

# Turning Whisper into Real-Time Transcription System

**Dominik Macháček** (ÚFAL), Raj Dabre (NICT), Ondřej Bojar (ÚFAL)

[machacek@ufal.mff.cuni.cz](mailto:machacek@ufal.mff.cuni.cz)

2.11.2023, Bali



# Whisper

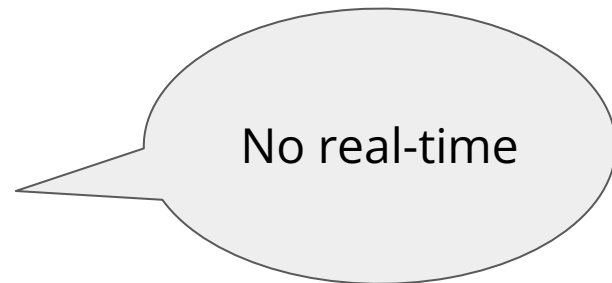
speech-to-text

[Radford et al., 2022]

# TL;DR: We made Whisper-Streaming in real-time mode

**Whisper**  
speech-to-text

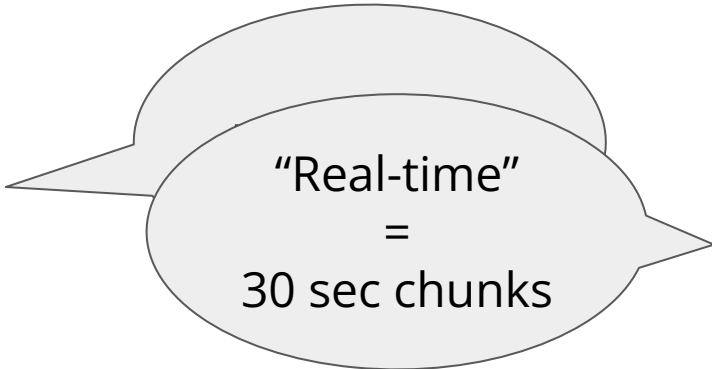
[Radford et al., 2022]



# TL;DR: We made Whisper-Streaming in real-time mode

**Whisper**  
speech-to-text

[Radford et al., 2022]



“Real-time”  
=  
30 sec chunks

# Streaming methods

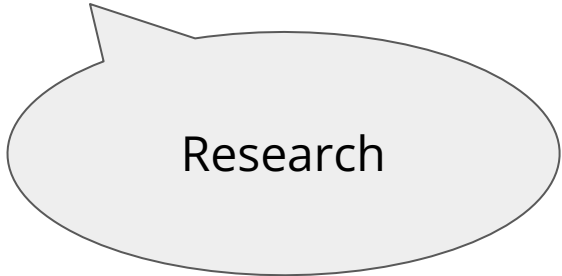
## Local-Agreement

[Liu et al., 2020, Polák et al., 2022, ...]

## Streaming methods

Local-Agreement

[Liu et al., 2020, Polák et al., 2022, ...]



Research

**TL;DR: We made Whisper-Streaming in real-time mode**

# Whisper-Streaming



[github.com/ufal/whisper\\_streaming](https://github.com/ufal/whisper_streaming)



