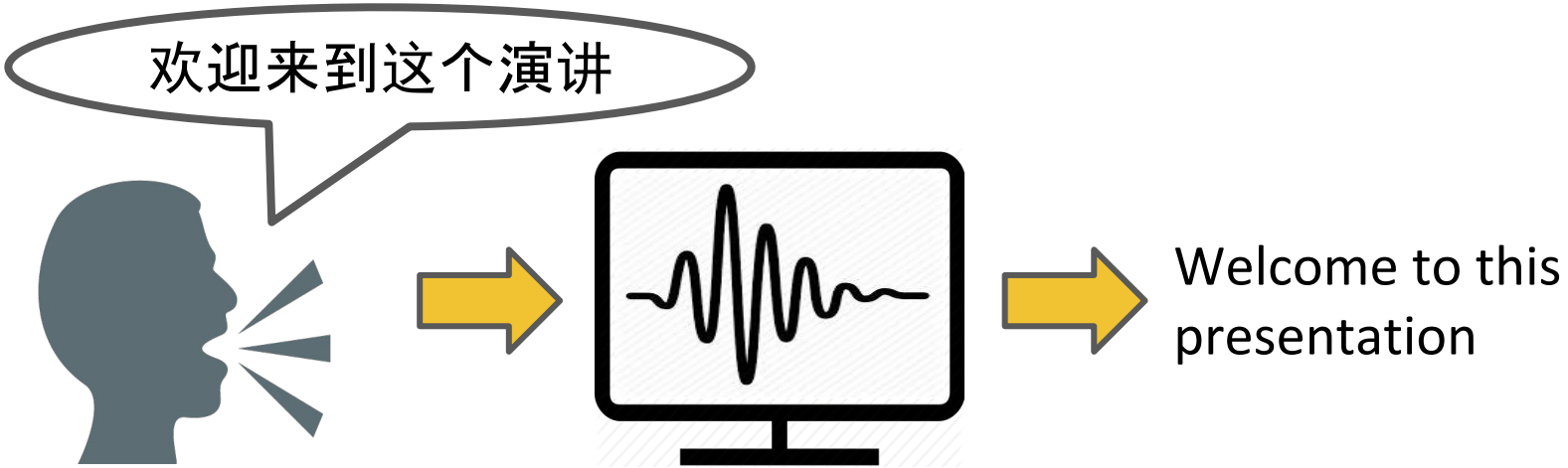
Recent Developments in Direct Speech Translation

Mattia A. Di Gangi, Marco Gaido, Matteo Negri, Roldano Cattoni
Luisa Bentivogli, Beatrice Savoldi, Marco Turchi
University of Trento & Fondazione Bruno Kessler

UPC Seminars
July 15 2020
Online

Speech-to-Text Translation

Speech translation is the process by which speech in one language is automatically translated into text of a second language.



Outline

- **Introduction** to speech translation
- **MuST-C** a Multilingual Speech Translation Corpus
- **S-Transformer** for speech translation + **Multilinguality**
- **IWSLT 2020** shared task
- **MuST-SHE** a test set for evaluating gender bias
- **Conclusions** and future directions

INTRODUCTION

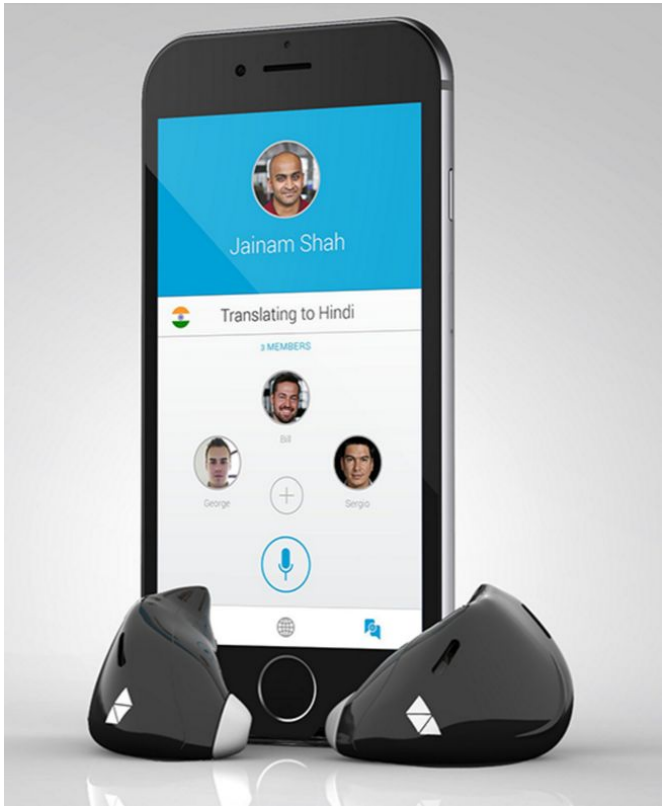
MOTIVATION

CHALLENGES

BACKGROUND

Motivation - Real-Time Communication

Motivation - Real-Time Communication



Motivation - Live Broadcasting

Motivation - Human-Machine Communication

Пожалуйста,
включи
музыку

Please,
play music

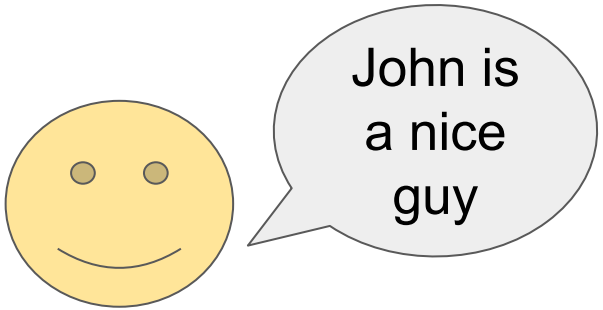
INTRODUCTION

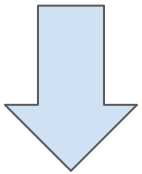
MOTIVATION

CHALLENGES

BACKGROUND

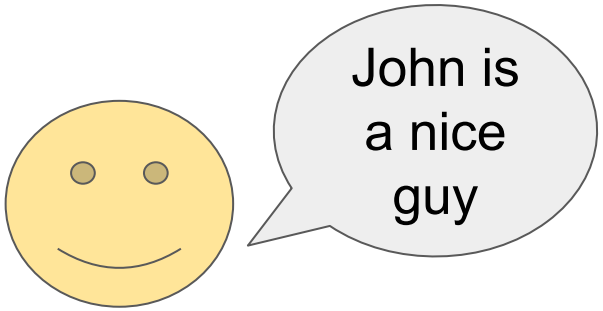
Challenges - Recognition Errors





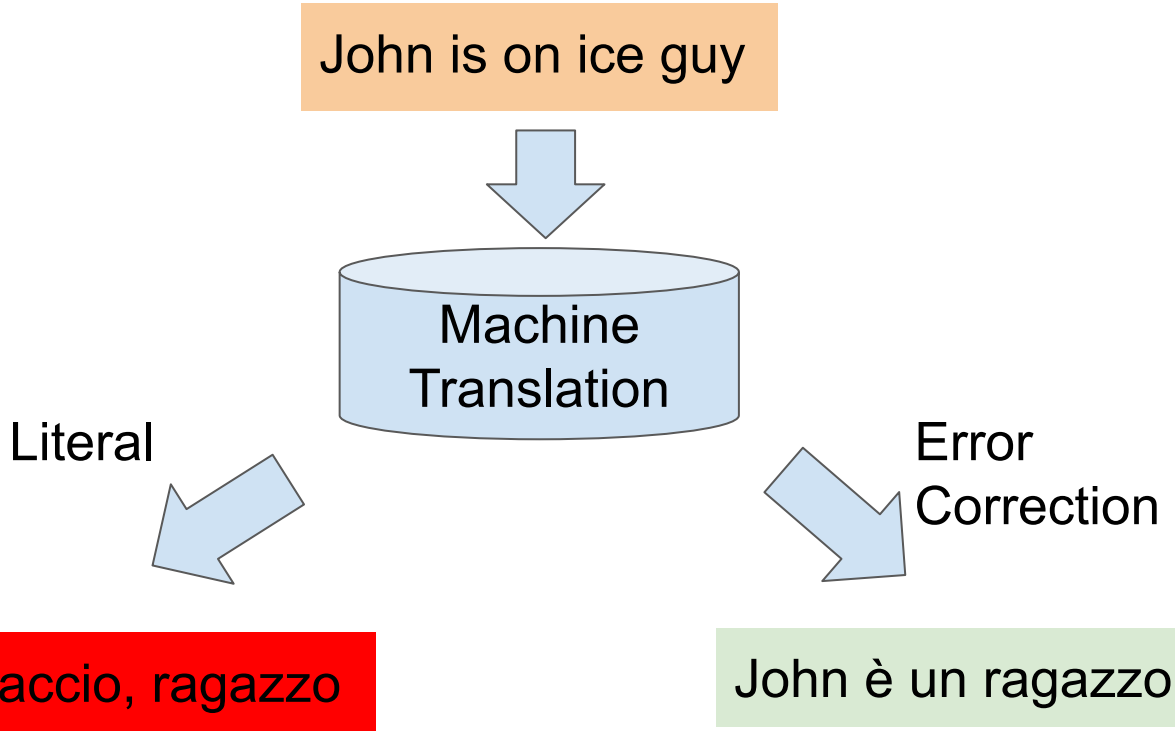
John is on ice guy

Challenges - Recognition Errors

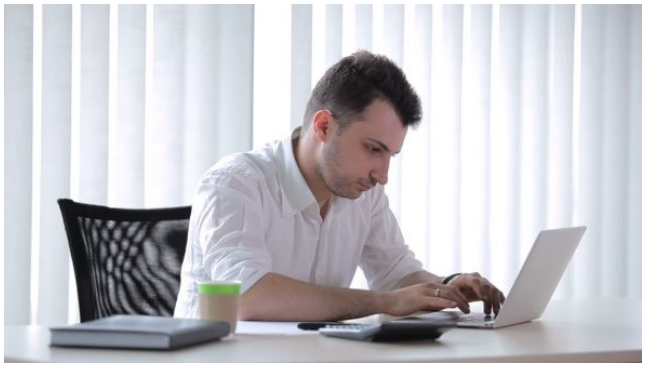


John is **on ice** guy

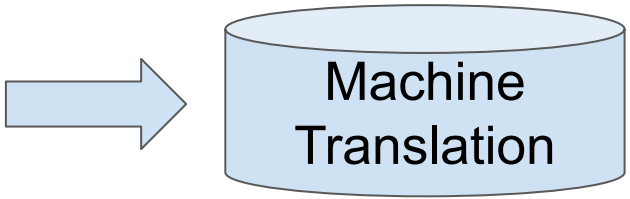
Challenges - Error Propagation



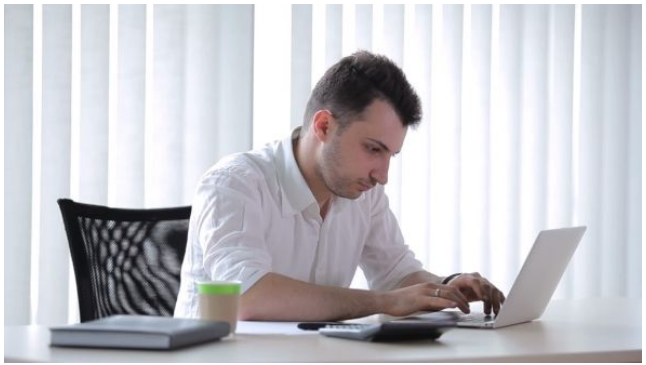
Challenges - Translation Errors



Work completed with only minor problems.



Challenges - Translation Errors

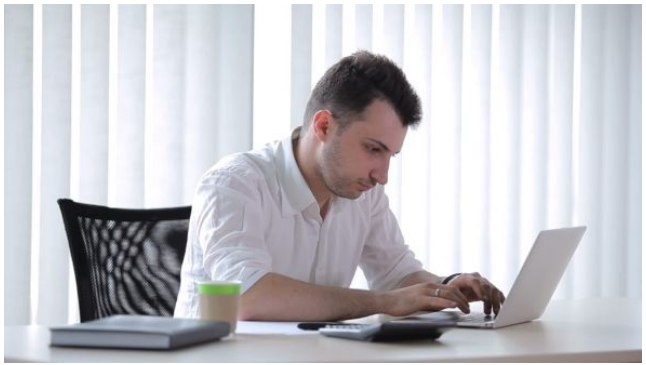


Il lavoro ha completato solo i problemi minori.

Work completed with only minor problems.

Machine Translation

Challenges - Translation Errors



Il lavoro ha completato solo i problemi minori.

Work completed with only minor problems.

