Multimodal LLMs: Speech and Real-Time

Peter Polák, Dominik Macháček
2/5/2024

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

After the class, you should be able to:

- Know motivation for speech in LLMs
- Know the basic and example speech-to-text methods
- Know real-time methods

Class outline

- Speech NLP tasks (ASR, translation, emotion recognition, ...)
- Speech in NNs (sound representation, MFCC, raw audio) and in LLMs (Wav2vec, HuBERT, Whisper)
- Simultaneous methods: re-translation vs. incremental
- Streaming policies wait-k and LocalAgreement
- Whisper-Streaming and ELITR demo

Outline

- Motivation
- Speech and NNs
  - Representation
  - Speech-to-text generation
- Speech and LLMs
- Simultaneous Methods
- Demo

Speech and LLMs

Why speech?

Why should LLMs work with speech … from **user perspective**?

- More natural interaction
- Accessibility
- Some languages do not have a written form
- Complementary with text
- New applications

Why speech?

Why should LLMs work with speech … from **technical perspective**?

- Using information beyond text
  - Prosody
    - Intonation, stress, and rhythm
  - Non-verbal language
  - Sentiment
  - Environment

- Avoiding error propagation

Speech and NNs

Speech NLP Tasks

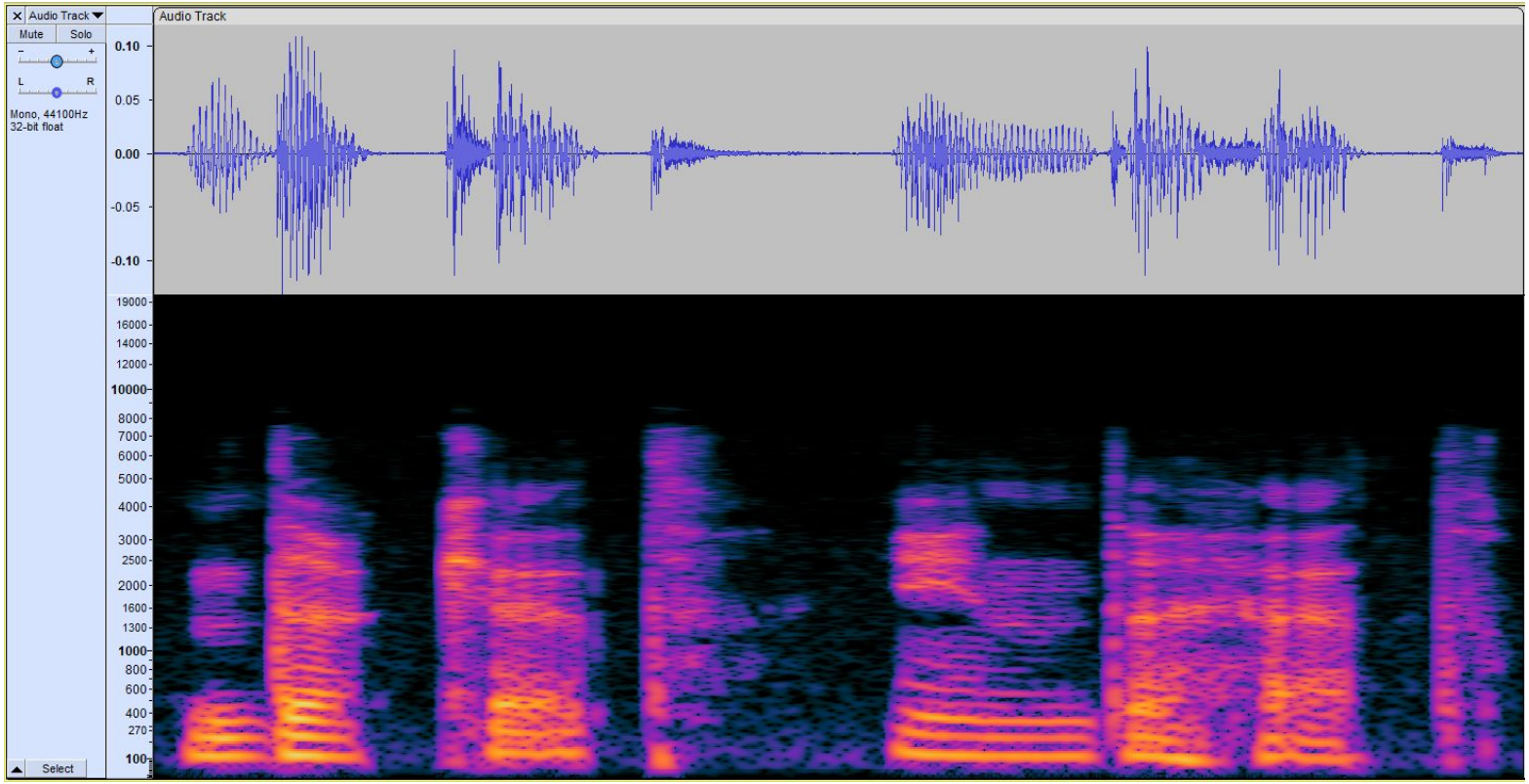
- Speech-to-text
  - Automatic Speech Recognition (ASR) 🗣️→🔤
  - Speech Translation 🗣️→ ㊙️
  - Summarization 🗣️→ □
- Emotion Recognition 🗣️→🤗
- Speaker Verification 🗣️+🆔→👍
- Speaker Identification 🗣️→🆔
- And many more …

Speech Representation for NNs

Speech Representations: Recap from Text

- How neural networks read?
- Tokenization
  - Break the text into smaller units = tokens [characters, words, sub-words]
- Translate tokens to indices in a vocabulary
- Embedding Layer
  - Translate each index into a vector

Speech Representation

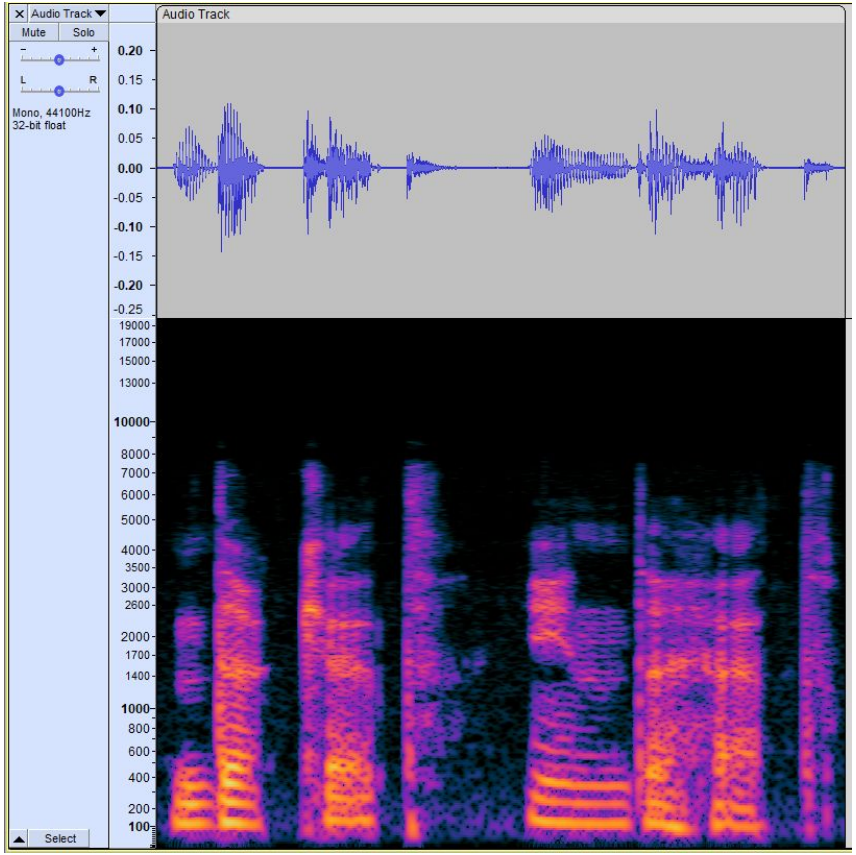


the cat in the hat

Speech Representations: Recap from Text

- How do we represent sound in computer?
  - Human speech 100 - 4 kHz, better to 8 kHz
  - 16 kHz wav = 16k floats/s