# TL;DR: We made Whisper-Streaming in real-time mode

## Whisper-Streaming

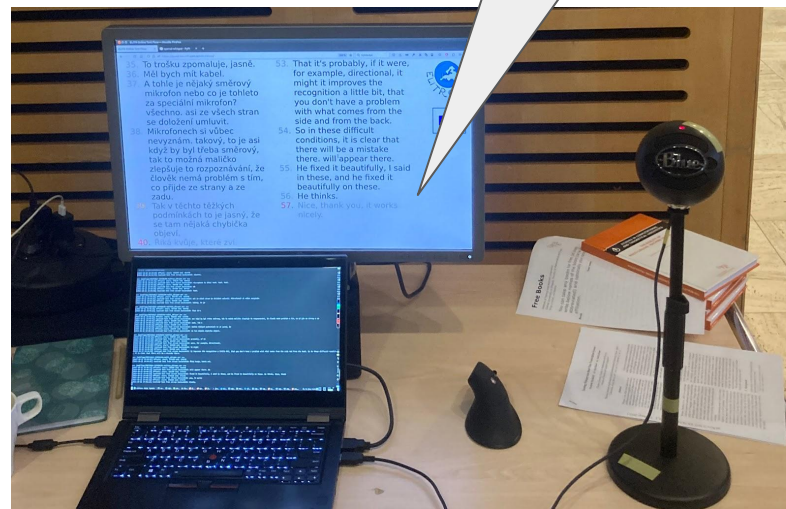


[github.com/ufal/whisper\\_streaming](https://github.com/ufal/whisper_streaming)



## Demo

Thanks, it works nicely!





# TL;DR: We made Whisper-Streaming in real-time mode

## Whisper-Streaming



[github.com/ufal/whisper\\_streaming](https://github.com/ufal/whisper_streaming)



## Demo

Thanks, it works nicely!

**Real-time speech-to-text**  
ASR + translation  
Videoconferencing  
Automatic minuting  
Voicebots

# How it works: Local Agreement-2

With a new audio chunk:



*audio buffer*

# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer



# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD)



# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript

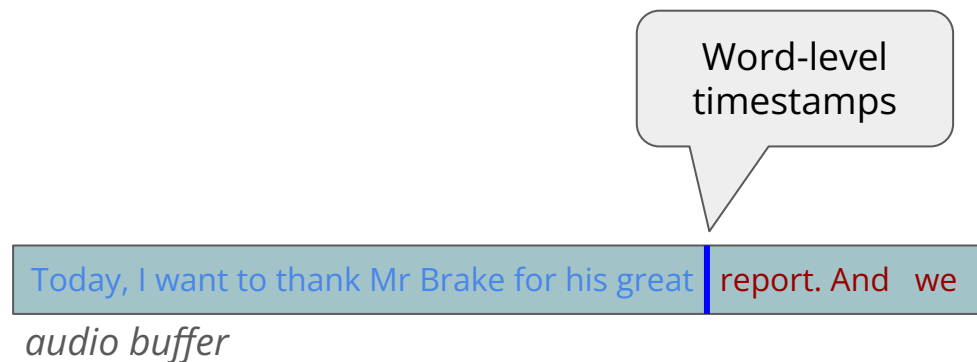
Today, I want to thank Mr Brake for his great report. And we

*audio buffer*

# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part



# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last **2** transcripts

Today, I want to thank Mr. Brejc for his great report. And

Today, I want to thank Mr Brake for his great report. And we

*audio buffer*

# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last 2 transcripts, confirm common prefix

Today, I want to thank Mr. Brejc for his great report. And

Today, I want to thank Mr Brake for his great report. And **we**

*audio buffer*



# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last 2 transcripts, confirm common prefix
5. Trim buffer at the end of sentence,

Today, I want to thank Mr. Brejc for his great report. And

Punctuation

And | we

*audio buffer*

# How it works: Local Agreement-2

With a new audio chunk:

1. Append to the audio buffer
2. Process buffer -> (VAD) -> text transcript
3. Skip previously confirmed part
4. Compare last 2 transcripts, confirm common prefix
5. Trim buffer at the end of sentence,  
update "prompt" = inter-sentence context

Today, I want to thank Mr. Brejc for his great report.