Machine Translation

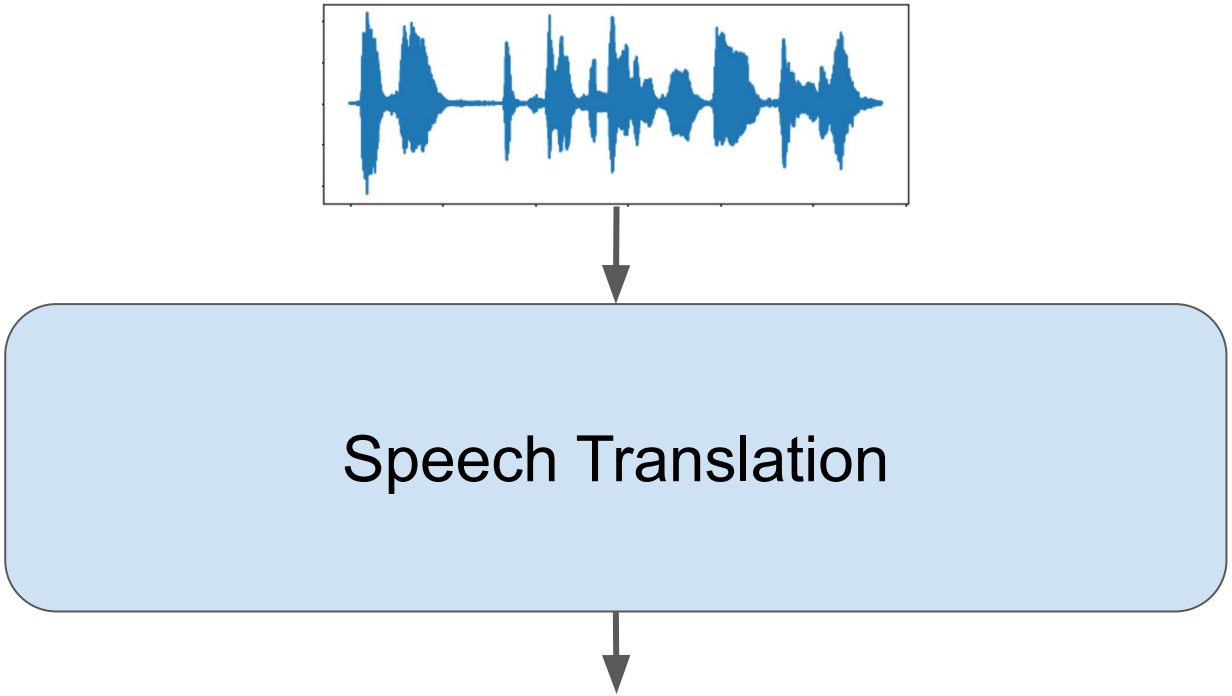
INTRODUCTION

MOTIVATION

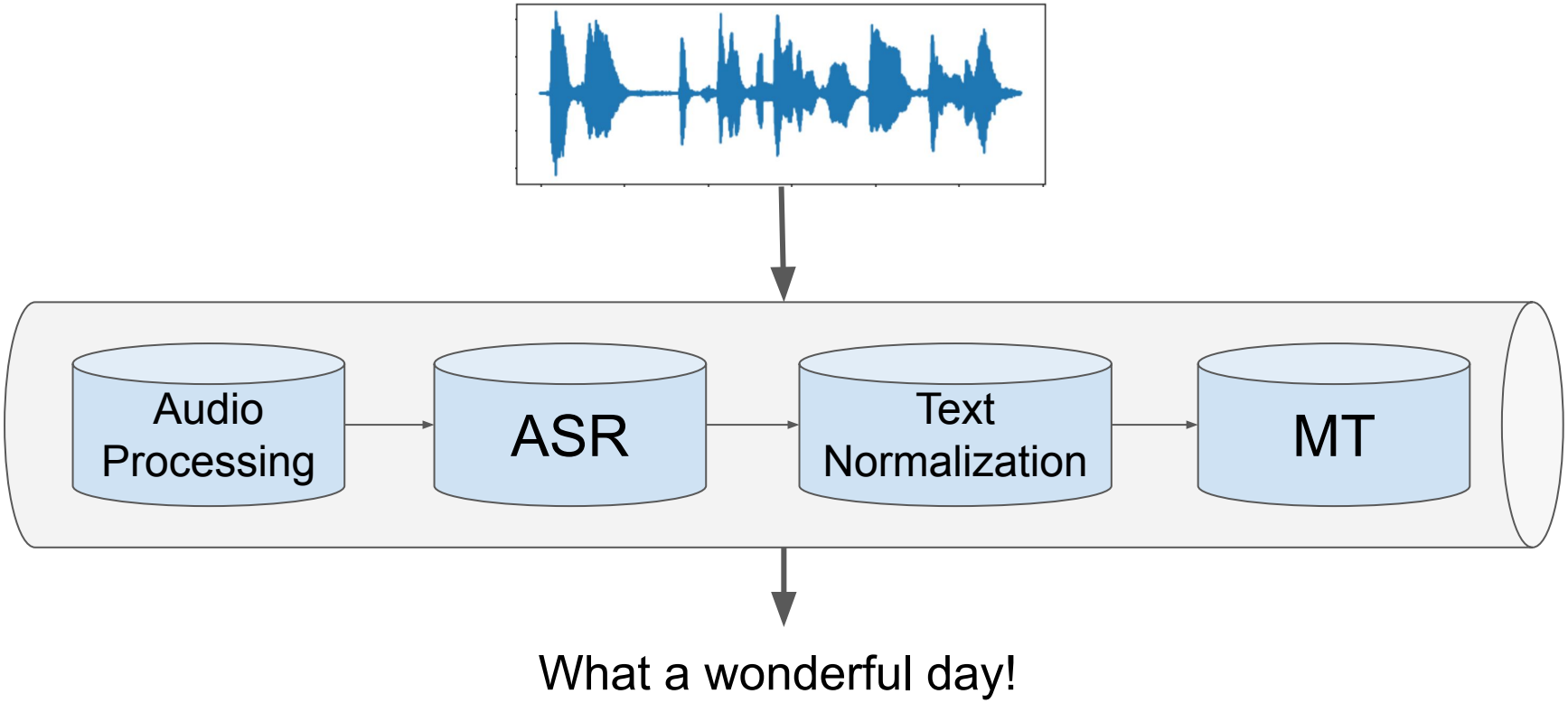
CHALLENGES

BACKGROUND

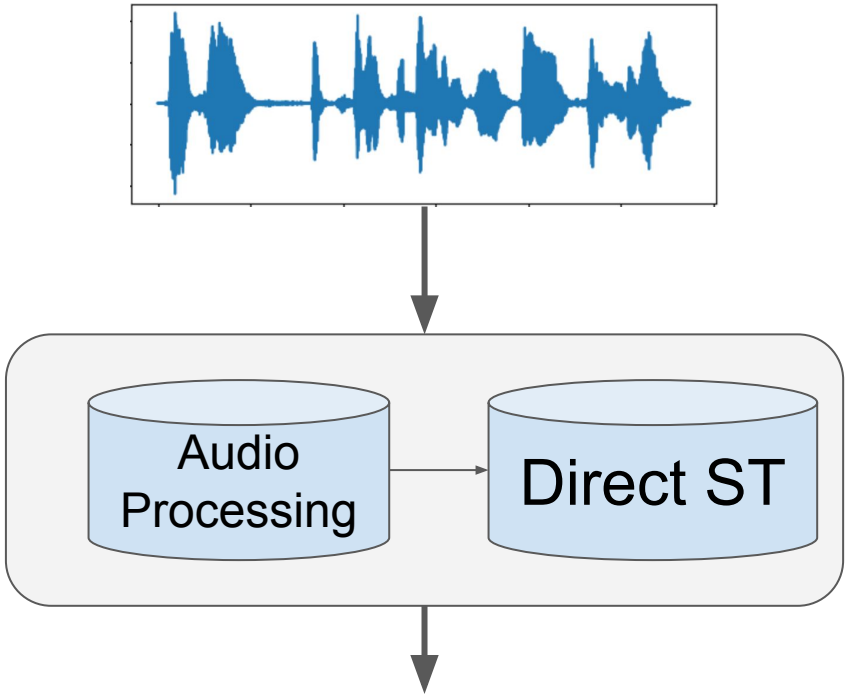
Background: Speech Translation



Background: Speech Translation



Background: Speech Translation



DIRECT SPEECH TRANSLATION

MUST-C

S-TRANSFORMER

MULTILINGUAL

Existing corpora in 2018 (>30 hours)

Corpus	Languages	Hours of speech
Sanabria et al. (2018)	En →Pt	300
Niehues et al. (2018)	En → De	273
Kocabiyikoglu et al. (2018)	En → Fr	236
Tohyama et al. (2005)	En ←→ Jp	182
Paulik and Waibel (2009)	En → Es	111
	Es → En	105
Post et al. (2013)	Es → En	38
Stüker et al. (2012)	En → Es	37

- Few languages
- Small size
- Specific domains

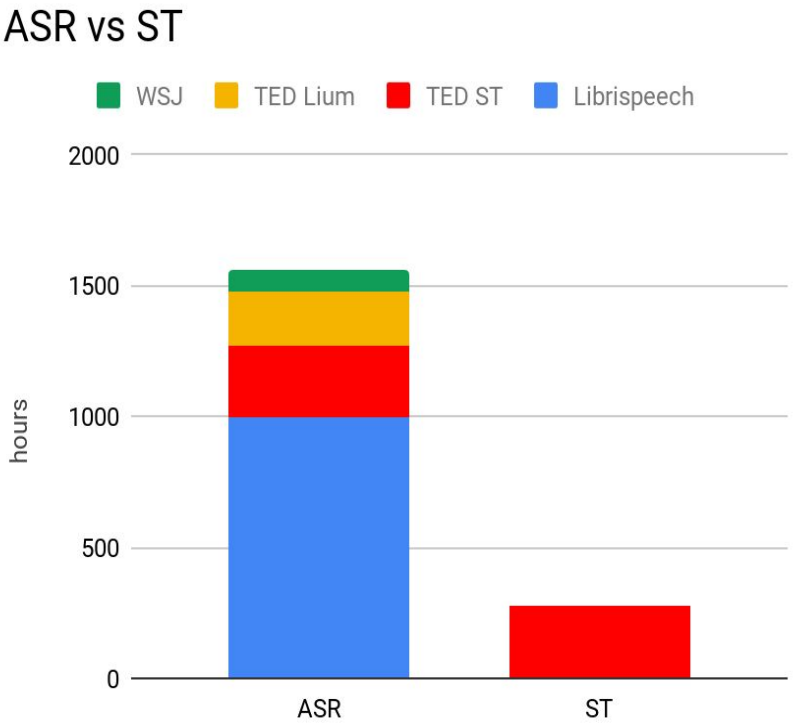
MuST-C: a **Mu**ltilingual **S**peech **T**ranslation **C**orpus

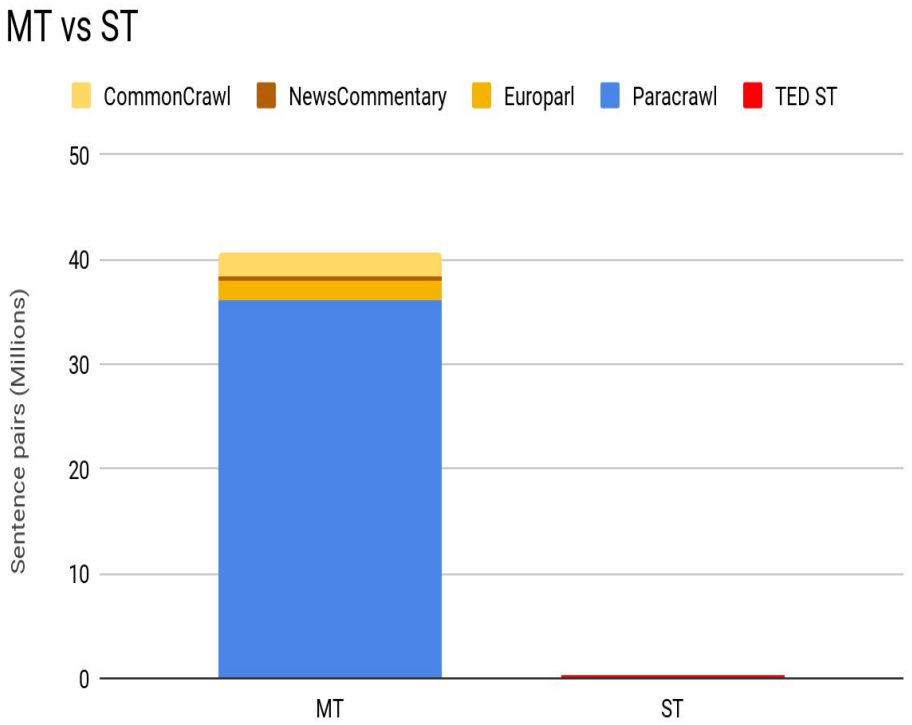
RQ: Can we build automatically a large and reliable ST corpus?

1. A large multilingual corpus
   - built from English TED talks
   - 8 target languages: Es, De, Fr, It, Nl, Pt, Ro, Ru
   - 385 to 504 hours of speech (> any existing resource)
2. With empirical and manual quality assessment
3. Train, dev and test split
4. And baselines ASR, MT and ST results

Di Gangi, M. A., et al. (2019). "MuST-C: a **Mu**ltilingual **S**peech **T**ranslation **C**orpus." *Proceedings of HLT-NAACL 2019.*

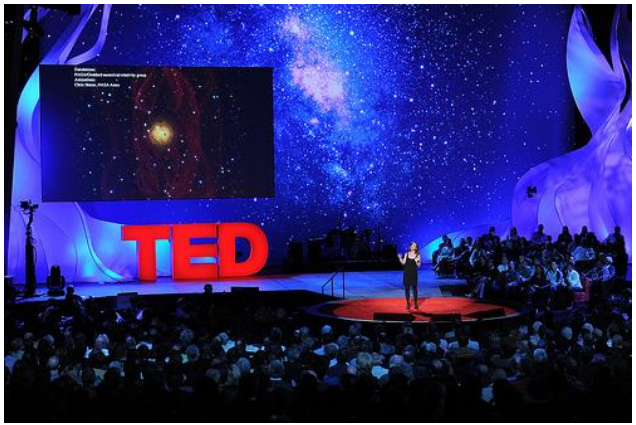
Example: data available for En→De





Ted Talks

TED
IDEAS WORTH SPREADING



- > 3,100 English talks as of May 2019
- High-quality speech and recordings
- High topic/speaker variability
- High-quality transcripts by volunteers
- > 140K translations in 116 languages
- Numbers are constantly growing...

Corpus creation pipeline

1. Align texts in English and target language
    - Tool: Gargantua (`http://github.com/braunefe/Gargantua`)
2. Align English text and audio
    - Tool: Gentle (`http://github.com/lowerquality/gentle`)
3. Filter unreliable alignments

Statistics

Target language	#Talks	#Sentences	Hours of speech	Source words	Target words
Dutch	2,267	253K	442	4.7M	4.3M
French	2,510	280K	492	5.2M	5.4M
German	2,093	234K	408	4.3M	4.0M
Italian	2,374	258K	465	4.9M	4.6M
Portuguese	2,050	211K	385	4.0M	3.8M
Romanian	2,216	240K	432	4.6M	4.3M
Russian	2,498	270K	489	5.1M	4.3M
Spanish	2,564	270K	504	5.3M	5.1M

Statistics

Target language	#Talks	#Sentences	Hours of speech	Source words	Target words
Dutch	2,267	253K	442	4.7M	4.3M
French	2,510	280K	492	5.2M	5.4M
German	2,093	234K	408	4.3M	4.0M
Italian	2,374	258K	465	4.9M	4.6M
Portuguese	2,050	211K	385	4.0M	3.8M
Romanian	2,216	240K	432	4.6M	4.3M
Russian	2,498	270K	489	5.1M	4.3M
Spanish	2,564	270K	504	5.3M	5.1M

Thousands of talks = thousands of topics and thousands of speakers!

Statistics

Target language	#Talks	#Sentences	Hours of speech	Source words	Target words
Dutch	2,267	253K	442	4.7M	4.3M
French	2,510	280K	492	5.2M	5.4M
German	2,093	234K	408	4.3M	4.0M
Italian	2,374	258K	465	4.9M	4.6M
Portuguese	2,050	211K	385	4.0M	3.8M
Romanian	2,216	240K	432	4.6M	4.3M
Russian	2,498	270K	489	5.1M	4.3M
Spanish	2,564	270K	504	5.3M	5.1M

Larger than any existing SLT resource (452 hours on average!)