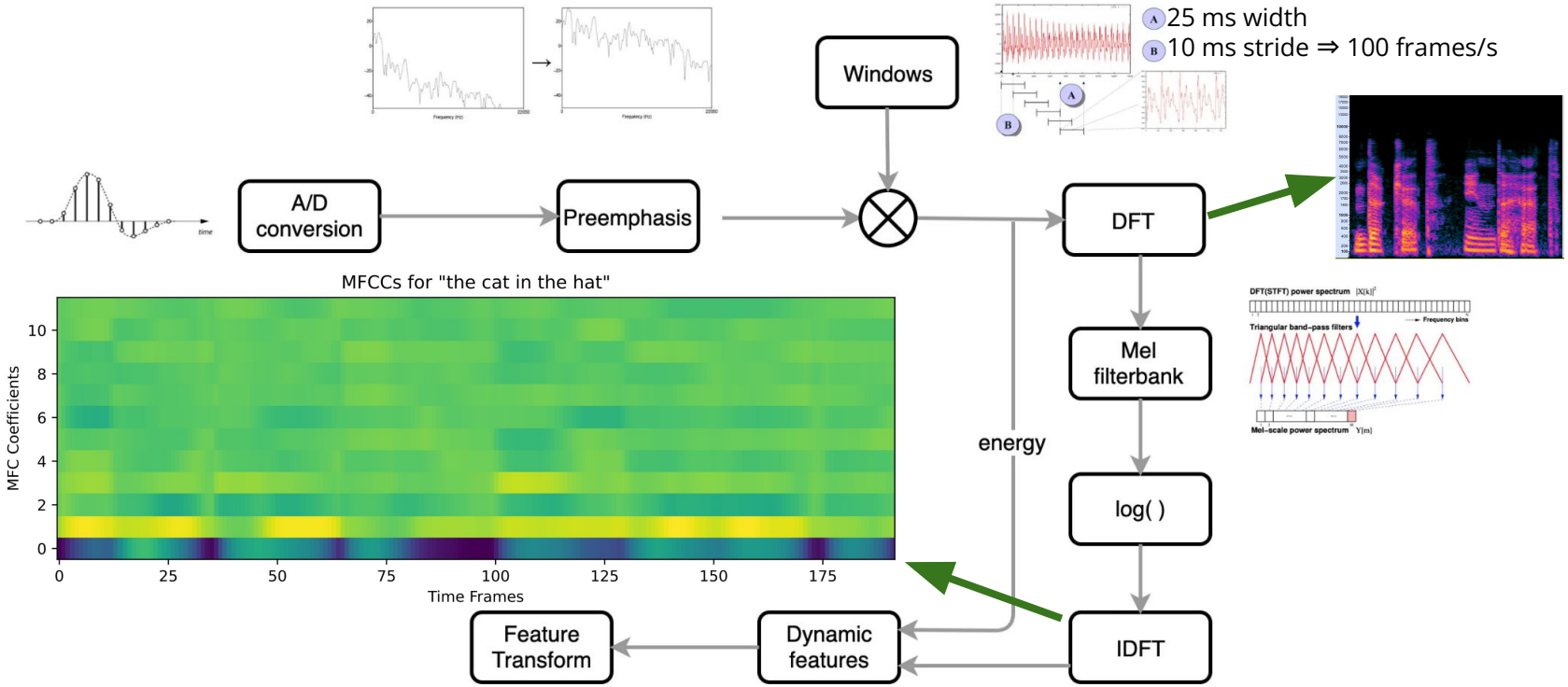
- How NNs understand speech?
  - Two approaches:
    - MFCCs
    - Raw audio (some tricks needed)



the cat in the hat

Speech Representations: MFCCs

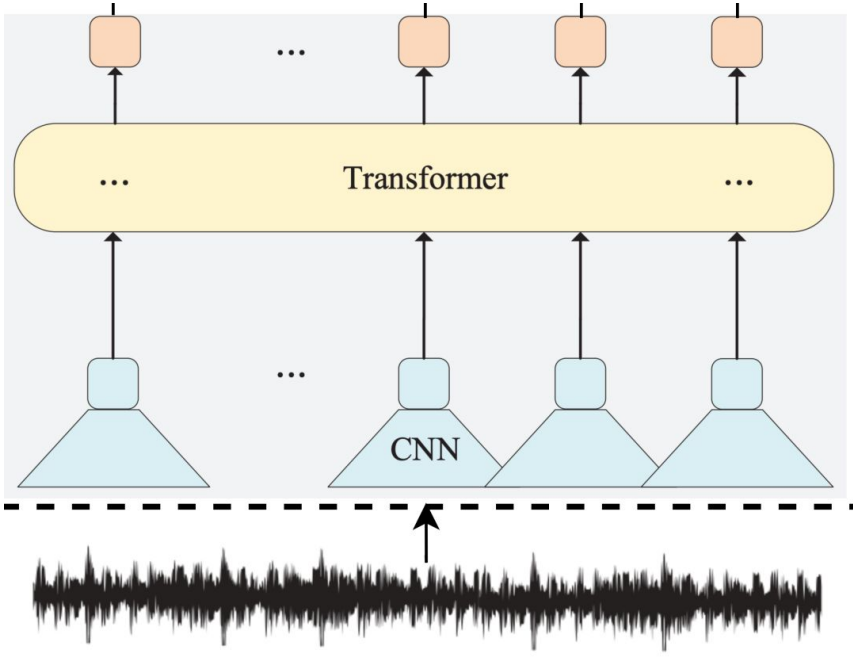
- Mel Frequency Cepstrum Coefficients



https://jonathan-hui.medium.com/speech-recognition-feature-extraction-mfcc-plp-5455f5a69dd9

Speech Representations: Direct Approach

- Feed directly 16 kHz to NN
  - We want to use Transformers - where is the problem?
    - Complexity of self-attention - $O(n^2)$
  - Solution
    - Downsample long input with CNNs
    - CNN serves as feature encoder
      - Similar to MFCCs
      - But the representation is **learned from data**
  - Typically a part of pre-trained models
    - Wav2vec 2.0, HuBERT, WavLM



From: SPEECH EMOTION DIARIZATION: WHICH EMOTION APPEARS WHEN?