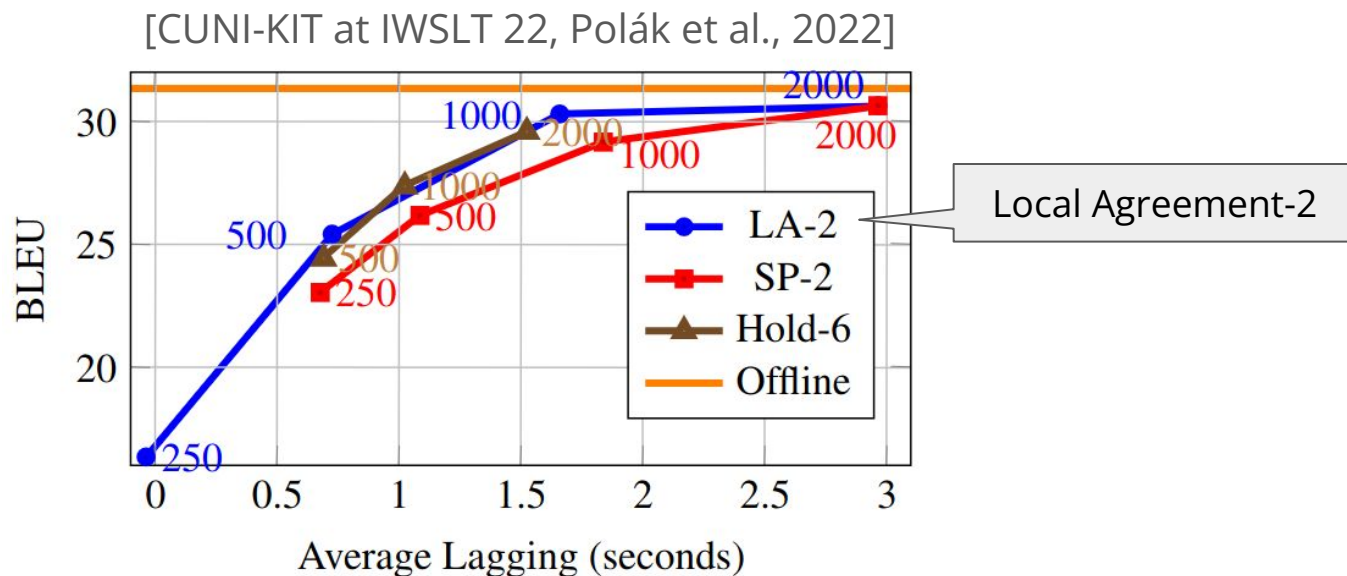
*prompt*

And we

*audio buffer*

# Why Local Agreement-2 [Liu et al., 2020]

- Self-adaptive latency = Waits by the uncertainty in language/content
- Best in IWSLT 2022 competition [Polák et al., 2022]
- Min. latency 2-times chunk-size
- Max. unlimited



# Real-time processing


- Invariant:
  - Buffer starts with a new sentence

# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds

# Real-time processing

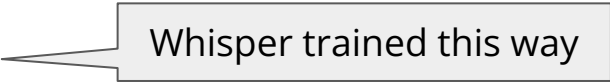
- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds



Whisper trained this way

# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds

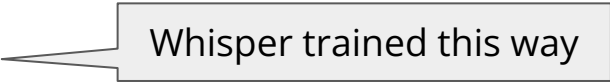


Whisper trained this way

- **faster-whisper** backend on GPU:

# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds
- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

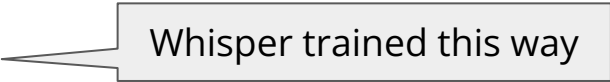


Whisper trained this way



# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds
- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time
- Parameter: Update with
  - [MinChunkSize] of new audio,



Whisper trained this way

# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds
  
- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time
  
- Parameter: Update with
  - [MinChunkSize] of new audio,

Whisper trained this way

0.25 sec/0.5 sec/1.0 sec/ ...

# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds
- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time
- Parameter: Update with
  - [MinChunkSize] of new audio, or whatever received.

Whisper trained this way

0.25 sec/0.5 sec/1.0 sec/ ...

# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds
- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time

Whisper trained this way

- Parameter: Update with
  - [MinChunkSize] of new audio, or whatever received.