Statistics

Target language	#Talks	#Sentences	Hours of speech	Source words	Target words
Dutch	2,267	253K	442	4.7M	4.3M
French	2,510	280K	492	5.2M	5.4M
German	2,093	234K	408	4.3M	4.0M
Italian	2,374	258K	465	4.9M	4.6M
Portuguese	2,050	211K	385	4.0M	3.8M
Romanian	2,216	240K	432	4.6M	4.3M
Russian	2,498	270K	489	5.1M	4.3M
Spanish	2,564	270K	504	5.3M	5.1M

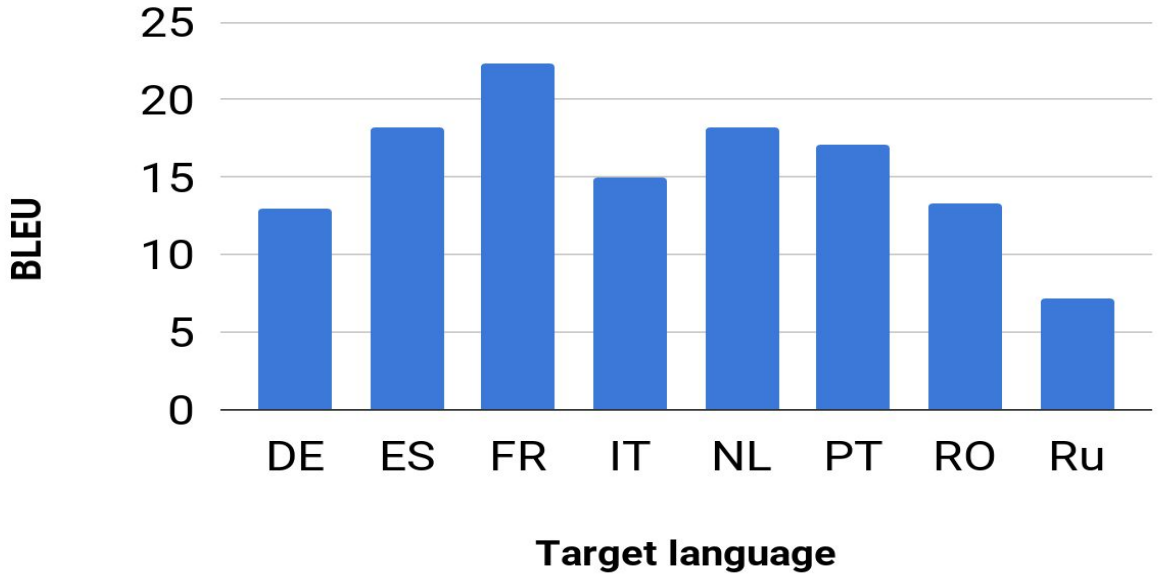
Millions of transcribed/translated words = useful also for ASR and MT!

Baseline results (8 languages)

- **8 Dev sets**: 1.4K sentences from common talks
- **8 Test set**: 2.5K sentences from common talks
- **8 Training sets**: the remaining material

NOTE FOR REPLICABILITY: These splits are preserved in the released corpus

Baseline ST results



Examples (EN-IT)

SRC: In 2010, it was viewed 23 million times. This is a chart of what it looked like when it first became popular last summer.

REF: Nel 2010, è stato visto 23 milioni di volte. Questo è un grafico delle visualizzazioni quando è diventato popolare l'estate scorsa.

ST: Nel 2010, era voi 23 milioni di volte. Perché il grafico sembrava che quando all' inizio diventava popolare, l' estate scorsa .

Examples (EN-IT)

SRC: In 2010, it was viewed 23 million times. This is a chart of what it looked like when it first became popular last summer.

REF: Nel 2010, è stato visto 23 milioni di volte. Questo è un grafico delle visualizzazioni quando è diventato popolare l'estate scorsa.

ST: **Nel 2010**, era voi **23 milioni di volte**. Perché il **grafico** sembrava che quando all'inizio diventava **popolare**, **l' estate scorsa.**

Impact of MuST-C

- Training set for IWSLT 2019-2020
- Enabled experiments on larger ST data
- Fostered research in the area
- New datasets followed:
  - MuST-Cinema
  - MuST-SHE
  - MuST-C++

Impact of MuST-C

- Training set for IWSLT 2019-2020
- Enabled experiments on larger ST data
- Fostered research in the area
- New datasets followed:
  - **MuST-Cinema**
  - MuST-SHE
  - MuST-C++

Translation of subtitles

Impact of MuST-C

- Training set for IWSLT 2019-2020
- Enabled experiments on larger ST data
- Fostered research in the area
- New datasets followed:
  - MuST-Cinema
  - **MuST-SHE**
  - MuST-C++

Test set for gender phenomena

Impact of MuST-C

- Training set for IWSLT 2019-2020
- Enabled experiments on larger ST data
- Fostered research in the area
- New datasets followed:
  - MuST-Cinema
  - MuST-SHE
  - ~~MuST-C++~~

Sorry, programmer's joke

DIRECT SPEECH TRANSLATION

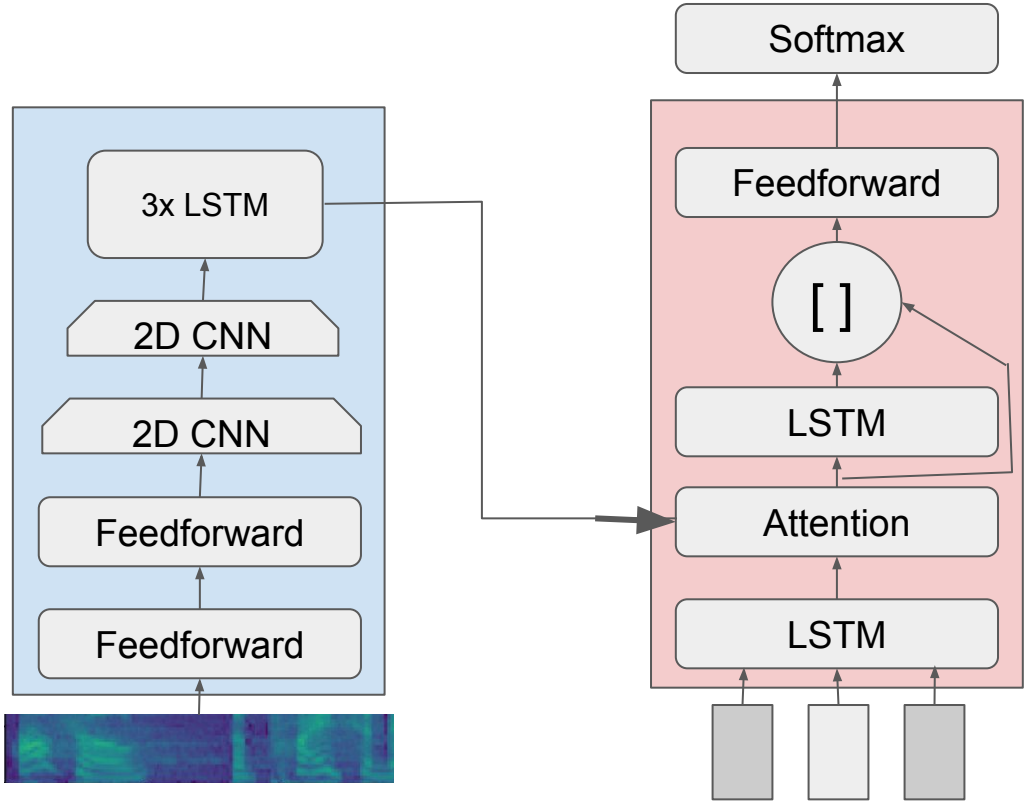
MUST-C

S-TRANSFORMER

MULTILINGUAL

Bérard's Model (CNN + LSTM)

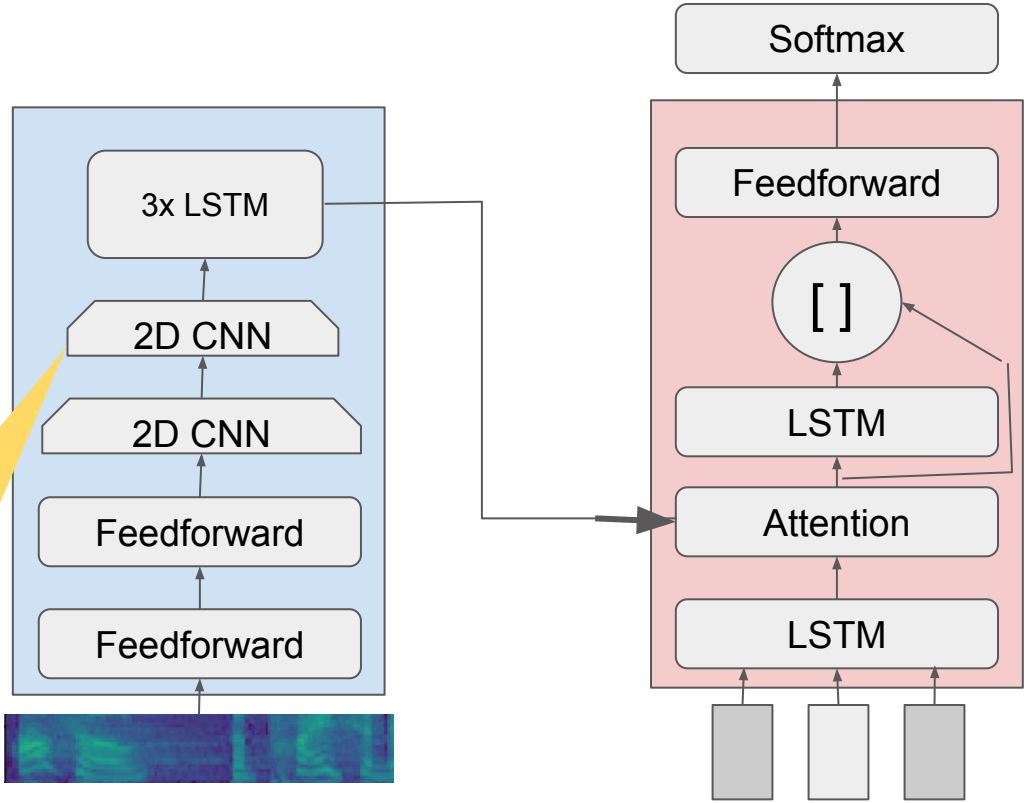
(Bérard et al., 2018)



Bérard's Model (CNN + LSTM)

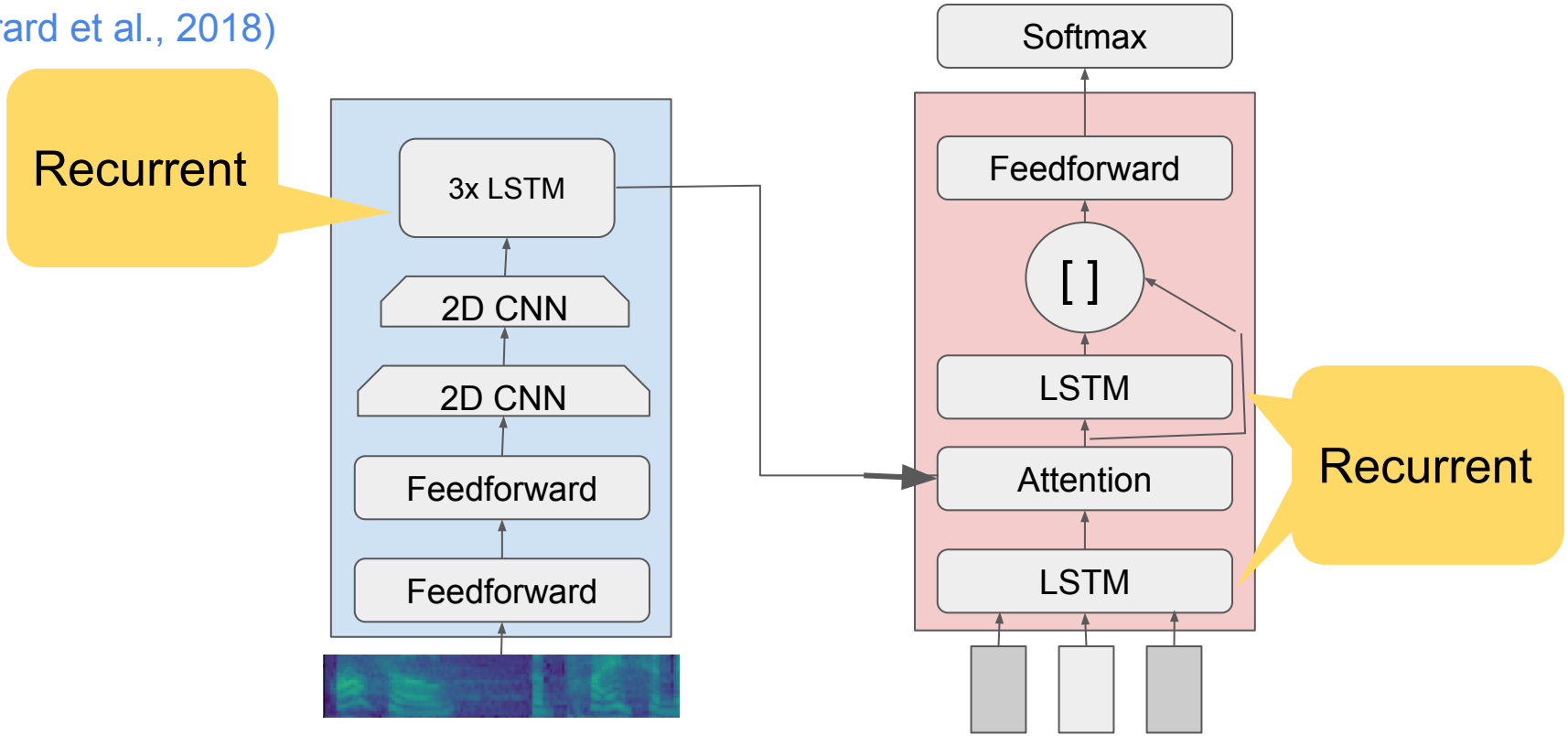
(Bérard et al., 2018)