Speech Representations: Comparison

- MFCCs
  - 👍 Efficient
  - 👍 Interpretable
  - 👎 Limited features
    - Ideal for ASR, not ER
  - 👎 Not robust to noise

- Direct approach
  - 👍 More complex features
    - speaker characteristics, emotions, and environmental noise
  - 🤔 Depends on training data
    - Can be robust
  - 👎 Computational cost
  - 👎 Interpretability

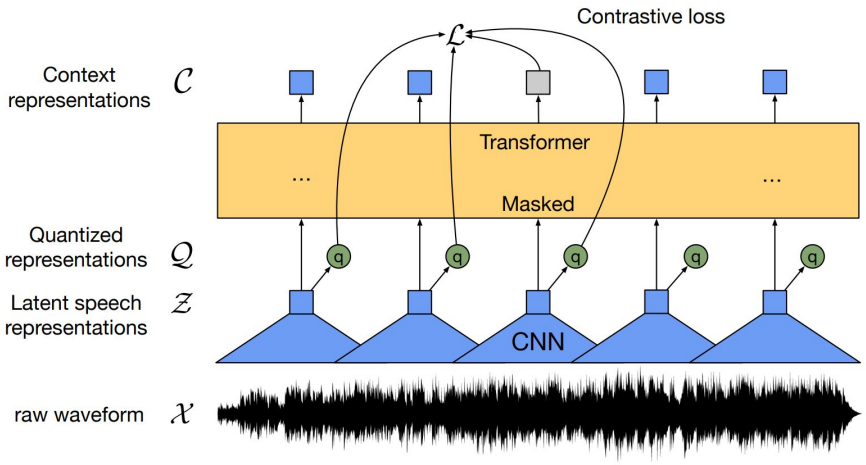
Leveraging Unlabeled Data: Self-Supervised Learning

- We have a lot of unlabeled data
  - Can we use them?

- Pseudo-labeling
  - Use existing ASR
  - 🤔 ASR quality
  - 👎 No ASR for some languages

- Self-supervised learning (SSL)
  - Wav2Vec 2.0, HuBERT, WavLM, …

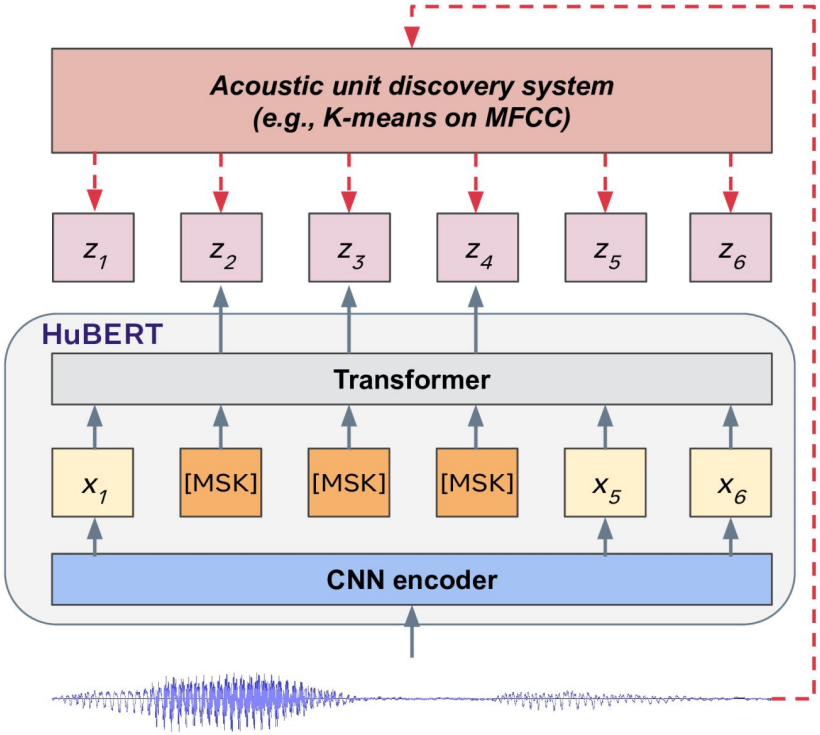
- Idea:
  - Model decides on representations based on data
  - Improved context representation using Transformer



From: wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Leveraging Unlabeled Data: Example

- HuBERT
  - CNN feature encoder
    - 320x downsampling = 50 frames/s
  - Transformer context encoder
- Training
  - Mask random spans
  - Predict acoustic units in masked regions
  - Several hours on up to 256 GPUs
- Acoustic units
  1. Unlabeled audio → MFCCs → k-means (k=100)
  2. Contextual features from Transformer from the first step → k-means (k=500)
- Finetune for downstream task
  - ASR, ER, …
  - Works even with 10 mins of data



From: wav2vec 2.0: HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units

Plugging Speech to LLMs

- Hot research topic
- Typically
  - Embed audio with some encoder
    - HuBERT, …
  - Interleave audio and prompts
  - Feed to LLM

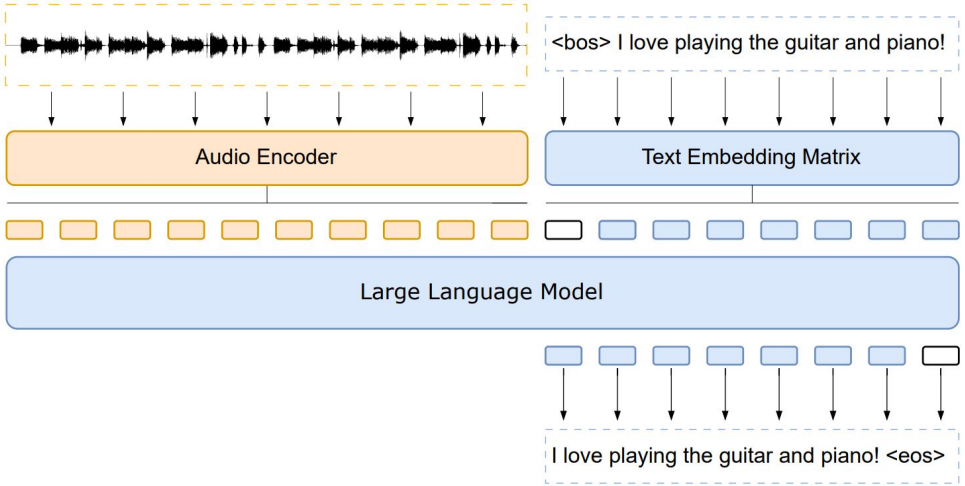
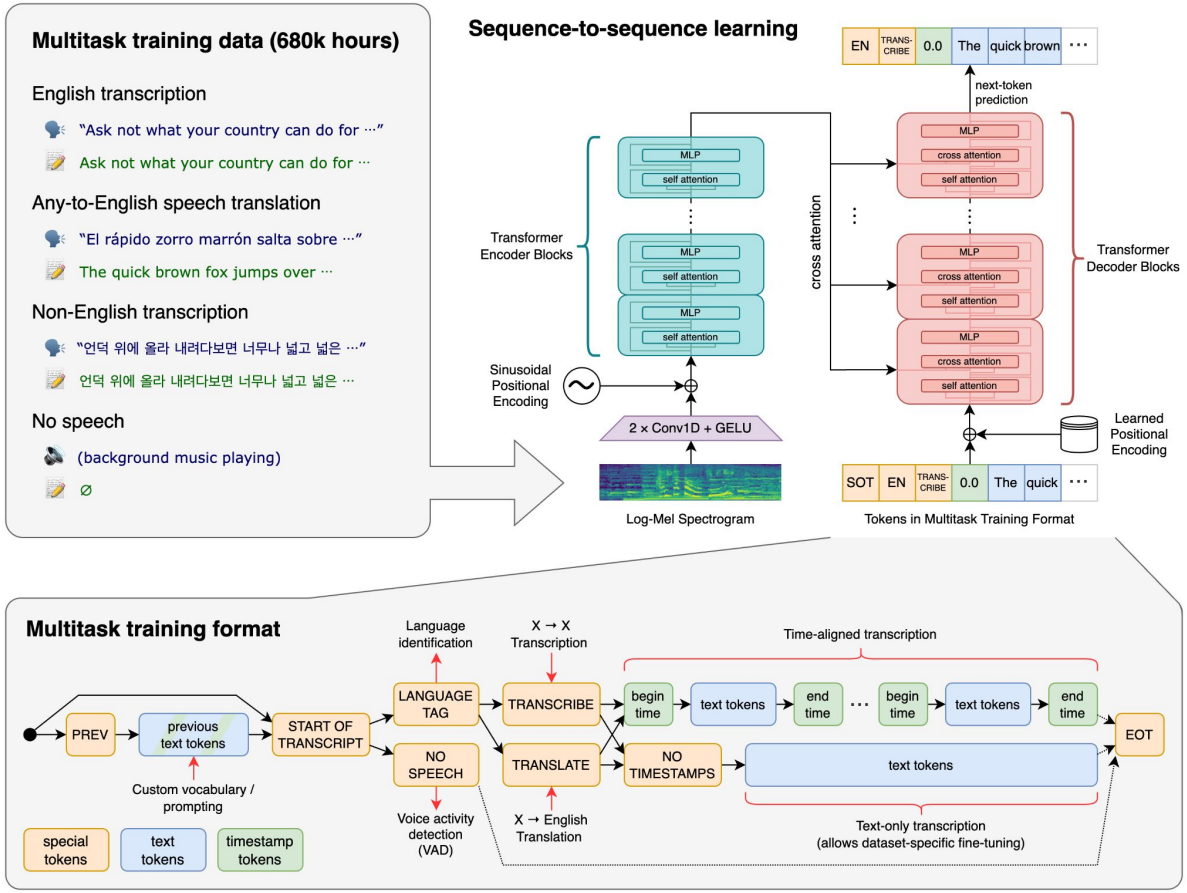