0.25 sec/0.5 sec/1.0 sec/ ...

Processing can take longer.

# Real-time processing

- Invariant:
  - Buffer starts with a new sentence
  - At most 30 seconds
  
- **faster-whisper** backend on GPU:
  - 30 seconds audio ~ in 1 second => real-time
  
- Parameter: Update with
  - [MinChunkSize] of new audio, or whatever received.
  - Controls the latency and quality

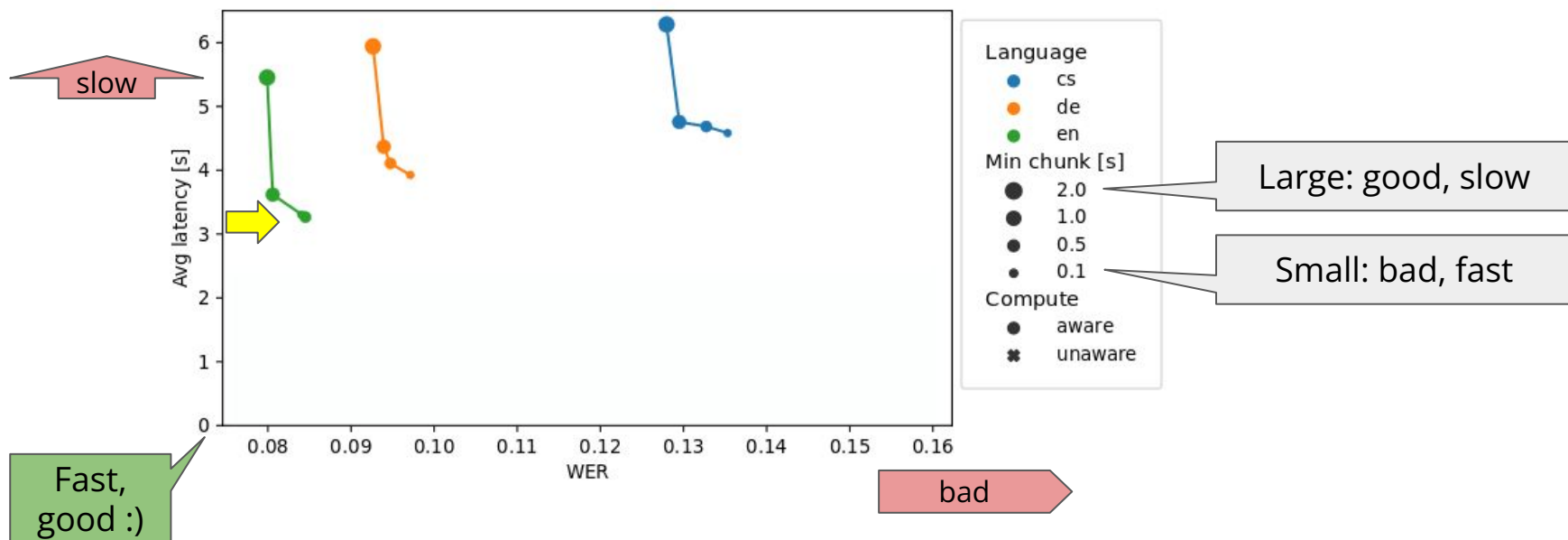
Whisper trained this way

0.25 sec/0.5 sec/1.0 sec/ ...

Processing can take longer.