Reduces seq length for computational and convergence reasons



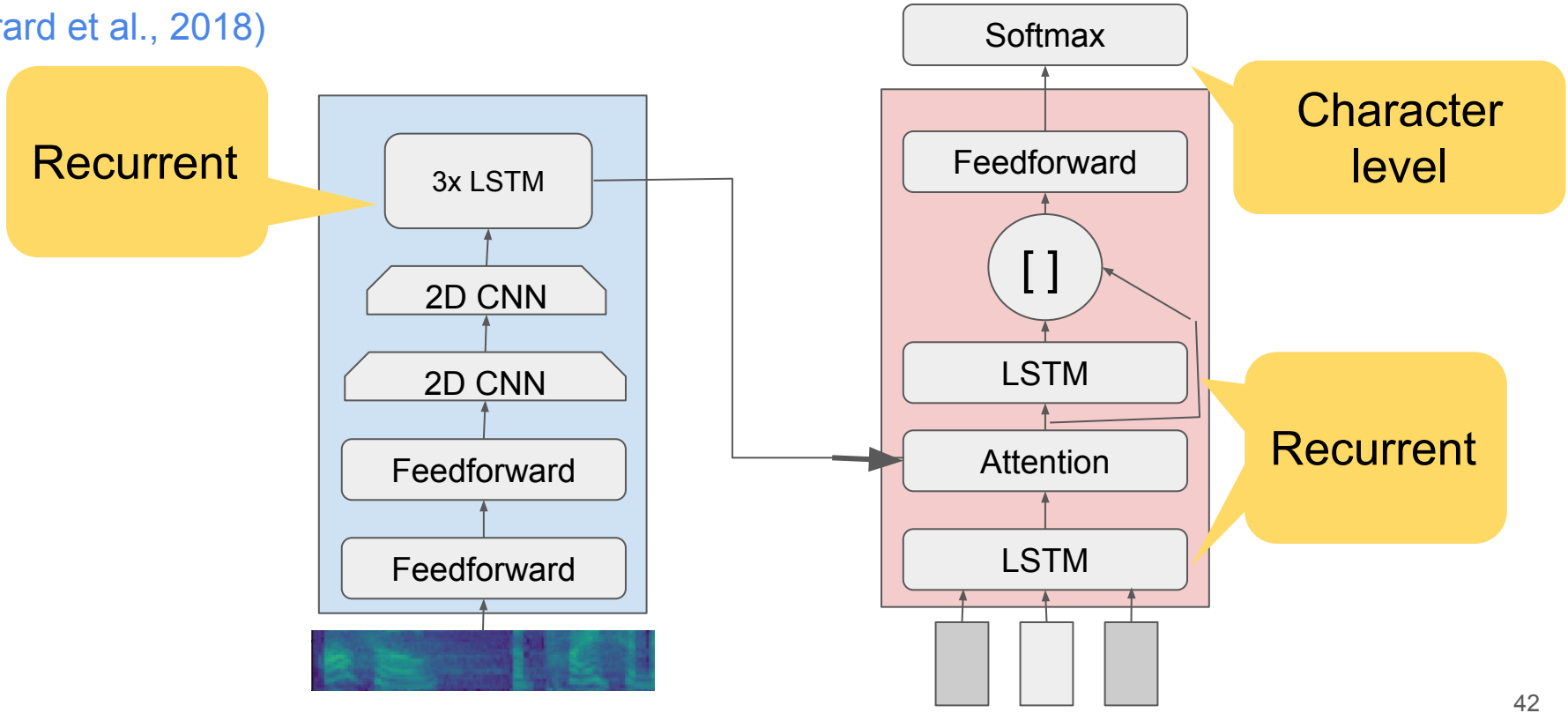
Bérard's Model (CNN + LSTM)

(Bérard et al., 2018)

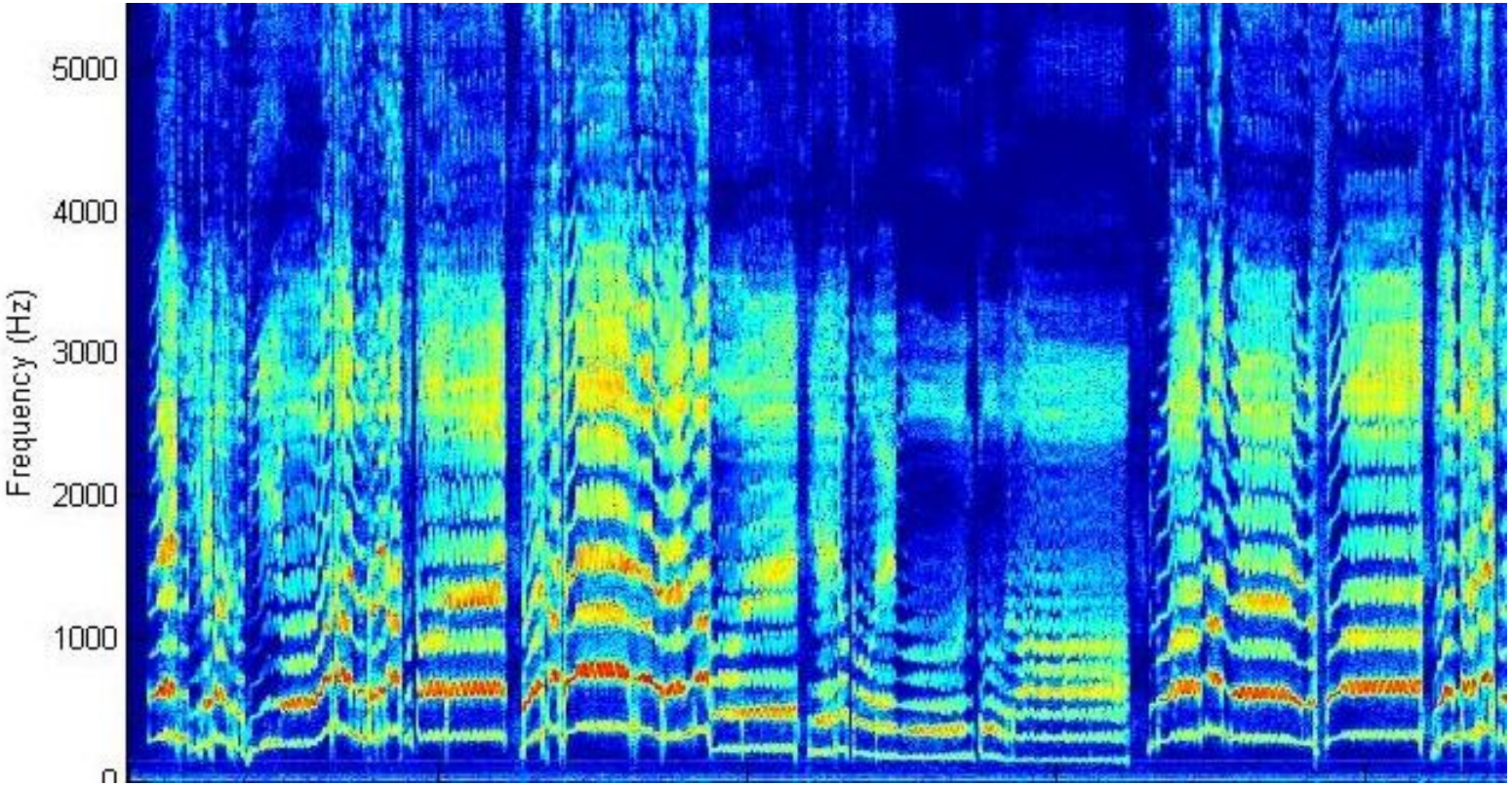


Bérard's Model (CNN + LSTM)

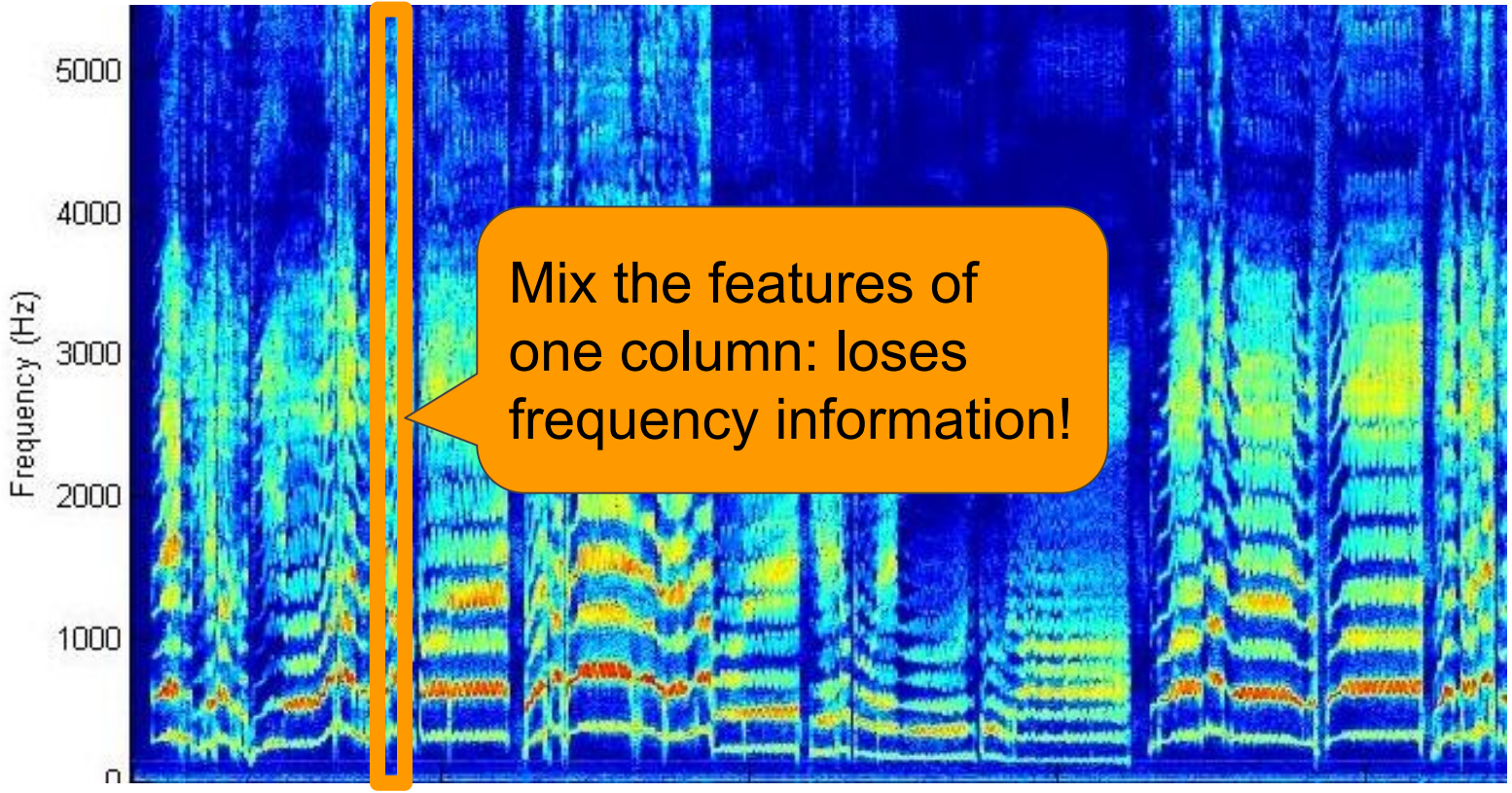
(Bérard et al., 2018)