Figure 2: Model architecture. The embedding sequence generated from the audio encoder is directly prepended to the text embeddings sequence. This is directly fed into the decoder-only LLM, tasked with predicting the next token. The LLM can be frozen, adapted with parameter efficient approaches such as LoRA or fully finetuned. This work will investigate the former two.

From: Prompting Large Language Models with Speech Recognition Abilities

Speech-to-Text Example: Whisper



LLMs in Real-Time: **Simultaneous Speech Translation**

Dominik Macháček

2/5/2024, recap from 11/4/2024 + new parts

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Problem:

- Seq-to-seq task, e.g. 🗣️→ 🔤/㊙️/□/...
- Continuous incremental input 🗣️🗣️🗣️...
  - = Don’t assume <end-of-sequence> mark
  - But expect it in a real-life app. Use e.g. Voice activity detection for 0.5s silence

- **Problem:** How to process the task fast?

  = In **real-time**? 🏄

  = Simultaneously? 👥

Simultaneous speech translation (recap 11/4/2024)

- **Simultaneous** = Live = Real-Time = Low-latency = Incremental
  - Source available continuously, one **chunk** at a time
  - The **chunk** can be:
    - audio segment … in the direct speech-to-text translation or <u>transcription = ASR</u>
    - or word (text) produced by incremental ASR … in a cascaded system = ASR + **MT**
  - Provide the target "<u>at the same time</u>" as the source is being produced

    <u>= simultaneously</u> = with a small additive delay

In the European Parliament:

-> English original source

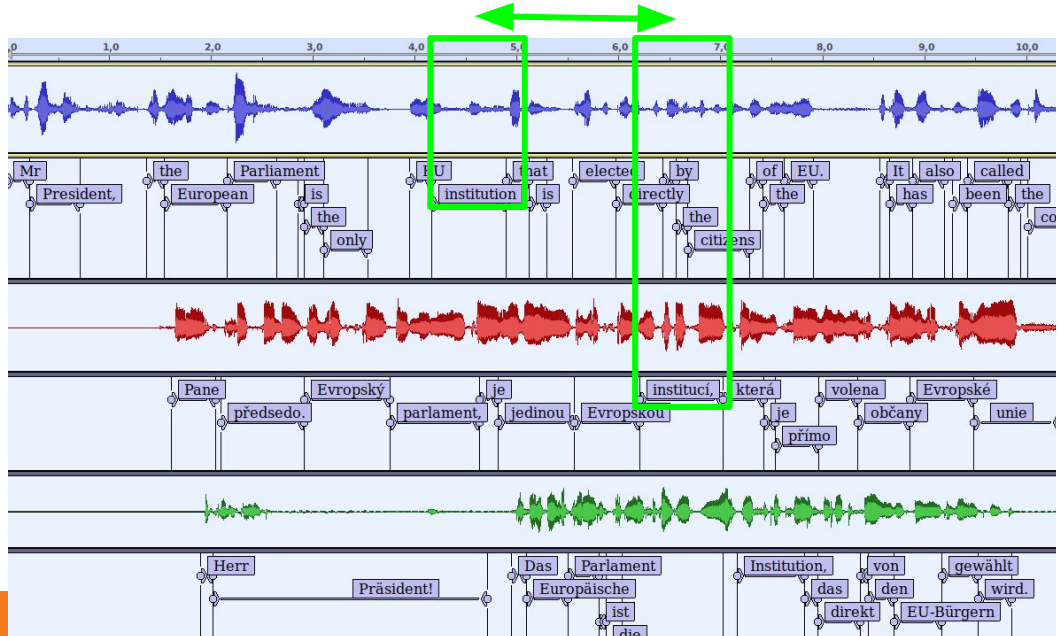
-> English-to-Czech Sim. Interpreting

-> English-to-German Sim. Intp.

Simultaneous speech translation (recap 11/4/2024)

- **Simultaneous** = Live = Real-Time = Low-latency = Incremental
  - Source available continuously, one **chunk** at a time
  - The **chunk** can be:
    - audio segment … in the direct speech-to-text translation or <u>transcription = ASR</u>
    - or word (text) produced by incremental ASR … in a cascaded system = ASR + **MT**
  - Provide the target "<u>at the same time</u>" as the source is being produced

    <u>= simultaneously</u> = with a small additive **delay**



In the European Parliament:

-> English original source

-> English-to-Czech Sim. Interpreting

-> English-to-German Sim. Intp.

Simultaneous approaches

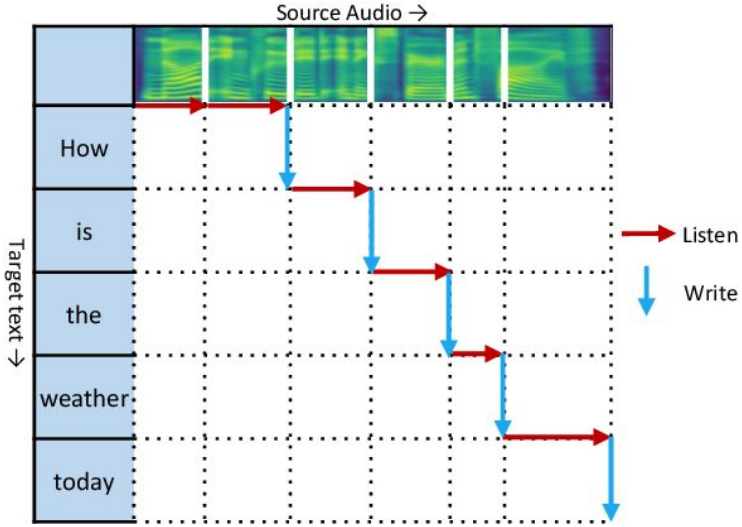
Re-Translation vs. Incremental

- Re-translate from beginning of sentence each time: **rewrite + append**

Source	Output								Erasure
1: Neue	New								-
2: Arzneimittel	New	Medicines							0
3: könnten	New	Medicines							0
4: Lungen-	New	drugs	may	be	lung				1
5: und	New	drugs	could	be	lung	and			3
6: Eierstockkrebs	New	drugs	may	be	lung	and	ovarian	cancer	4
7: verlangsamen	New	drugs	may	slow	lung	and	ovarian	cancer	5
Content Delay	1	4	6	7	7	7	7	7	

Source: [Arivazhagan et al., 2020]

- Alternates between reading from ASR and translating: **no rewrites, only append**



Source: [Ren et al., 2020]

Re-Translation vs. Incremental

	Re-translation	Incremental
Latency	👍	👎
Quality	👍 as in offline mode	👎 lower than in offline mode
Stability (flicker)	👎 may be unreadable	👍 only appends to the end
Presentation space	Needs large space 👎	Suitable for 2-line subtitles 👍
Versatility	Can serve many users at once 👍 Medium src. lang. proficiency … needs fast outputs No src. Lang. proficiency … slower OK, high quality needed	Compromise 👎 latency and quality
Easy to plug-in any offline model?	👍 YES. May be unstable on prefixes but high quality in the end.	Not easy but possible.