# ASR Performance tests

- ESIC – Europarl, English orig., German, Czech interpreting [Macháček et al., 2021]
  - NVIDIA A40 GPU, Whisper large-v2
- ➔ English 0.5s m.ch.: 8.5% WER, **3.3s** avg. latency

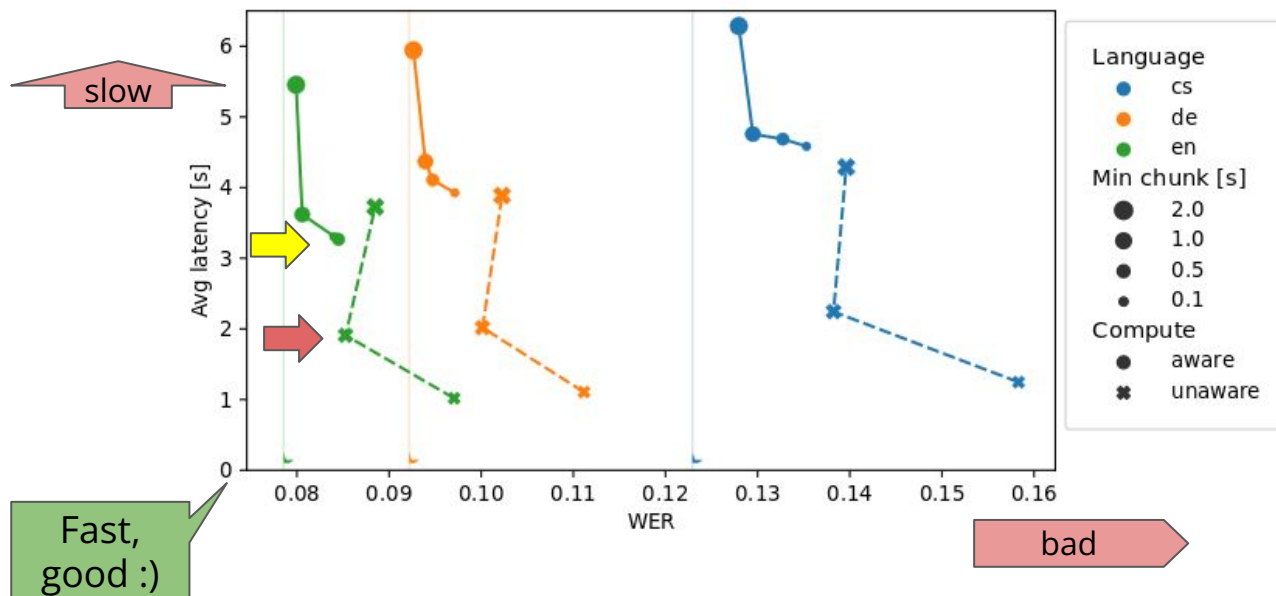


# ASR Performance bounds

- Computationally unaware = “optimal hardware speed”
- Offline ASR quality = “optimal quality”
- ➔ English 0.5s m.ch.: 8.5% WER, **3.3s** avg. latency -> unaw. +1.2% WER, -2.5s = 1.7s
- Offline: -1.8% WER

Model, language

Implementation, hardware




slow

Fast,  
good :)

bad

# Demonstration

- Integration with ELITR live speech translation framework
- Evaluation event – one day conference in    -> excellent quality 😊
- **Interactive** demo
  - Speak in any of **96** langs.! Observe the quality-latency! Have a chat!
  - AACL: ✗ slot in program ✗ desk ✗ Internet cable ✗ screen 😞 => in breaks on request

powerful experience.

80. That was Josef Pazdlerka from Czech Radio Plus.

81. And now, the President of the Czech Republic, Petr Pavel.

82. Please come up here on stage, and present your opening speech to start the first session of this conference.

83. Mr. President.

84. Good morning, ladies and gentlemen, guests here and listeners and viewers on the other platforms.

85. When I was asked by the Czech radio to take over the auspices of this event, I did not hesitate for a second, because the topics that we are discussing here today are very important to me.

86. This is the 100th anniversary since the start of the regular broadcast of the Czech radio, which also tells us about the importance of freedom of speech, of talking without censorship, without limitations, the freedom to accept information, to seek information, to spread information, the freedom that in many parts of the world is restricted very strongly, and a freedom... people keep giving their lives for.