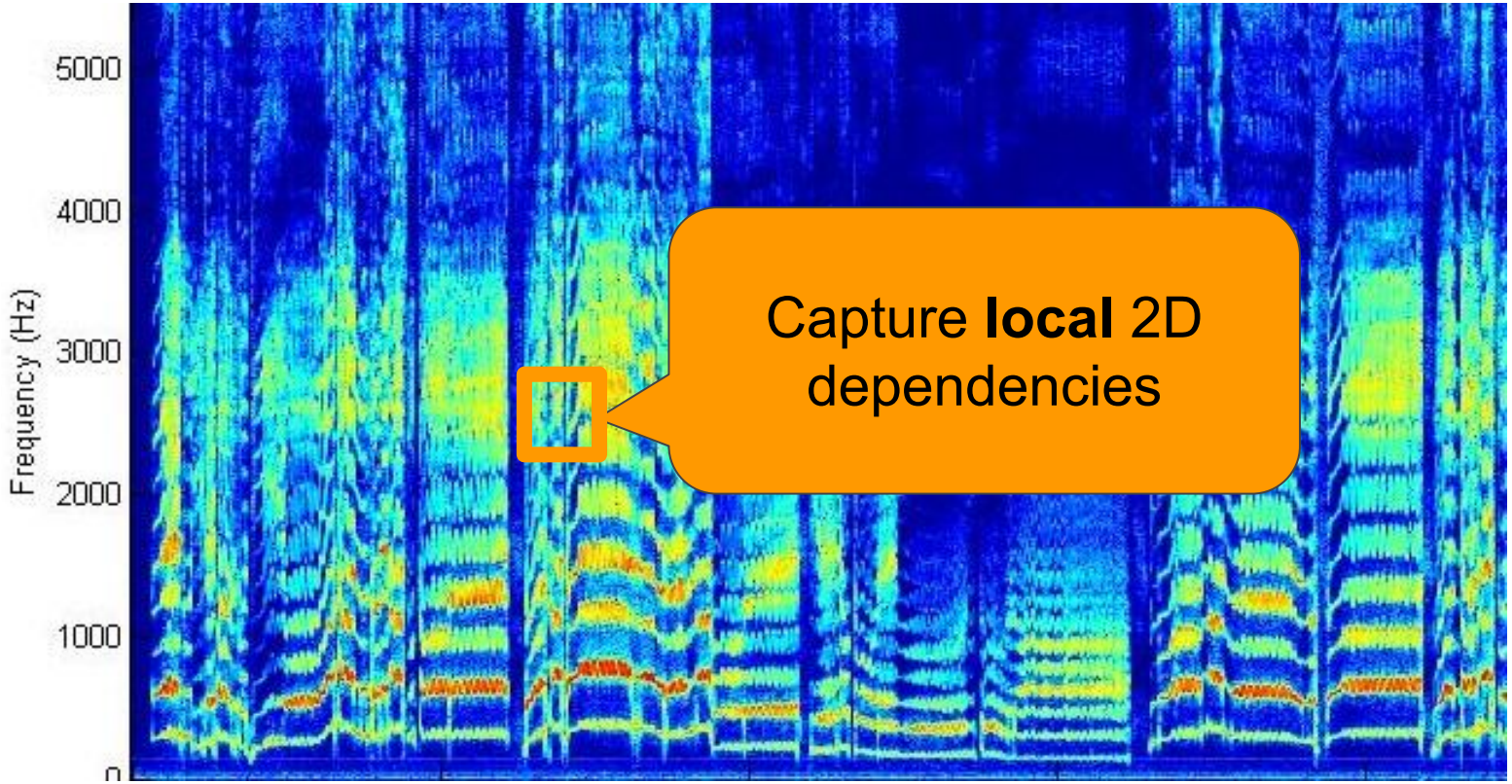
Spectrograms



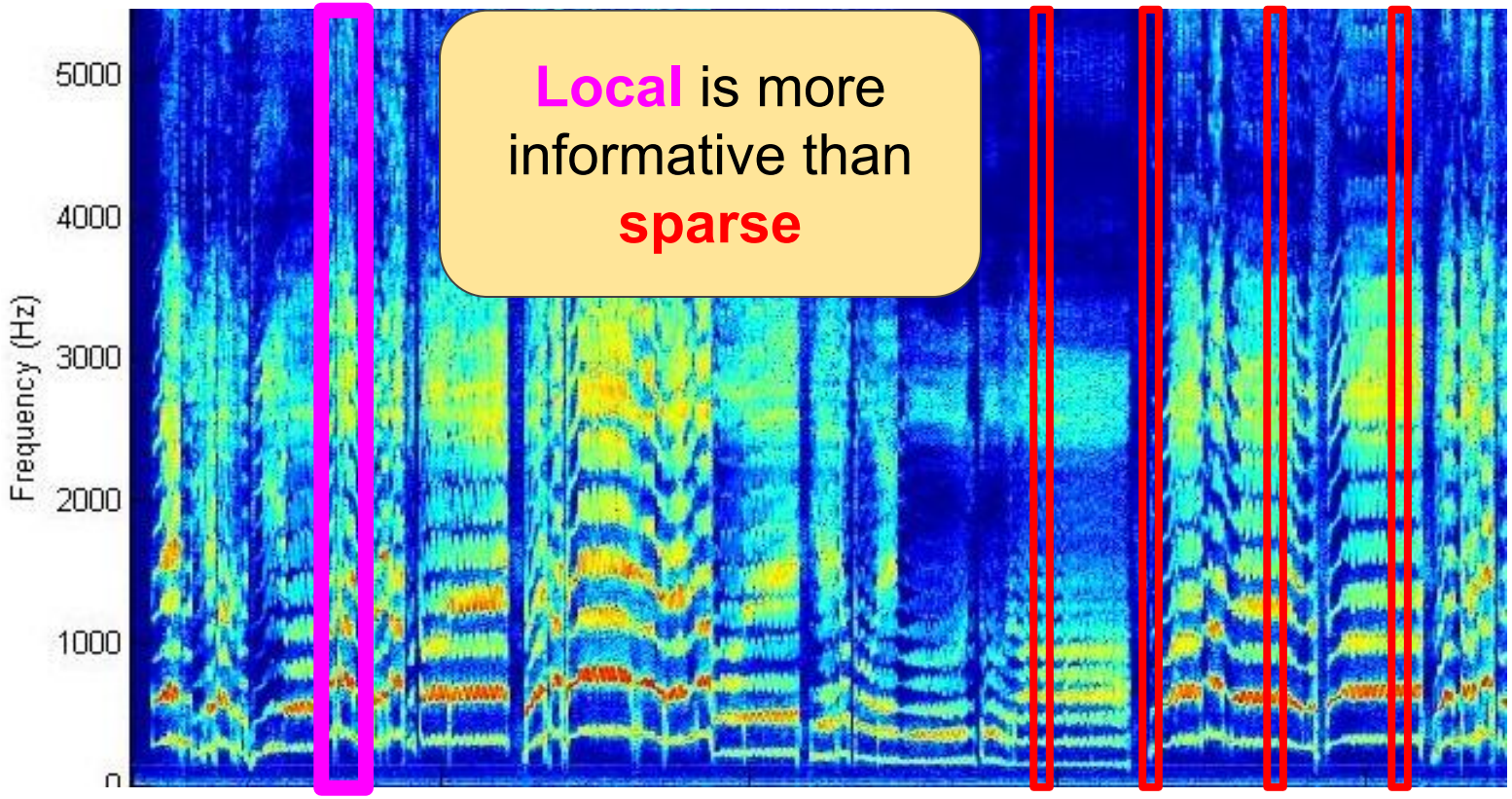
Spectrograms - Time-distributed Dense Layers



Spectrograms - 2D Convolutions



Spectrograms - Locality



S-Transformer

RQ: Can we build a more effective and efficient model for direct ST?

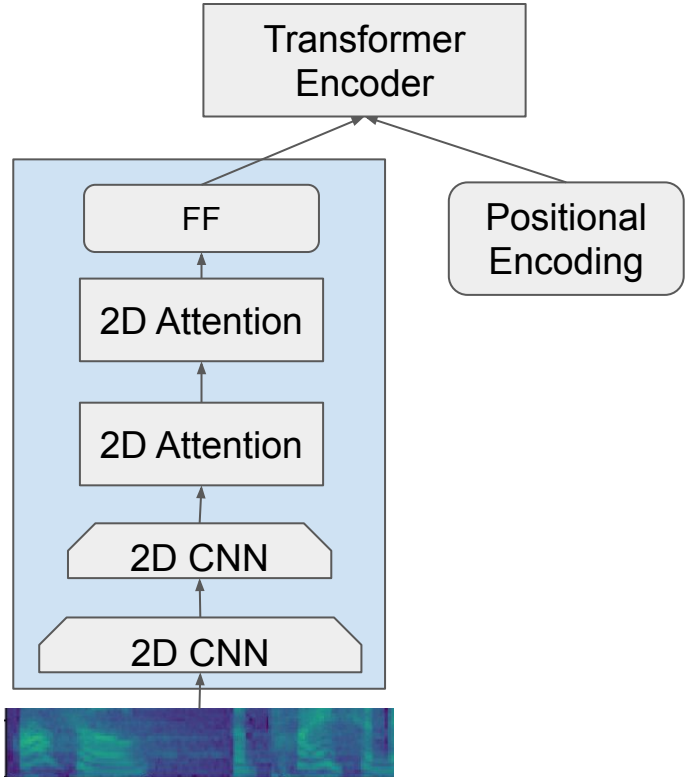
Adaptation of Transformer

- 2D-oriented encoder computation
- Locally-biased self-attention

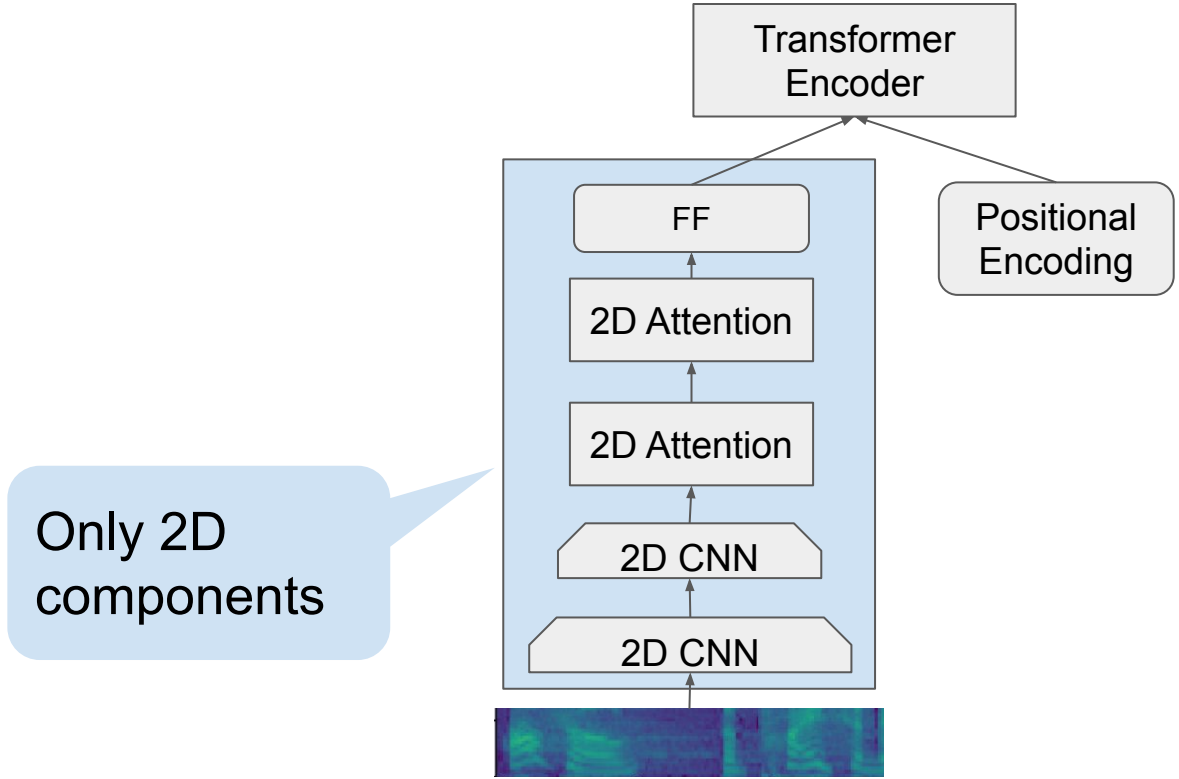
Di Gangi, M. A., et al. (2019). "Adapting transformer to end-to-end spoken language translation." *Proceedings of INTERSPEECH 2019*.

Di Gangi, M. A., et al. (2019). “Enhancing transformer for end-to-end speech-to-text translation.” *Proceedings of Machine Translation Summit XVII.*

