
        - Choice of model size
    - What if we want to use a different STT model in the future?

Azure batch endpoint



Infrastructure



Demo

Transcription (CMS) - Link demo

Possibility of fine-tuning a Whisper model

- **Québec French**
  - We found several patterns of errors in transcriptions

Names of places
Lanaudière → *L'anodiale* L'Île-du-Prince-Édouard → *Lille du pras-sédoire*

Surnames
Drouin-Brisebois → *Brisebois* Roy-Desmarais → *Desmarais*

Expressions
t'sais → *tu OU sais OU c'est* savais-tu → *tu*

Vocabulary
omicron → *micron* tiktok → *tok*

Next steps

- **Transcribe more and more content**
    - All radio shows
    - Podcasts
    - Videos

- **Give access to an on-demand transcription service**

Merci !

Médias

Transcription

Média 1 ORIGINALE

AJUSTER

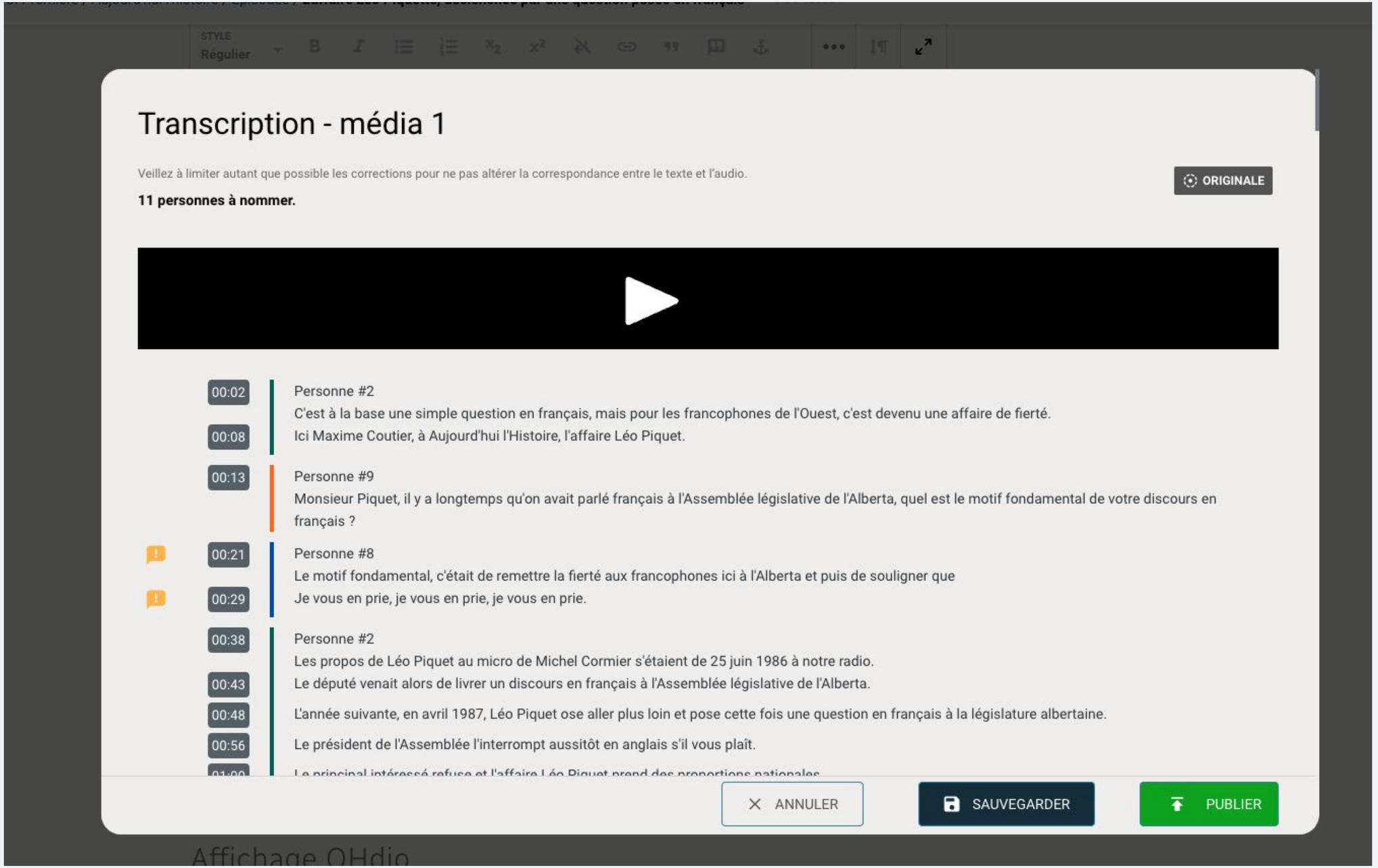
Ajuster la transcription pour la publier.