Stability in Re-Translation

How to make re-translation more stable?

Baseline (“standard” offline MT)

Source	Output								Erasure
1: Neue	New								-
2: Arzneimittel	New	Medicines							0
3: könnten	New	Medicines							0
4: Lungen-	New	drugs	may	be	lung				1
5: und	New	drugs	could	be	lung	and			3
6: Eierstockkrebs	New	drugs	may	be	lung	and	ovarian	cancer	4
7: verlangsamen	New	drugs	may	slow	lung	and	ovarian	cancer	5

Stability measure: **13** erasures for **8** generated tokens = **1.625**

Example: [Arivazhagan et al., 2020].

How to make re-translation more stable?

Baseline (“standard” offline MT)

Source	Output								Erasure
1: Neue	New								-
2: Arzneimittel	New	Medicines							0
3: könnten	New	Medicines							0
4: Lungen-	New	drugs	may	be	lung				1
5: und	New	drugs	could	be	lung	and			3
6: Eierstockkrebs	New	drugs	may	be	lung	and	ovarian	cancer	4
7: verlangsamen	New	drugs	may	slow	lung	and	ovarian	cancer	5

Stability measure: **13** erasures for **8** generated tokens = **1.625**

Improvement

Source		Erasure
1: Neue	New	-
2: Arzneimittel	New drugs	0
3: könnten	New drugs may	0
4: Lungen-	New drugs may lung	0
5: und	New drugs may lung and	0
6: Eierstockkrebs	New drugs may lung and ovarian cancer	0
7: verlangsamen	New drugs may slow lung and ovarian cancer	5

4 erasures for **8** generated tokens = **0.5**

We learned: Proportional prefix training

Full	Source	Die Führungskräfte der Republikaner rechtfertigen ihre Politik mit der Notwendigkeit , den Wahlbetrug zu bekämpfen [*15 tokens*]
	Target	Republican leaders justified their policy by the need to combat electoral fraud [*12 tokens*]
Prefix	Source	Die Führungskräfte der Republikaner rechtfertigen [*5 tokens*]
	Target	Republican leaders justified their [*4 tokens*]

Table 2: An example of proportional prefix training. Each example in the minibatch has a 50% chance to be truncated, in which case, we truncate its source and target to a randomly-selected fraction of their original lengths, 1/3 in this example. No effort is made to ensure that the two halves of the prefix pair are semantically equivalent.

1. Train a standard offline MT
2. Finetune on 1:1 mix of full sent. pairs and src-target prefixes
3. Create the prefixes from the length proportion,
4. do not care about the parallel words in the truncated suffix => anticipation
5. Measure the MT quality and erasures
6. Select a suitable trade-off

Example: [Arivazhagan et al., 2020]. The method first proposed by Niehues et al., 2018.

Demo time

Demo: Notice the flicker

- From our paper:
  Presenting Simultaneous Translation in Limited Space, Macháček and Bojar, 2020

- How to present re-translation? **Learn this:**
  - If possible, **use large text output** size, e.g. paragraph view.
    - The end users choose where to look
  - When only few lines are possible:
    - Re-writes can happen in the scrolled away part
    - => **ASR** re-writes usually small ... **OK for few lines**
    - => **MT** re-writes too large
      ... too much flicker or large presentation delay, **not usable for few lines**

ELITR demo

- On your device: quest.ms.mff.cuni.cz/elitr/demo/ or QR code →
- Lang. options:
  - **EN** = any lang. translation … Whisper-Streaming = voice activity controller + auto lang. detection +
  - **ASR** = any lang. ASR + W. large-v3 + long-form streaming with LocalAgreement-2, 1 sec. min. chunk size
  - Others* = En -> 44 langs. **MT** with re-translation … UEDIN English-all rainbow model, 44 l. variants, 12-layer enc., 2021-12-10, v4
  - *ZH = now blank, ready to connect another MT
- Wrapped in the **ELITR** pipeline (HE project 2019-2022, CUNI+UEDIN+KIT+others)

80. That was Josef Pazderka 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. conference, Ukraine as a Shared Responsibility.
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,

59. republiky, Petra Pavla, aby přiseł 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. Prosím, pane prezidente.
62. Dobrý den, dámy a pánové, vážení hosté zde v sále, posluchači, ale také diváci na ostatních platformách.
63. Když mě vedení Českého rozhlasu požádalo o záštitu nad dnešní konferenci, nemusel jsem dlouho váhat, protože témata, kterými se tady zabýváme, jsou pro mě velice důležitá.
64. Připomínáme si 100. výročí odzahájení 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 omezována 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, sídli i Radio Sobotná Evropa.
71. Jehož tři novináři jsou dnes vězněni. – Jihard Losik a Andrej Kuznečík v Bělorusku a Vladislav Jesipenko na ruském okupovaném Krymu.
72. V únoru tohoto roku jsme si připomněli

79. Is-Sur President.
80. Filgħodu tajjeb, nisa u mara, mistiednija hawn u dawk li jisimgħu u l-ispetturi fuq il-
81. Meta ntalabni mir-radju Ċeka biex tieħu l-awditi ta' dan l-avveniment, ma stajtx għal sekonda, minħabba li s-suġġetti li qed niddiskutu hawn llum huma importanti ħafna għalija.
82. 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 mingħajr ċensura, mingħajr limitazzjonijiet. il-libertà li jaċċettaw informazzjoni, li jfittxu informazzjoni, li jinfirxu informazzjoni, il-libertà li f'ħafna partijiet tad-dinja hija ristretta ħafna b'saħħitha, u libertà... in-n
87. U eżempji speċifiċi mhumiex bogħod.
88. Aħna għandna fostna t-tifla ta' Boris Nemtsov, il-politika ta' l-oppożizzjoni Russa maqtula, Zhanna Nemts
89. Fuq it-triq ta' Vyhorodska, qrib ħafna mill-kwartieri ġenerali tar-Radju Ċeka, hemm ir-Radju Ħielsa ta' l-Ewropa, u

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 цієї 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 річниць від початку трансляції Чеського радіо.
76. І це також нагадування про важливість свободи слову.
77. Свободи висловлювати свою думку без обмежень.
78. Свободу приймати інформацію та думки, шукати їх та поширювати.
79. Свободу, яка у різних частинах світу досі підається 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 Жанна Німцова.
83. У вулиці Виноградська, 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