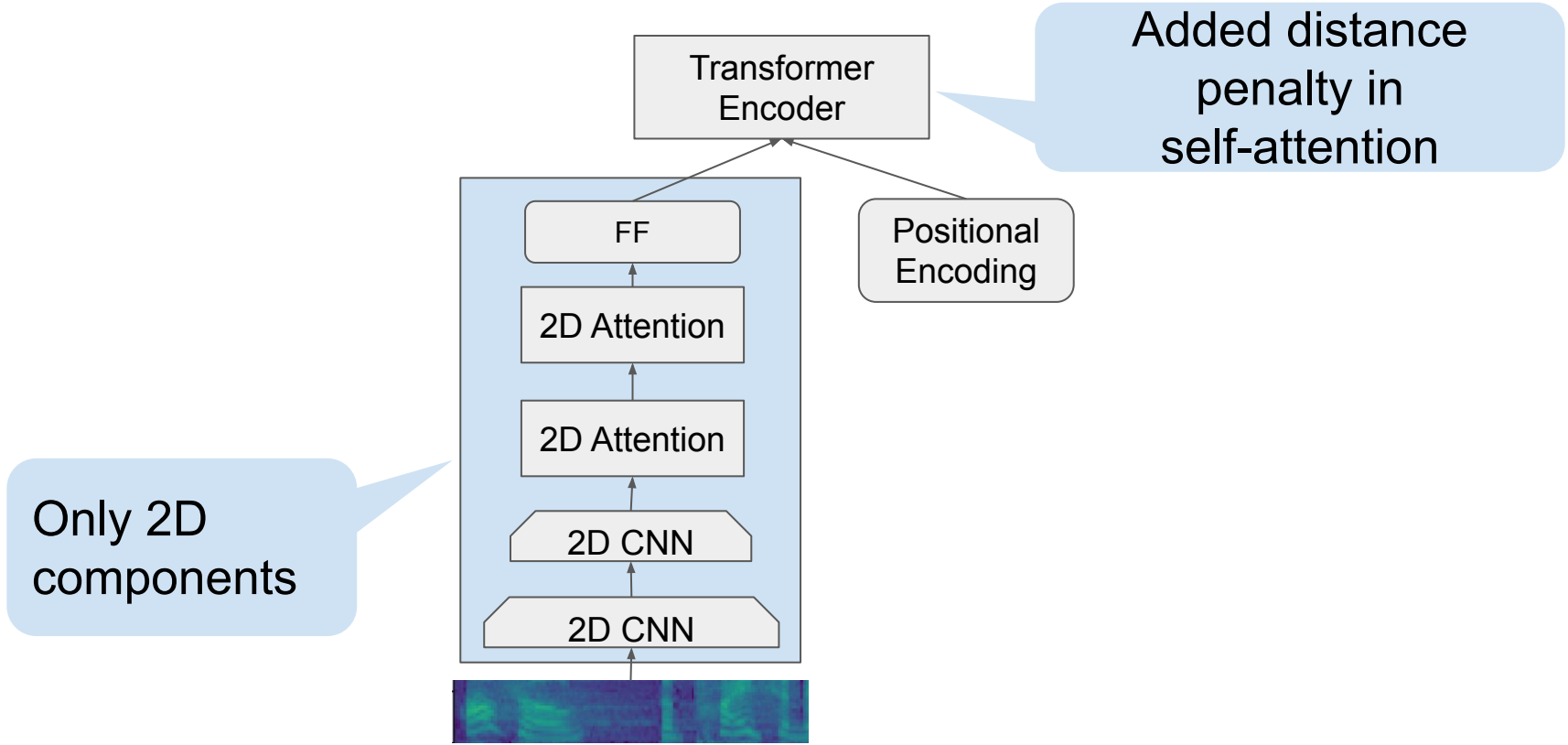
S-Transformer



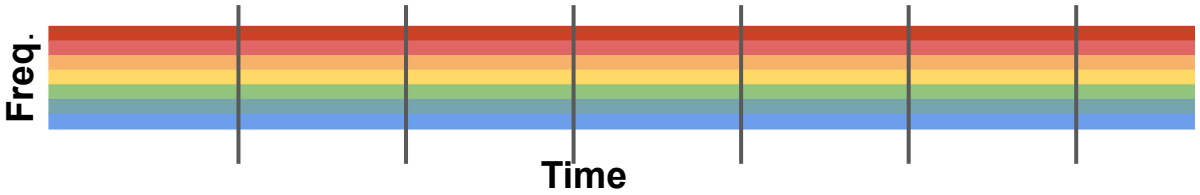
S-Transformer



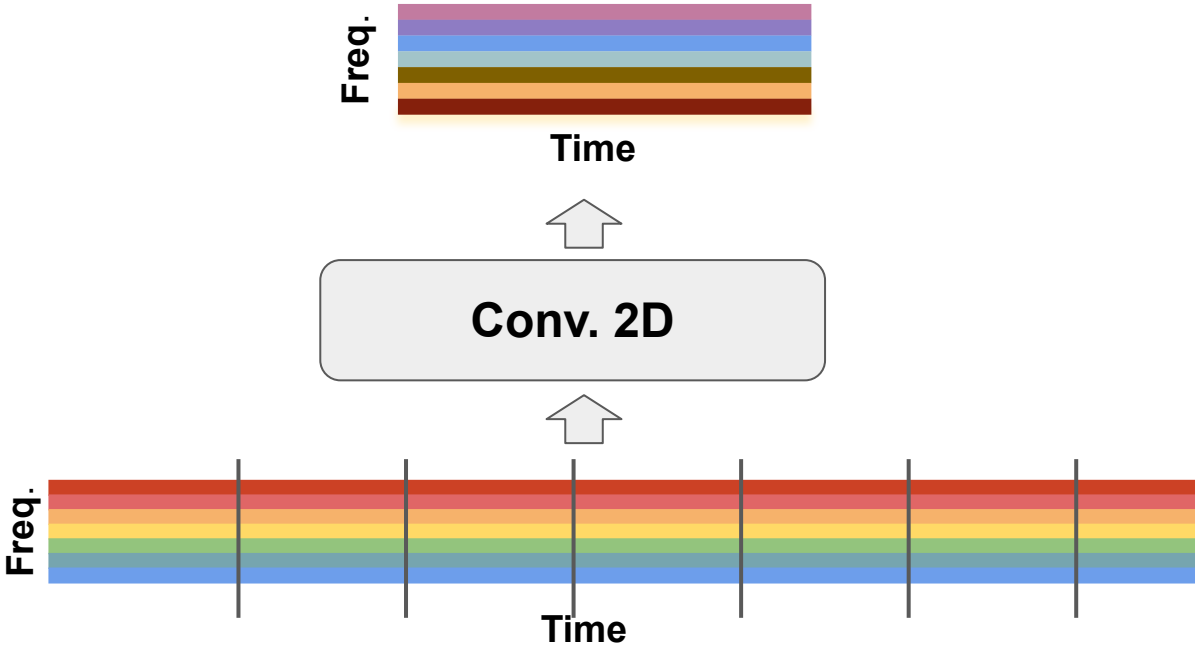
S-Transformer



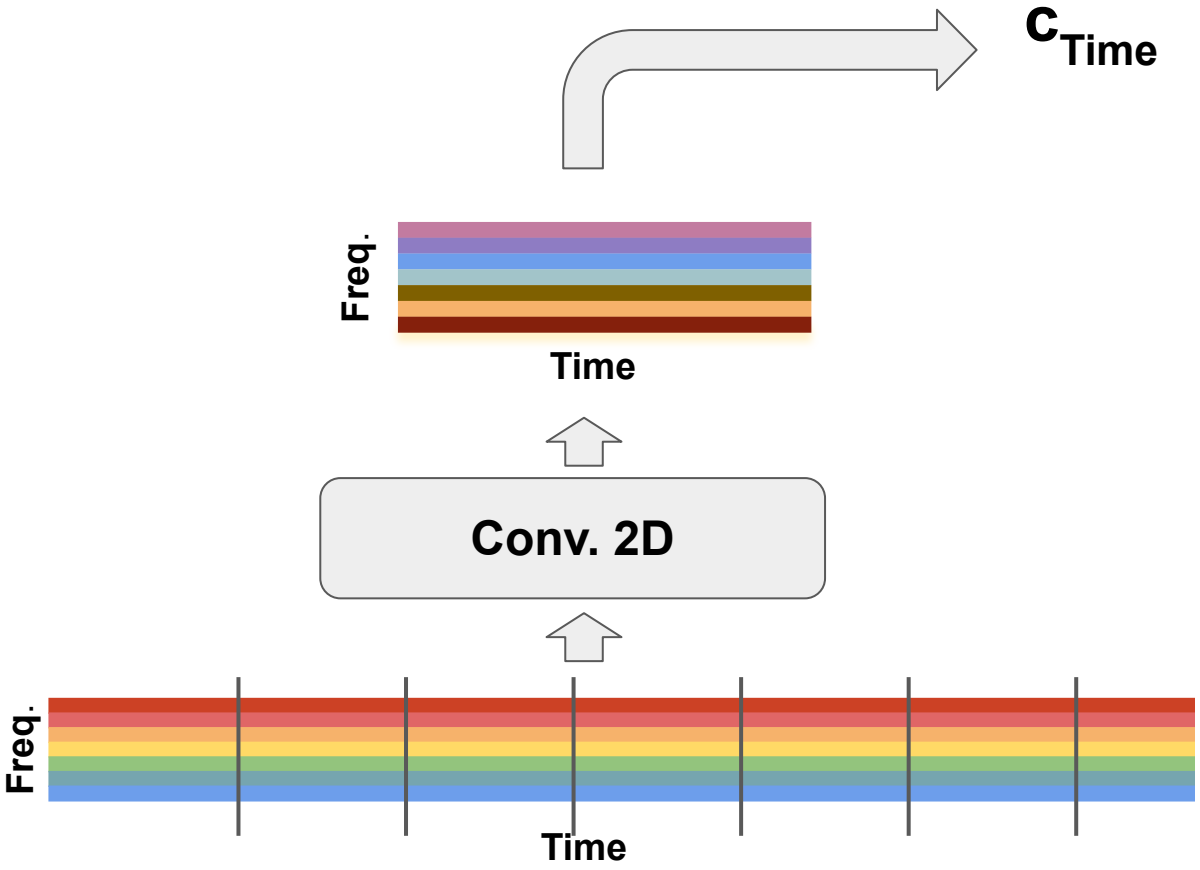
Preserving spectrogram properties: 2D Attention



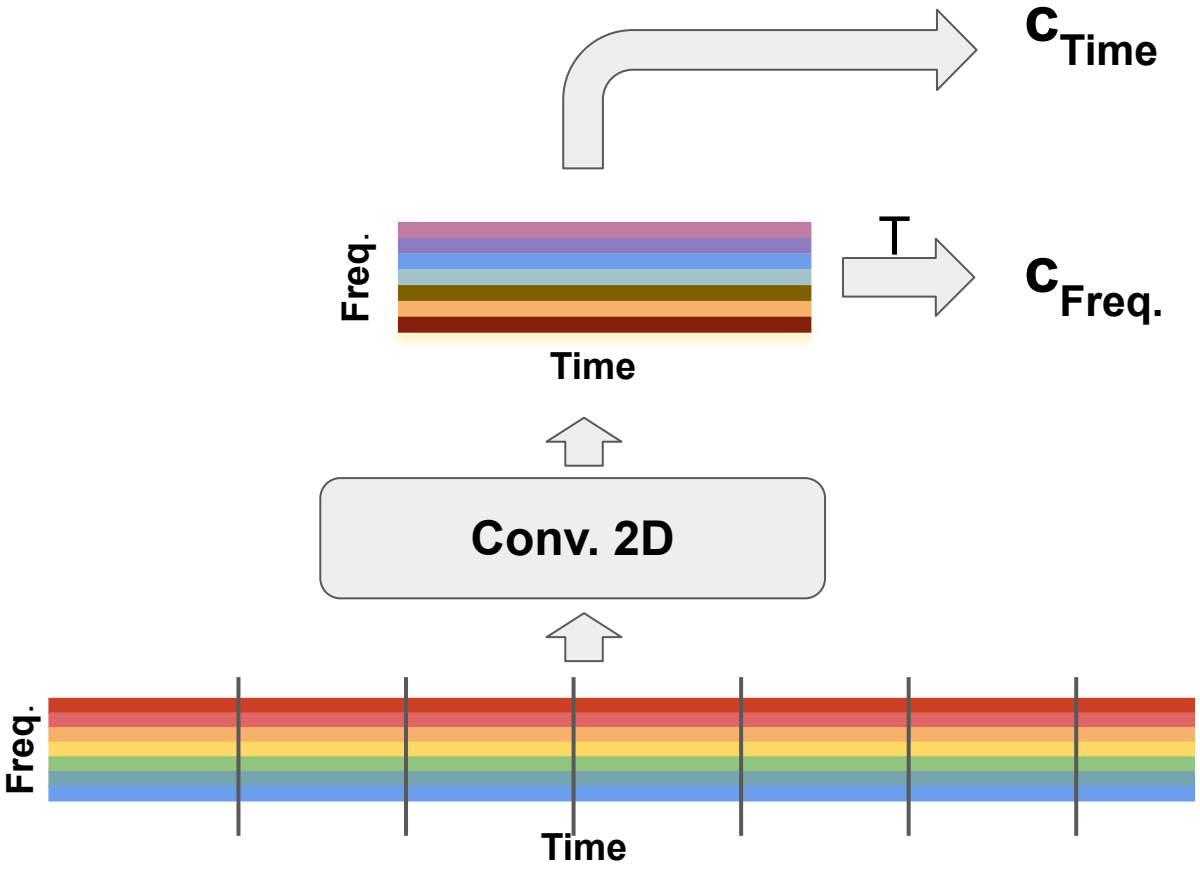
Preserving spectrogram properties: 2D Attention



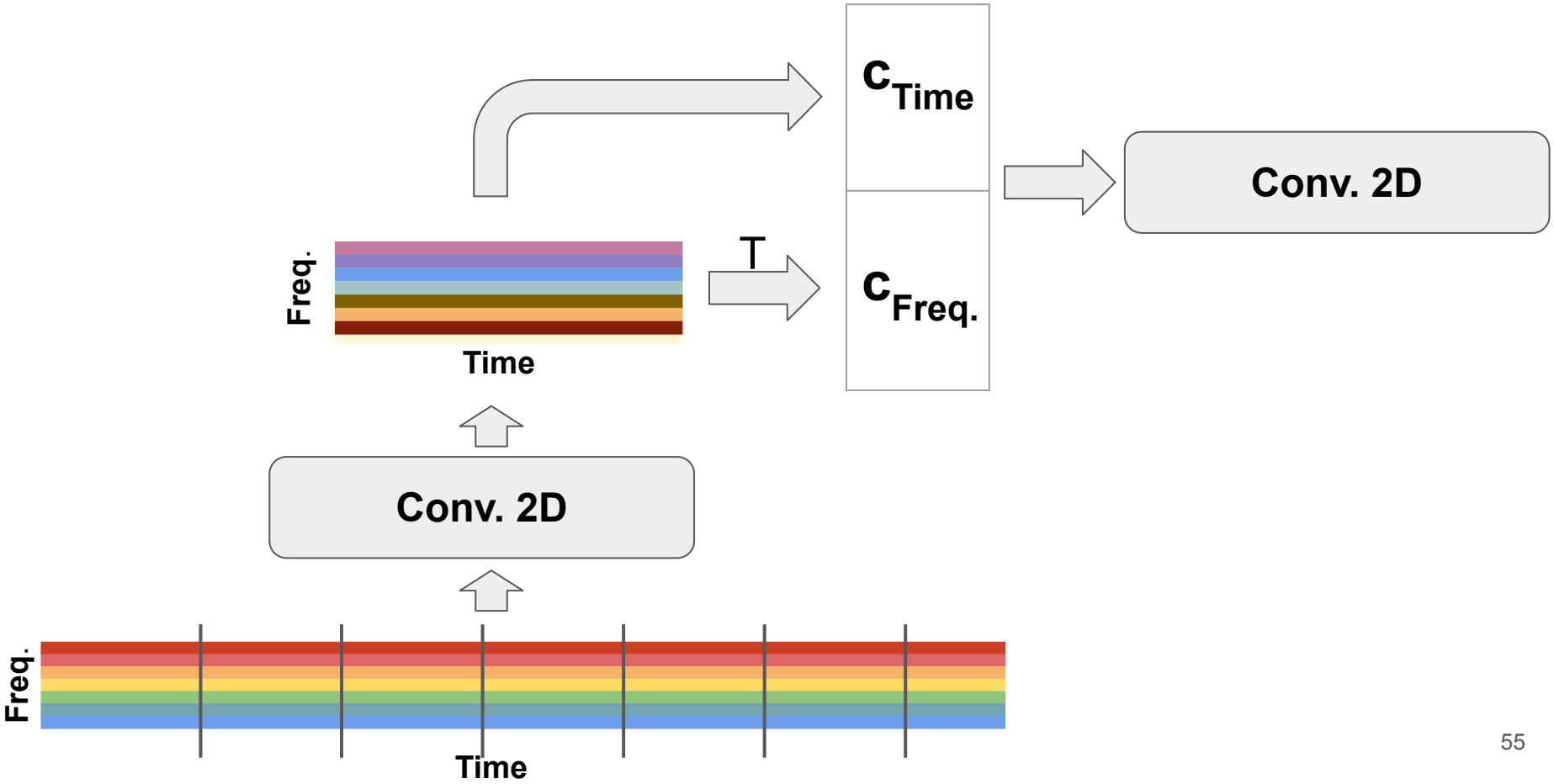
Preserving spectrogram properties: 2D Attention



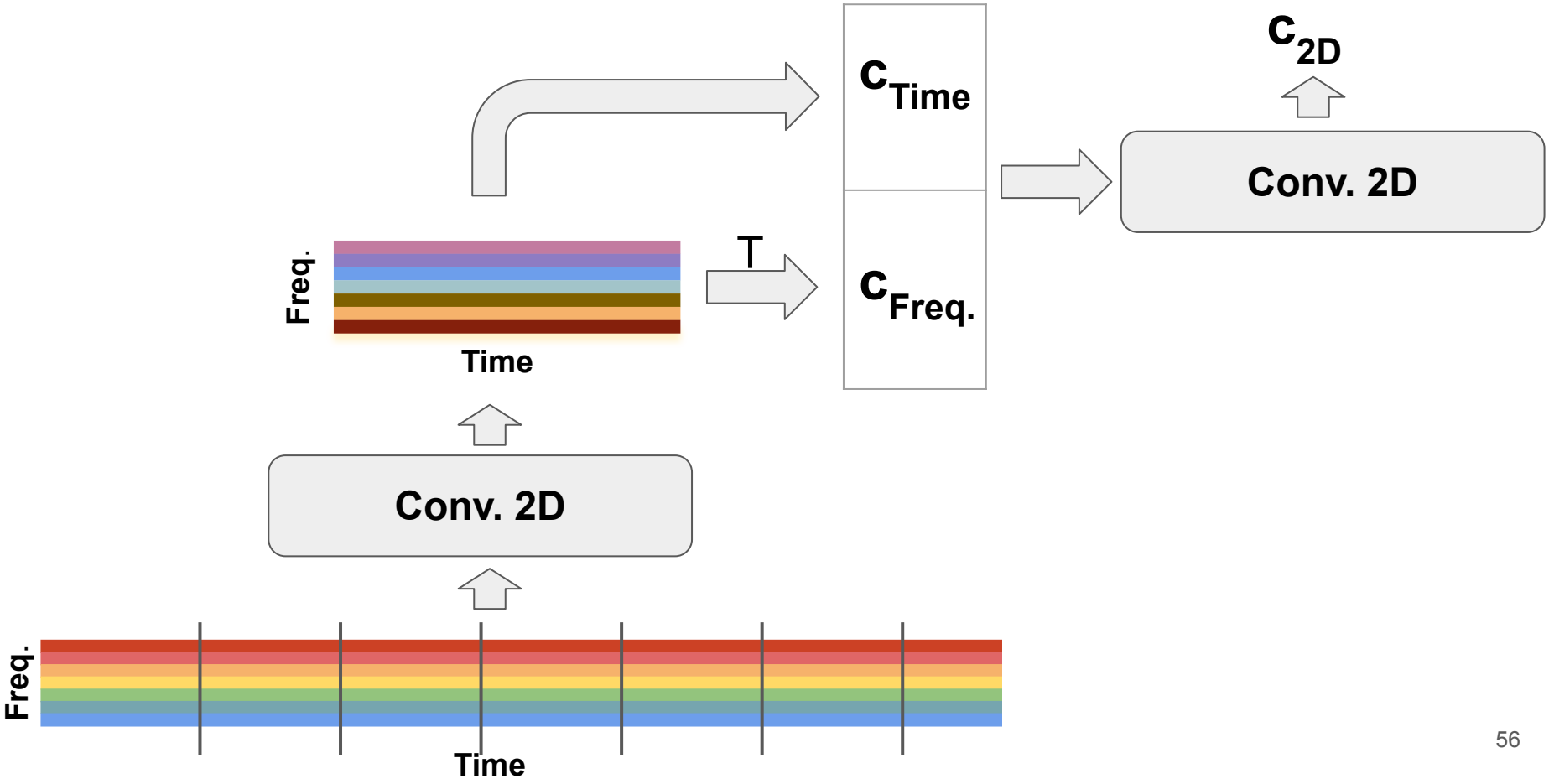
Preserving spectrogram properties: 2D Attention



Preserving spectrogram properties: 2D Attention



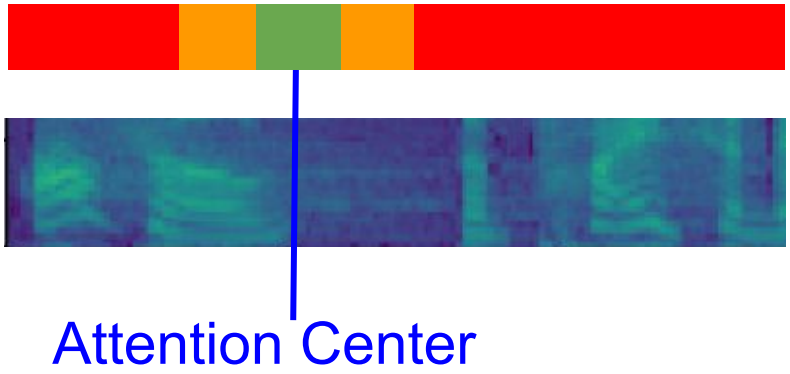
Preserving spectrogram properties: 2D Attention