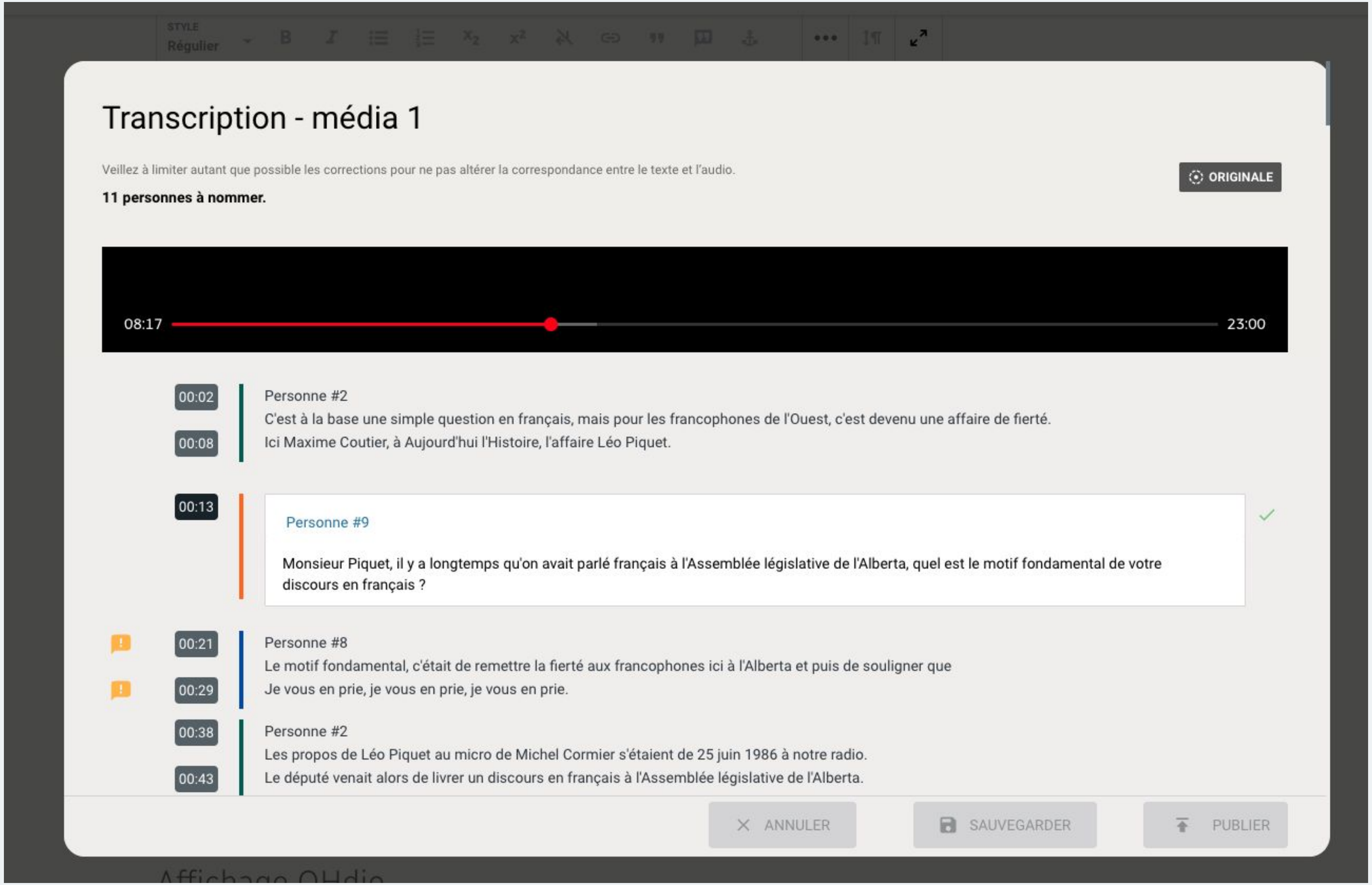
Il reste 11 personnes à nommer

Plateformes externes

Sélectionner un média (MP3) à intégrer

Comme l'émission est diffusée sur les plateformes externes, un fichier MP3 est requis.