87. And specific examples are not far away.

88. We have among us the daughter of Boris Nemtsov, the murdered Russian opposition politician, Zhanna Nemtsova.

89. On Vyhorodska street, quite close to the headquarters of the Czech Radio, there is Radio Free Europe, and three of its journalists are now in prison,

tepujinky. Petra Pavla, auy přisel sem k nám a přednesl svůj úvodní projev a vlastně tak otevřel ten první blok celé konference.

60. Blok nazvaný Ukrajina jako společná odpovědnost.

61. Prosim, pane prezidente.

62. Dobry den, damy a panove, vazeňi hoste zde v sále, posluchači, ale také diváci na ostatních platformách.

63. Když mě vedení Českého rozhlasu požádalo o zástihu na dnešní konferenci, nemusel jsem dlouho váhat, protože témata, kterými se tady zabýváme, jsou pro mě velice důležité.

64. Připomínáme si 100. výročí odzahžení pravidelného rozhlasového vysílání a to je zároveň i připomínkou významu svobody slova.

65. Svobody vyjadřovat se bez cenzury a bez omezení.

66. Svobody přijímat informace a myšlenky, vyhledávat je a šířit.

67. Svobody, která je v různých koutech světa stále výrazně omezoována a za její šprosazování lidé i dnes platí tu nejvyšší cenu.

68. Pro konkrétní příklady nemusíme vůbec chodit daleko.

69. Mezi námi je dnes dcera zavražděného ruského opozičního politika Borise Němcova, žena Němcovová.

70. Na ulici Vinohradská, jen kousek od sídla Českého rozhlasu, sídlí i Radio Sobotná Evropa.

71. Jehož tři novináři jsou dnes vězněni. - Jihrad Losik a Andrej Kuzněčik v Bělorusku a Vladislav Jespenko na ruském okupovaném Krymu.

72. V únoru tohoto roku jsme si připomněli pět let od vraždy slovenského



is-sur president.

84. Filghodu tajeb, nisa u mara, mistiednija hawn u dawk li jisimghu u l-ispetturi fuq il-

85. Meta ntablani mir-radjuj Ċeka biex tiehu l-awdi ta' dan l-avveniment, ma stajtx ghal sekonda, minhabba li s-sugġetti li qed niddiskutu hawn llum huma importanti hafna għajja.

86. Dan huwa l-100 anniversarju mill-bidu tat-trażmissjoni regolari tar-radju Ċek, li jgħidilna wkoll dwar l-importanza tal-libertà tal-kunsiderazzjoni, ta' tkellem minghajr censura, minghajr limitazzjonijiet. Il-libertà li jaċċettaw informazzjoni, li jiftixu informazzjoni, li jinfixu informazzjoni, li-libertà li f'hafna partijiet tad-dinja hija ristretta hafna b'saħħitha, u libertà... in-n-

87. U eżempji speċifiċi mhumiex bogħod.

88. Ahna għandna fostna t-tifla ta' Boris Nemtsov, il-politika ta' l-oppożizzjoni Russa maqta, Zhanna Nemts

89. Fuq il-triq ta' Vyhorodska, għib hafna mill-kwartieri ġenerali tar-radju Ċeka, hemm ir-radju Hlelsa ta' l-Ewropa, u

вступиле слово. Відкрив перший блок конференції.

70. Блок під назвою «Україна як спільна відповідальність».

71. Прошу пана президента.

72. Доброго дня, дами та панове, доброго дня гості в залі, слухачі, а також глядачі на інших платформах.

73. Коли керівництво Чеського радіо попросило мене взяти патронат на цій конференції, я не вагався.

74. Тому що на теми, про які ми сьогодні будемо говорити, це теми дуже важливі для мене.

75. Сьогодні ми пригадуємо соту річницю від початку трансляції Чеського радіо.

76. І це також нагадування про важливість свободи слова.

77. Свободи висловлювати свою думку без обмежень.