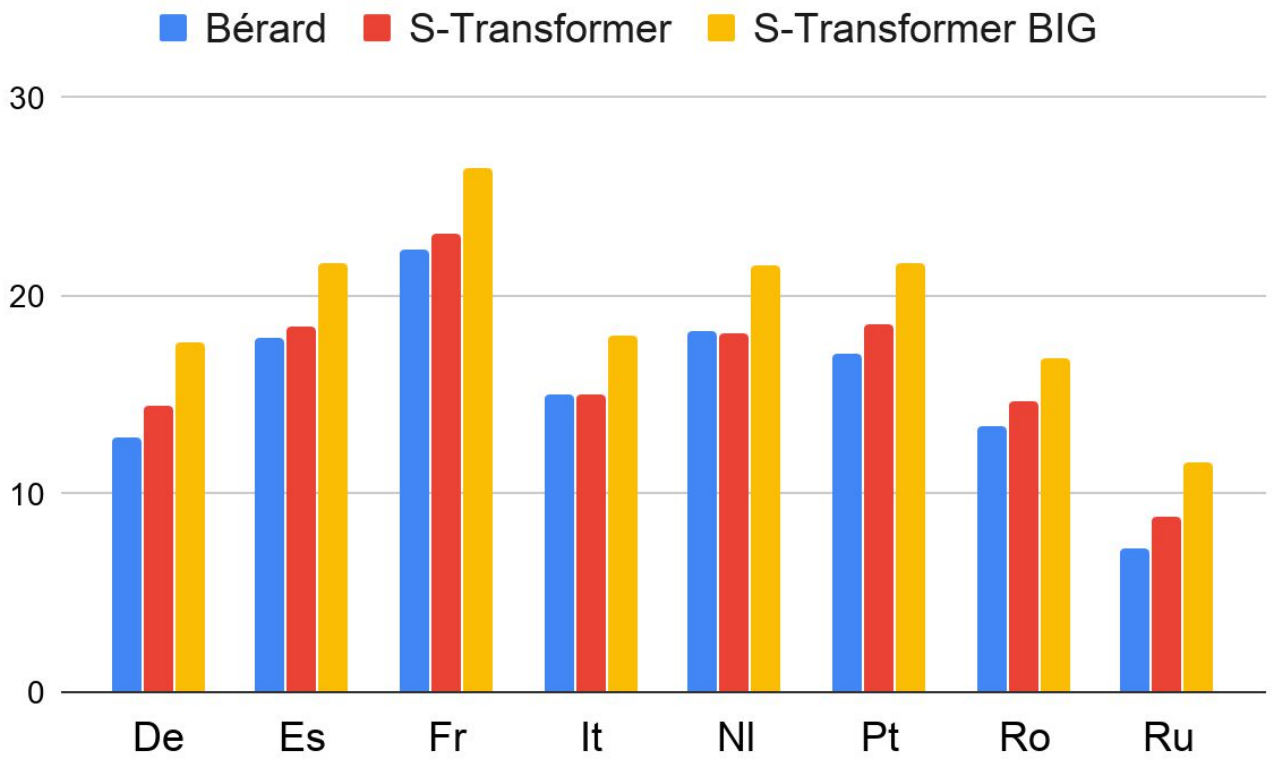
Distance Penalty

Very long input signal and with low “information density”

We bias the self-attention towards the local context with the goal of identifying low-level linguistic constituents.



Results on MuST-C



DIRECT SPEECH TRANSLATION

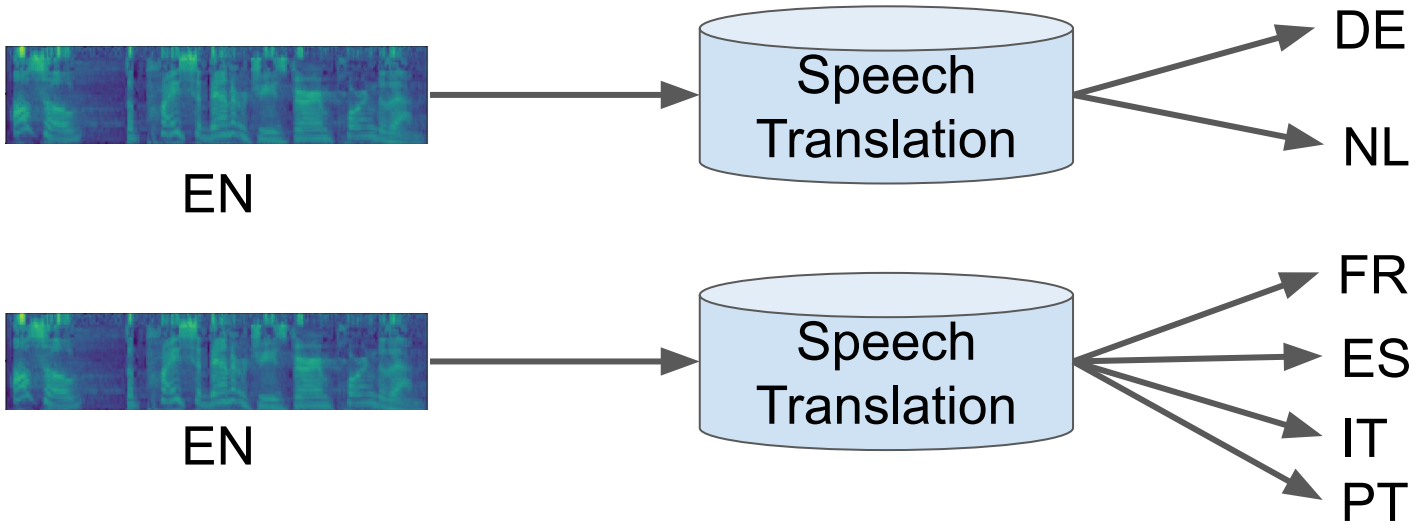
MUST-C

S-TRANSFORMER

MULTILINGUAL

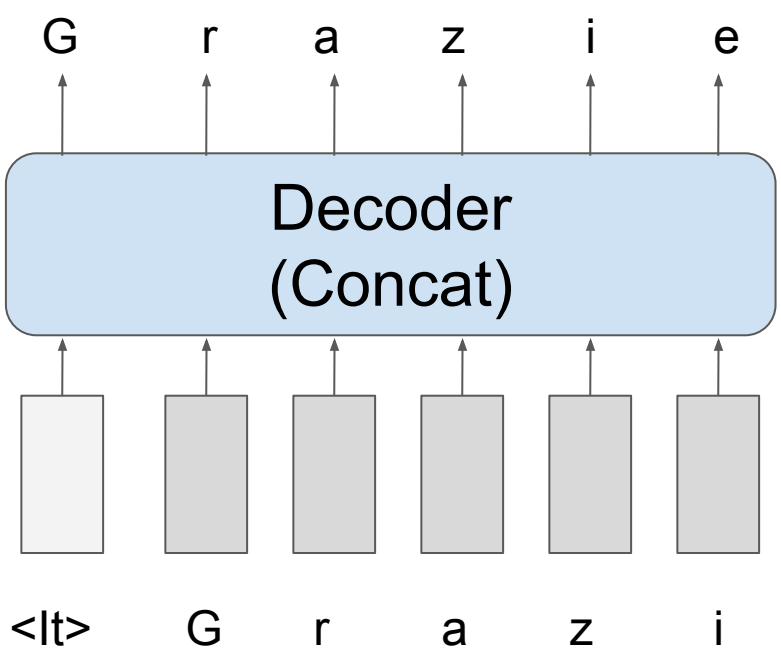
Multilingual ST

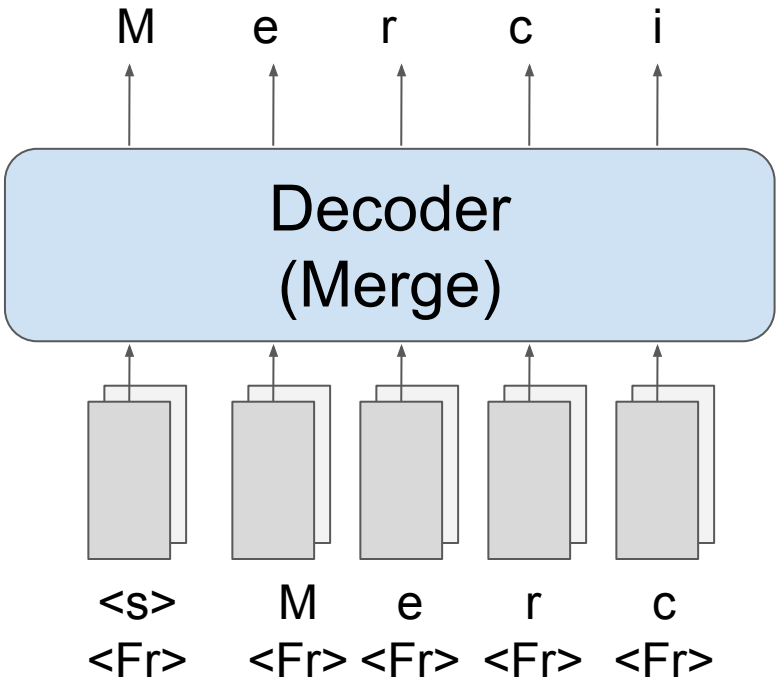
RQ: Can ST models benefit from training on multiple languages?



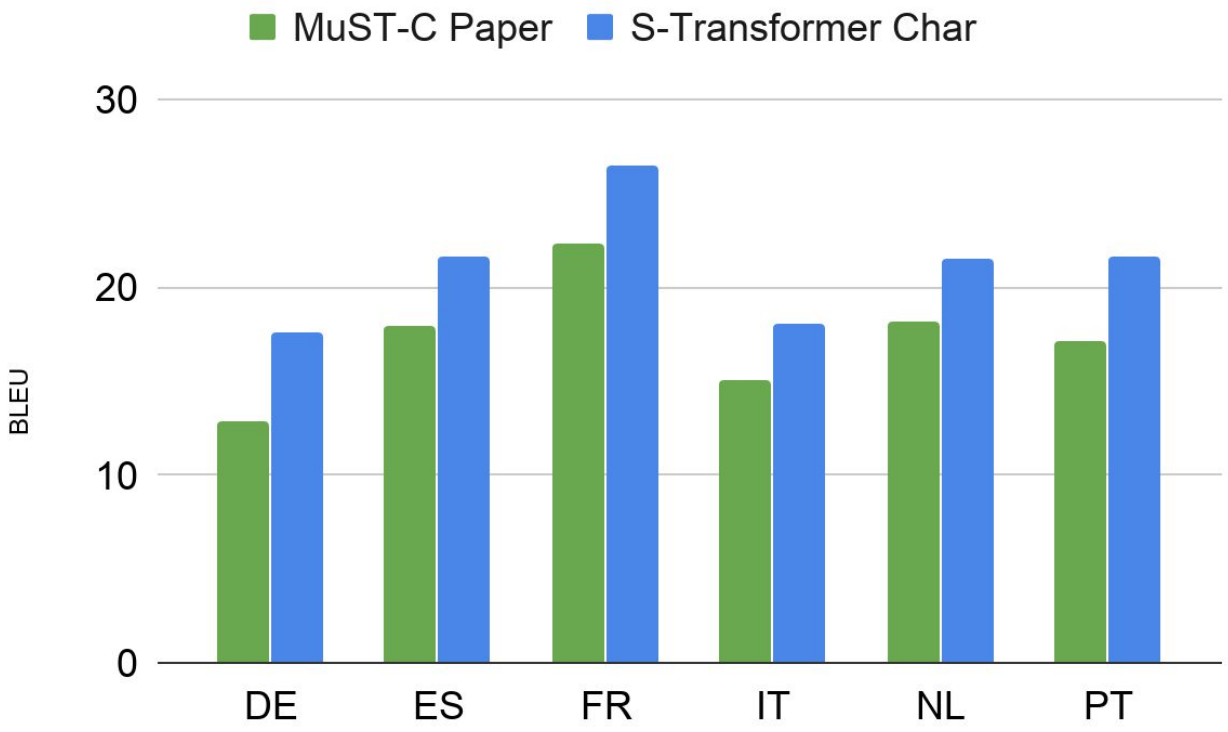
Di Gangi, Mattia Antonino, Matteo Negri, and Marco Turchi. (2019). "One-to-many multilingual end-to-end speech translation." *Proceedings of ASRU 2019*.