STYLE
Régulier
Transcription - média 1
Veillez à limiter autant que possible les corrections pour ne pas altérer la correspondance entre le texte et l'audio.
ORIGINALE
11 personnes à nommer.
00:02
Personne #2
C'est à la base une simple question en français, mais pour les francophones de l'Ouest, c'est devenu une affaire de fierté.
00:08
Ici Maxime Coutier, à Aujourd'hui l'Histoire, l'affaire Léo Piquet.
00:13
Personne #9
Monsieur Piquet, il y a longtemps qu'on avait parlé français à l'Assemblée législative de l'Alberta, quel est le motif fondamental de votre discours en français ?
00:21
Personne #8
Le motif fondamental, c'était de remettre la fierté aux francophones ici à l'Alberta et puis de souligner que
00:29
Je vous en prie, je vous en prie, je vous en prie.
00:38
Personne #2
Les propos de Léo Piquet au micro de Michel Cormier s'étaient de 25 juin 1986 à notre radio.
00:43
Le député venait alors de livrer un discours en français à l'Assemblée législative de l'Alberta.
00:48
L'année suivante, en avril 1987, Léo Piquet ose aller plus loin et pose cette fois une question en français à la législature albertaine.
00:56
Le président de l'Assemblée l'interrompt aussitôt en anglais s'il vous plaît.
ANNULER
SAUVEGARDER
PUBLIER
Affichage OHdio


STYLE
Régulier
Transcription - média 1
Veillez à limiter autant que possible les corrections pour ne pas altérer la correspondance entre le texte et l'audio.
ORIGINALE
11 personnes à nommer.
08:17
23:00
00:02
Personne #2
C'est à la base une simple question en français, mais pour les francophones de l'Ouest, c'est devenu une affaire de fierté.
00:08
Ici Maxime Coutier, à Aujourd'hui l'Histoire, l'affaire Léo Piquet.
00:13
Personne #9
Monsieur Piquet, il y a longtemps qu'on avait parlé français à l'Assemblée législative de l'Alberta, quel est le motif fondamental de votre discours en français ?
00:21
Personne #8
Le motif fondamental, c'était de remettre la fierté aux francophones ici à l'Alberta et puis de souligner que
00:29
Je vous en prie, je vous en prie, je vous en prie.
00:38
Personne #2
Les propos de Léo Piquet au micro de Michel Cormier s'étaient de 25 juin 1986 à notre radio.
00:43
Le député venait alors de livrer un discours en français à l'Assemblée législative de l'Alberta.
ANNULER
SAUVEGARDER
PUBLIER

Comparison with Scribe - TV shows

	Faster Whisper			Scribe
	medium	large-v1	large-v2	
L'épicerie	82.7%	82.8%	85.5%	85.3%
Les grands reportages	83.1%	80.5%	86.4%	59.9%
Oniva	76.1%	77.2%	77.5%	72.4%
Bonsoir bonsoir!	56.5%	64.4%	61.7%	48.3%
On va se le dire	61.1%	55.7%	62.1%	57.3%
Average	71.9%	72.1%	74.6%	66.9%

- **Observations**
  - The **large-v2** model has the best performance overall
  - There are some **short** episodes where **Scribe** performs better
    - However, sometimes **Scribe** returns transcriptions that are **cut off**
  - **Trade-off** between the **medium** and the **large** Whisper models

Comparison with Scribe - OHdio episodes

	Faster Whisper			Scribe	Azure's API
	medium	large-v1	large-v2		
Aujourd'hui l'histoire	89.1%	90.7%	92.7%	86.2%	92.5%
Les faits d'abord	90.4%	88.9%	91.1%	84.8%	91.5%
Average	89.7%	90.4%	91.7%	85.8%	92.2%

- **Observations**
  - Scribe's performance is not as good as that of Faster Whisper and Azure's API
  - None of the tested models output transcriptions that were cut off or presented hallucinations
  - **Trade-off** between the **medium** and the **large** Whisper models