78. Свободи приймати інформацію та думки, шукати їх та поширювати.

79. Свободу, яка у різних частинах світу досі піддається переслідуванням.

80. І за неї люди і сьогодні платять найвищу ціну.

81. За такими прикладами нам не треба ходити далеко.

82. Сьогодні серед нас є донька вбитого російського політика Жінна Німцова.

83. У вулиці Віноградська, зовсім недалеко від місця, де знаходиться Чеське радіо, знаходиться і радіо «Свобода».

84. Три журналіста, якого зараз знаходяться за ґратами.

85. У лютому цього року

موضوع هذا المؤتمر واضح تماما. للتأكد على نوعية المعلومات التي تأتي إلى الجمهور التشيكي. من المهم ما نوع المعلومات التي يستهلكها.

75. ونابأ، لم يرغب، بعد هذا أكثر من سنة الصراع، لم يرغب في أن يُنظر إلى هذا كالألعاب فيديو.

76. بعض الحركات على الخريطة وما زال هناك مصير فرادي الناس، معاناة فرادي ما يحدث.

77. طوال هذه الألواع، طوال اليوم، يجب أن تكون قادراً على رؤية على الأقل نظره على ذلك.

79. وأمل أن تكون متحررة للاهتمام وافية.

80. كذلك (جورجف باردركا) من الراديو التشيكي (ب)

81. والرئيس الجمهورية التشيكية، بنتر بل.

82. الافتتاحية لهذه الدورة الأولى لهذا المؤتمر، المؤتمر، أوكرانيا كمتسؤولية مشتركة.

83. سيدي الرئيس،

84. صباح الخير يا سيداتي وسادة صوف هنا والمستمعين والمناضين على التبريرات الأخرى وعندما طلبت من الإذاعة التشيكية أن تنولي رعاية هذا الحدث، لم أتردد لعده نايه، لأن المواضيع التي ناقشنا اليوم هامة جدا بالنسبة لي.

86. هذا هو الذكرى السنوية منذ بداية البت المنظم للإذاعة التشيكية، التي تخبرنا أيضاً بأهمية حرية التعبير، والتحدث بدون رقابة، دون قيود.

87. ونسبر المعلومات، والحرية التي نقدر في العديد من أنحاء العالم بقوة جداً، والحرية... الناس يستمعون في إعطاء جانبهم من أجلها.

88. وهناك أصلة محددة ليست بعيدة عن ذلك لدينا بناه ابنة بوبريس نامتسوف، سياسة المعارضة الروسية المنقلة، رانا نامتسوف.

89. في شارع فيهورودسكا، قريبا جداً من مقر الإذاعة التشيكية، هناك إذاعة أوروبا الحرة، وثلاثة من صحفيا في السجن الآن.





## Summary

# We made Whisper-Streaming in real-time mode

- Speech-to-text with Local-Agreement policy
- Interactive demo on request
- Simple and robust implementation



[https://github.com/ufal/whisper\\_streaming](https://github.com/ufal/whisper_streaming)

# References

[Liu et al., 2020] Low-Latency Sequence-to-Sequence Speech Recognition and Translation by Partial Hypothesis Selection. Proc. Interspeech 2020, 3620-3624, doi: 10.21437/Interspeech.2020-2897

[Macháček et al., 2021] Lost in Interpreting: Speech Translation from Source or Interpreter? Proc. Interspeech 2021, 2376-2380, doi: 10.21437/Interspeech.2021-2232

[Radford et al., 2022] Robust Speech Recognition via Large-Scale Weak Supervision, <https://cdn.openai.com/papers/whisper.pdf>

[Polák et al., 2022] CUNI-KIT System for Simultaneous Speech Translation Task at IWSLT 2022, In Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pages 277–285, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

ELITR – European Live Translator, [elitr.eu](http://elitr.eu)

**Dominik Macháček** – [ufal.cz/dominik-machacek](http://ufal.cz/dominik-machacek)