74. للتأكيد على نوعية المعلومات التي نأتي إلى الجمهور التشيكي. من المهم ما نوع المعلومات التي نستهلكها.
75. وثانياً، لم نرغب، بعد هذا أكثر من سنة الصراع، لم نرغب في أن يُنظر إلى هذا كألعاب فيديو.
76. بعض الحركات على الخريطة
77. وما زال هناك مصير فرادى الناس، معاناة فرادى ما يحدث.
78. طوال هذه الألواح، طوال اليوم، يجب أن نكون قادراً على رؤية على الأقل نظرة على ذلك.
79. وآمل أن نكون تجربة مثيرة للاهتمام وقوية.
80. كان ذلك (جوزيف بازدركا) من الراديو التشيكي (ب
81. والآن، رئيس الجمهورية التشيكية، بيتر بابل.
82. رجاءً تعال إلى هنا على المسرح. وعرض خطابك الافتتاحي لبدء الدورة الأولى لهذا المؤتمر. المؤتمر، أوكرانيا كمسؤولية مشتركة.
83. سيدي الرئيس
84. صباح الخير يا سيداتي وسادة ضيوف هنا والمستمعين والمشاهدين على المنبرات الأخرى
85. وعندما طلبت من الإذاعة التشيكية أن تتولى رعاية هذا الحدث، لم أتردد لمدة ثانية، لأن المواضيع التي نناقشها اليوم هامة جدا بالنسبة لي.
86. هذا هو الذكرى السنوية المئوية منذ بداية البث المنتظم للإذاعة التشيكية، التي تُخبرنا أيضاً بأهمية حرية التعبير، والتحدث بدون رقابة، دون قيود. حرية قبول المعلومات، والبحث عن المعلومات، ونشر المعلومات، والحرية التي تُقيَّد في العديد من أنحاء العالم بقوة جداً، والحرية... الناس يستمرون في إعطاء حياتهم من أجلها.
87. وهناك أمثلة محددة ليست بعيدة عن ذلك.
88. لدينا بيننا ابنة بوريس نامتسوف، سياسة المعارضة الروسية المقتلة، زانا نامتسوفا.
89. في شارع فيهوروديسكا، قريب جداً من مقر الإذاعة التشيكية، هناك إذاعة أوروبا الحرة. وثلاثة من صحفيها في السجن الآن.

Incremental

Problem: When to wait or translate

Chinese source:

jǐngfāng	xiàzhōu	jiāng	duì	bù fèn	shè àn	rén yuán	tí qǐ gōng sù
警方	下周	将	对	部分	涉案	人员	提起 公诉
police	next week	will	for	part	involved	people	**accuse**

Simultaneous intp.:

Next week, police

will

accuse some of the people involved in the case.

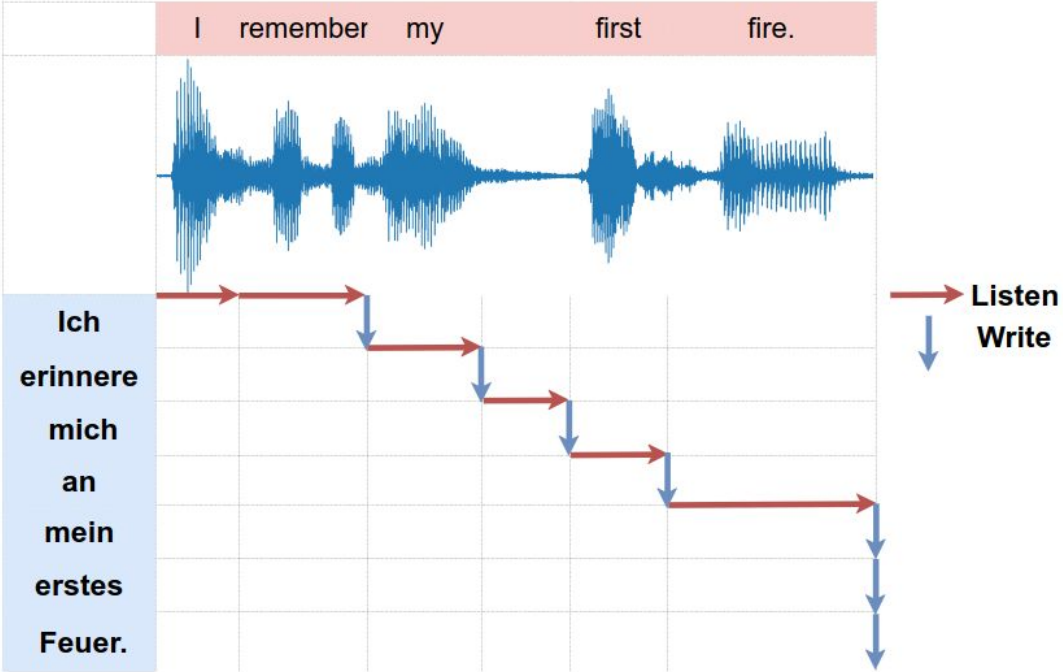
Source: https://aclanthology.org/2020.emnlp-main.178.pdf

Simultaneous policies

Sim. **policy** for an incremental input:

Given a seq-to-seq model, make continuous predictions.

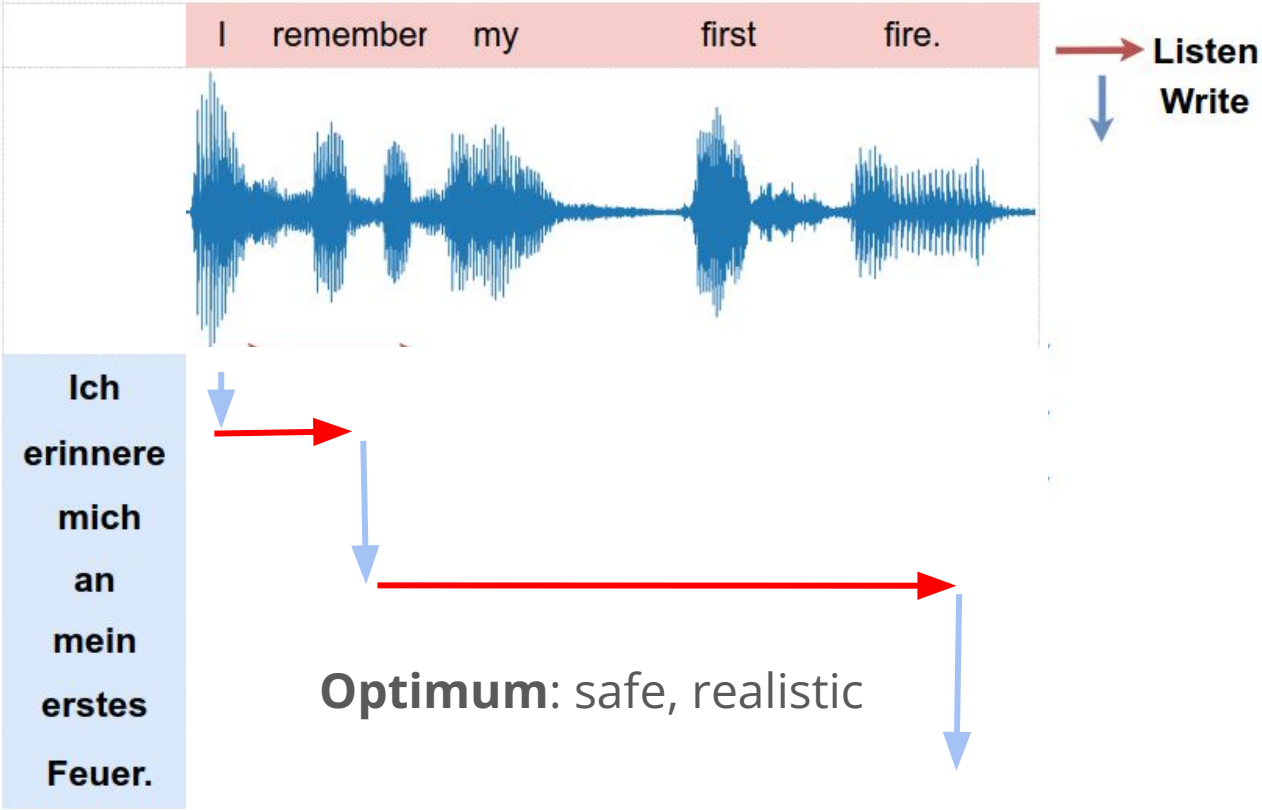
Objectives: latency and quality. Control them by one parameter.



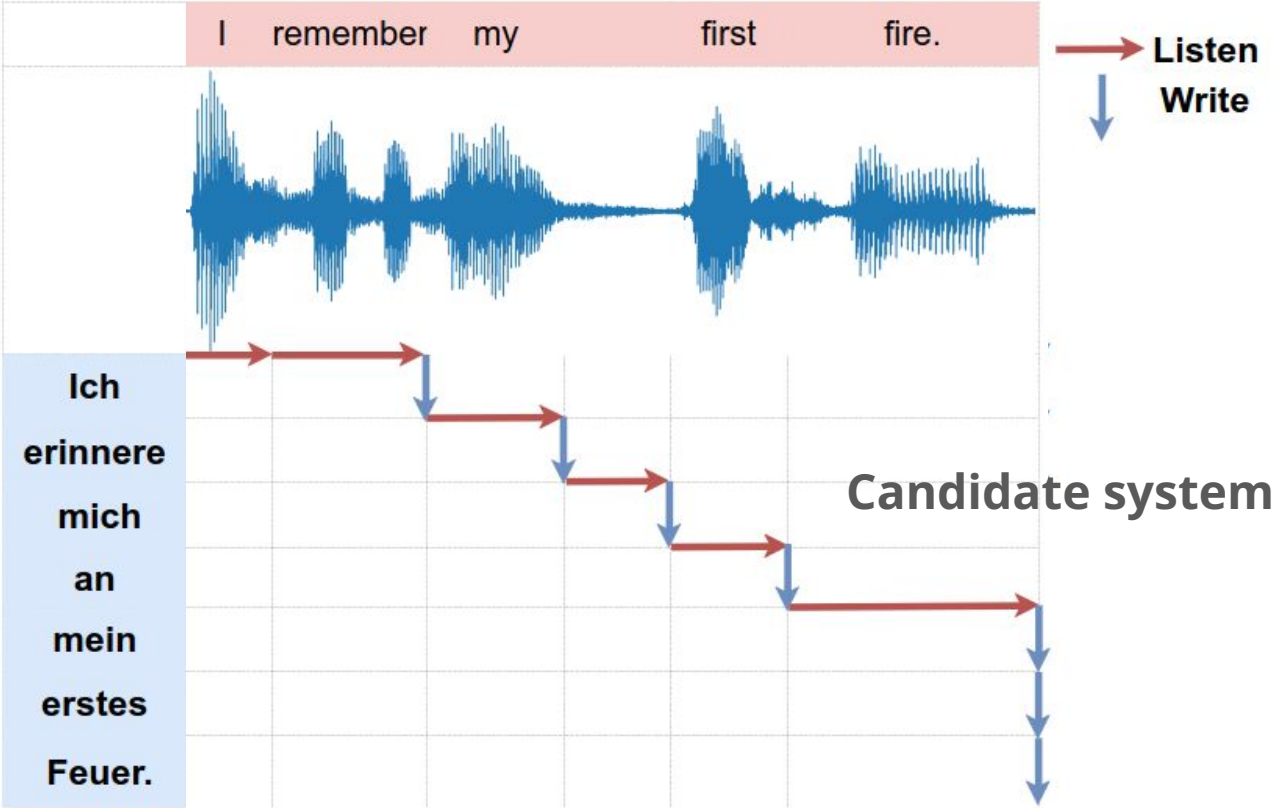
Simultaneous policies

Simultaneous policies

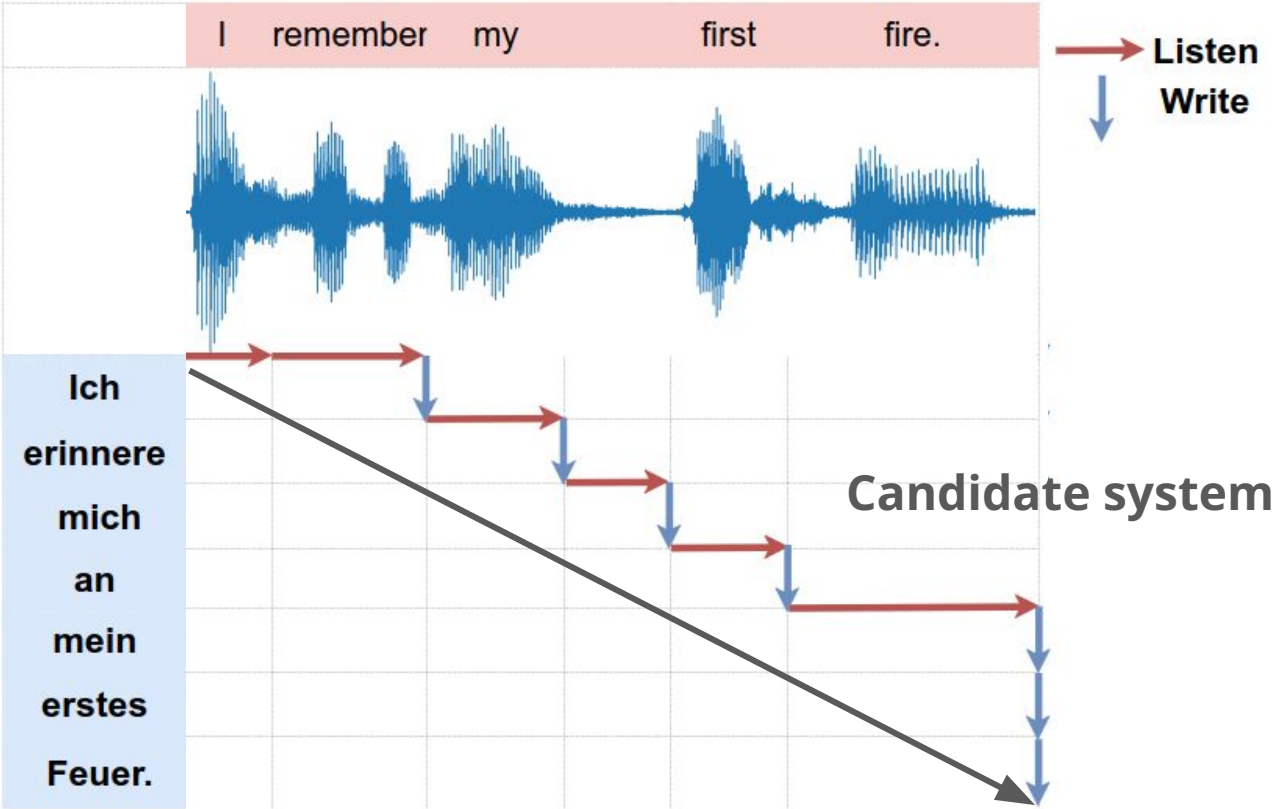
Simultaneous policies



Simultaneous policies



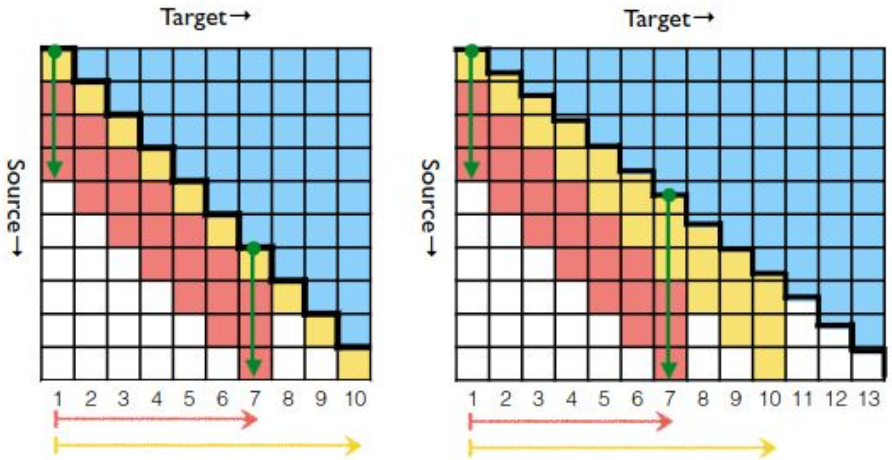
Simultaneous policies



Diagonal: Latency reference

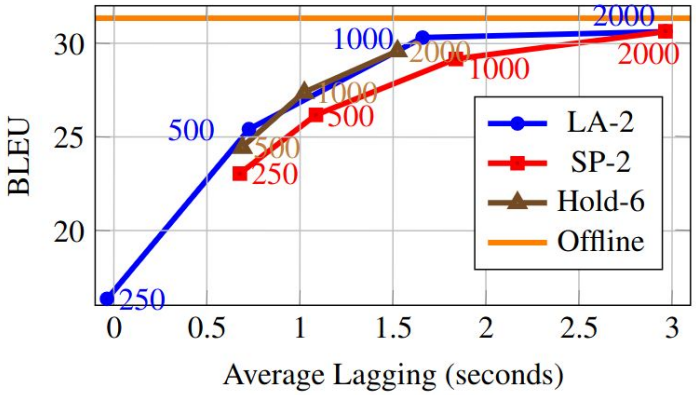
Simultaneous Policies

- **Wait-k**: Be always **k** words behind.
  (+ catch-up factor because src-tgt length ratio.)
  - Simple baseline
  - Small k might be OK for some lang. pairs
  - Not self-adaptable by lang. or content



Simultaneous Policies

- LocalAgreement-N (Polák et al., 2022):
  - With every new src chunk, decode internally.
  - Output **Agreement** = common prefix of the last **N** chunks.
  - Use it for forced decoding the next chunks.
  - => **Self-adaptable** by lang. and source complexity
  - Parameters: N, the chunk size (e.g. x-times 330ms = appx. x words)
- LocalAgreement-2 (Liu et al., 2020):
  - Best in IWSLT 2022 (Polák et al., 2022)
  - Min. latency 2x chunk size, max unlimited
  - In Whisper-Streaming



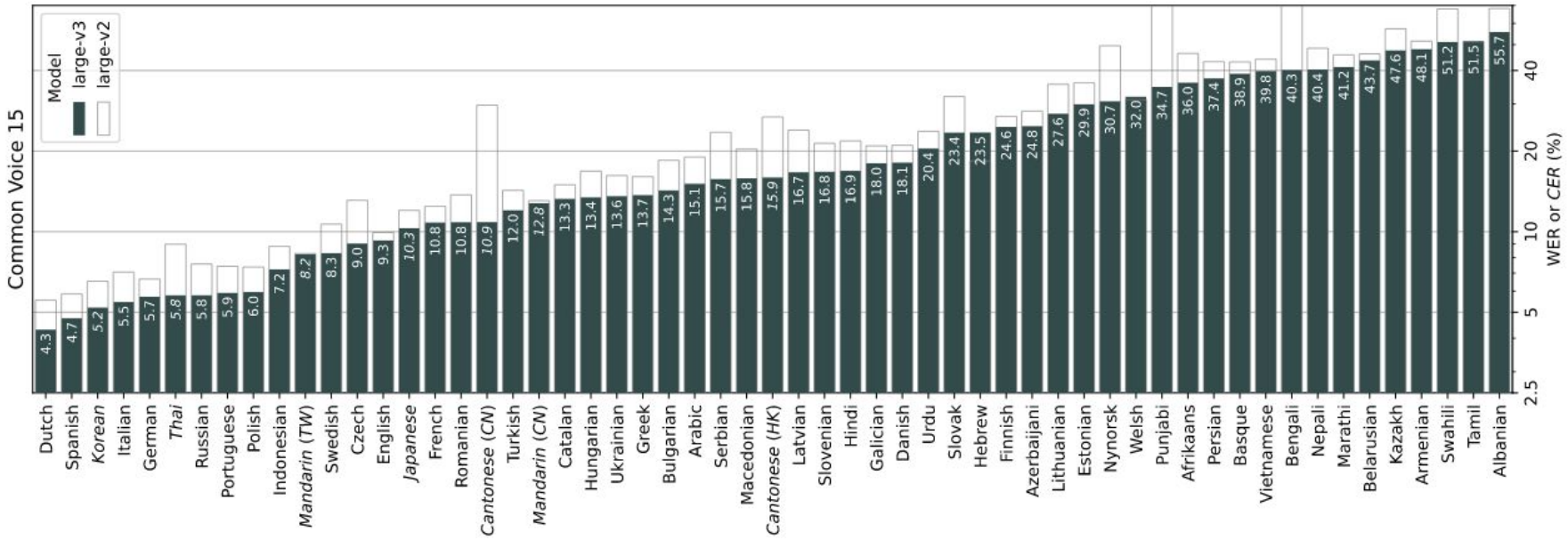
Whisper-Streaming

github.com/ufal/whisper_streaming

- Practical tool, demonstrates SoTA
- Includes:
  - Voice activity controller = end after 0.5s silence
  - + automatic language detection (99 langs)
  - + Whisper, with fast implementation
  - + LocalAgreement-2
  - + heuristics for continuous speech
  - Wrapped in ELITR pipeline

Whisper Languages

- Let’s chat in 99 langs.



List of Whisper’s languages

Afrikaans Albanian Amharic Arabic Armenian Assamese Azerbaijani Bashkir Basque Belarusian Bengali Bosnian Breton Bulgarian Burmese Cantonese Castilian Catalan Chinese Croatian Czech Danish Dutch English Estonian Faroese Finnish Flemish French Galician Georgian German Greek Gujarati Haitian Haitian Creole Hausa Hawaiian Hebrew Hindi Hungarian Icelandic Indonesian Italian Japanese Javanese Kannada Kazakh Khmer Korean Lao Latin Latvian Letzeburgesch Lingala Lithuanian Luxembourgish Macedonian Malagasy Malay Malayalam Maltese Maori Marathi Moldavian Moldovan Mongolian Myanmar Nepali Norwegian Nynorsk Occitan Panjabi Pashto Persian Polish Portuguese Punjabi Pushto Romanian Russian Sanskrit Serbian Shona Sindhi Sinhala Sinhalese Slovak Slovenian Somali Spanish Sundanese Swahili Swedish Tagalog Tajik Tamil Tatar **Telugu** Thai Tibetan Turkish Turkmen Ukrainian Urdu Uzbek Valencian Vietnamese Welsh Yiddish Yoruba

Summary

- You learned how to represent speech
- How to leverage unlabeled speech data
- How to integrate speech into LLMs
- Simultaneous methods and policies