Multilingual ST

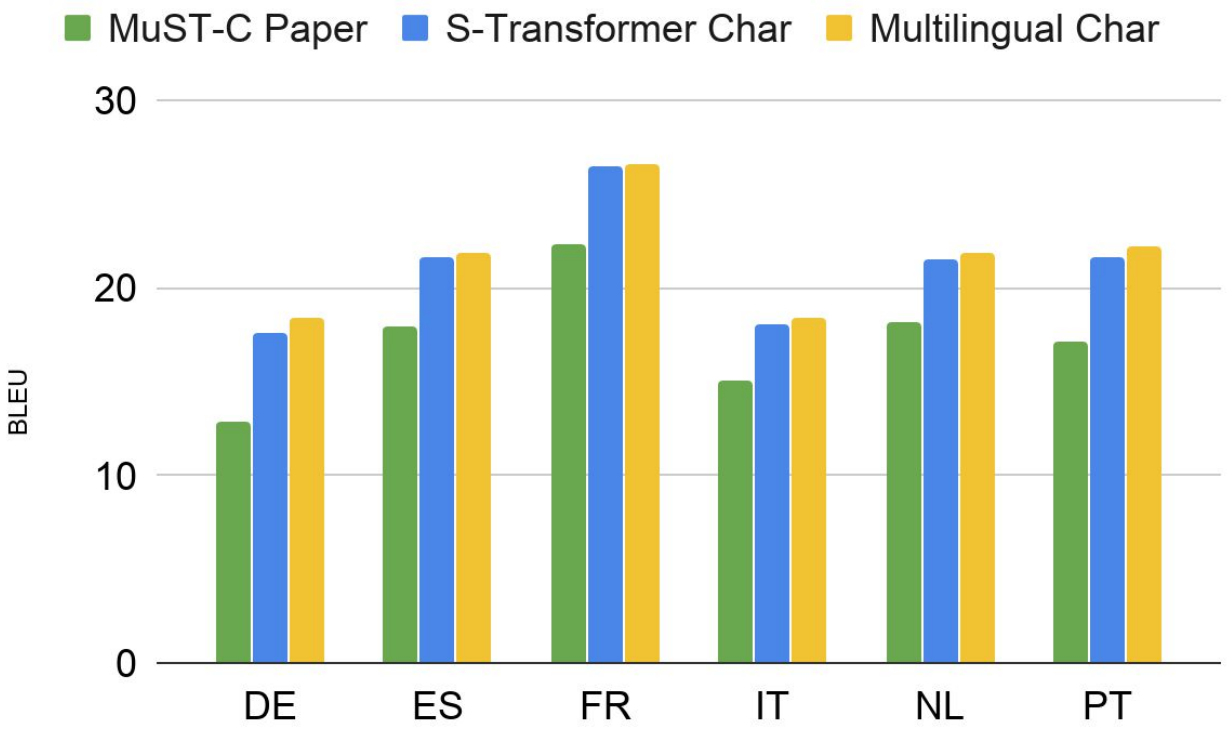




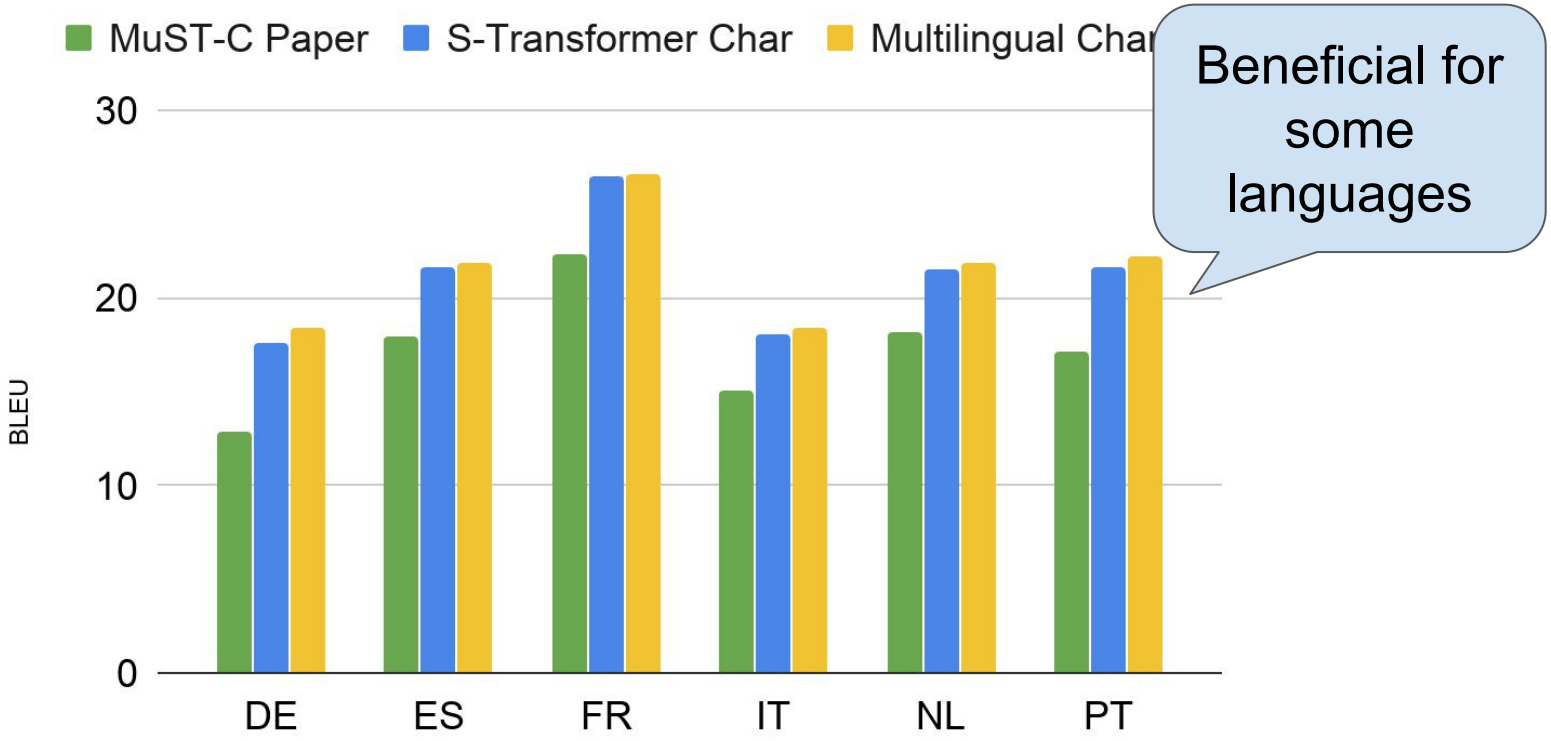
Results on MuST-C



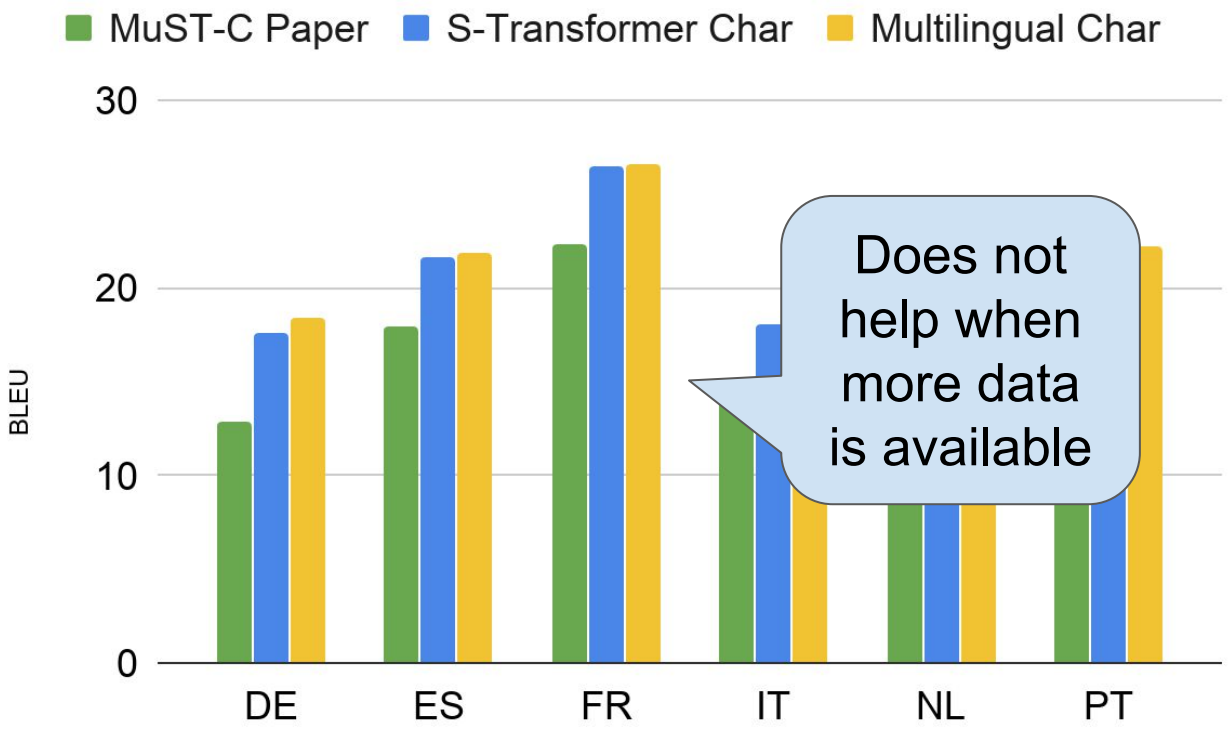
Results on MuST-C



Results on MuST-C



Results on MuST-C



Example (EN-IT)

EN: In 2010, it was viewed 23 million times. This is a chart of what it looked like when it first became popular last summer.
REF: Nel 2010, è stato visto 23 milioni di volte. Questo è un grafico delle visualizzazioni quando è diventato popolare l'estate scorsa.

PREV: Nel 2010, era voi 23 milioni di volte. Perché il grafico sembrava che quando all'inizio diventava popolare, l'estate scorsa.

BEST: Nel 2010, è stato visualizzato 23 milioni di volte. Questo è un grafico di quello che è diventato popolare l'estate scorsa.

Example (EN-IT)

EN: In 2010, it was viewed 23 million times. This is a chart of what it looked like when it first became popular last summer.
REF: Nel 2010, è stato visto 23 milioni di volte. Questo è un grafico delle visualizzazioni quando è diventato popolare l'estate scorsa.

PREV: **Nel 2010**, era voi **23 milioni di volte**. Perché il **grafico** sembrava che quando all'inizio diventava **popolare**, **l'estate scorsa.**

BEST: **Nel 2010, è stato visualizzato 23 milioni di volte. Questo è un grafico** di quello che **è diventato popolare l'estate scorsa.**

IWSLT Participation

2020

System Description and Findings

Goal: Approach Cascade's Results

RQ: How to build a state-of-the-art DST model?

S-Transformer with:

- Many training sets
- Deep encoder
- Multitask learning with CTC loss
- Data augmentation
- Knowledge distillation

Gaido, Marco, et al. "End-to-End Speech-Translation with Knowledge Distillation: FBK@ IWSLT2020." *Proceedings of IWSLT 2020*.

Goal: Approach Cascade's Results

S-Transformer with:

- **Many training sets**
- Deep encoder
- Multitask learning with CTC loss
- Data augmentation
- Knowledge distillation

With German translations:

- MuST-C En-De
- iwslt18
- Europarl-ST En-De

Goal: Approach Cascade's Results

S-Transformer with:

- **Many training sets**
- Deep encoder
- Multitask learning with CTC loss
- Data augmentation
- Knowledge distillation

Only English:

- TED LIUM 3
- Librispeech
- Mozilla Common Voice
- How2
- MuST-C En-Fr

Goal: Approach Cascade's Results

S-Transformer with:

- Many training sets
- **Deep encoder**
- Multitask learning with CTC
- Data augmentation
- Knowledge distillation

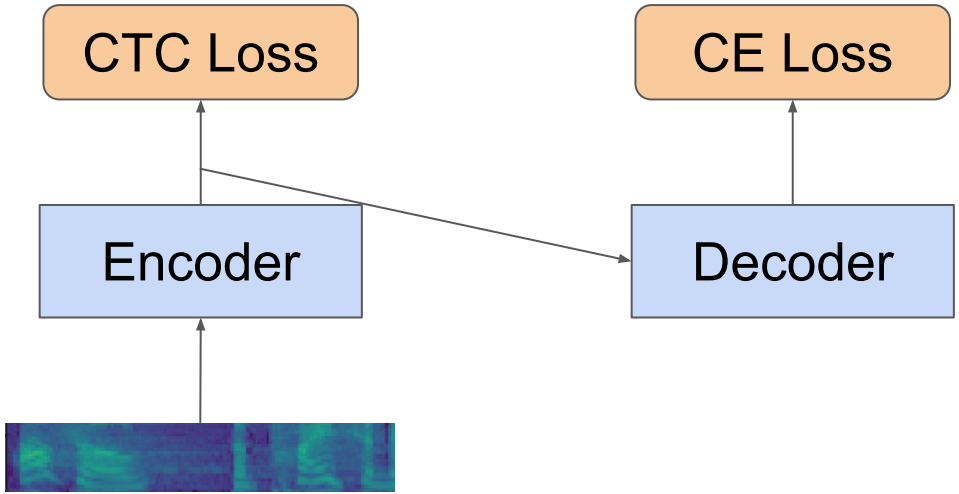
- 12 layers encoder
- 4 layers decoder
- NO 2D attention

Goal: Approach Cascade's Results

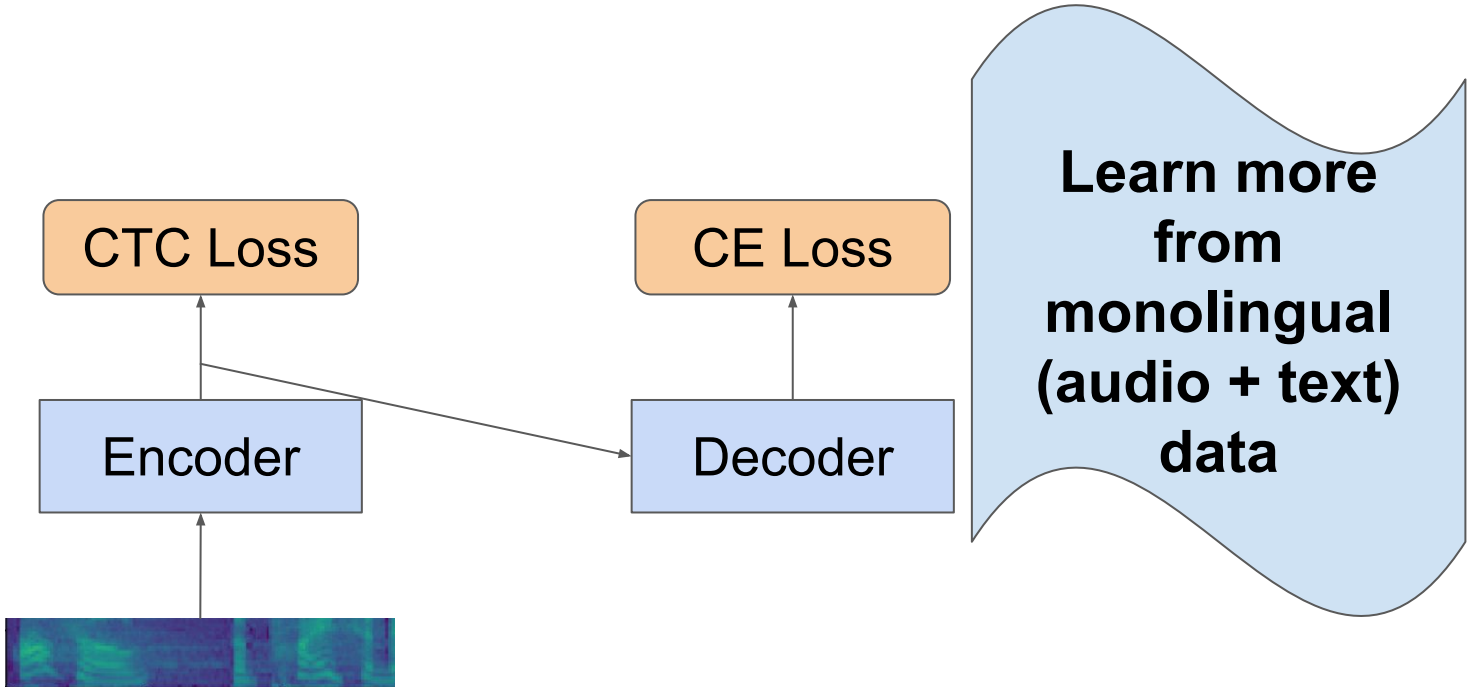
S-Transformer with:

- Many training sets
- Deep encoder
- **Multitask learning with CTC loss**
- Data augmentation
- Knowledge distillation

Multitask Learning with CTC Loss



Multitask Learning with CTC Loss



Goal: Approach Cascade's Results

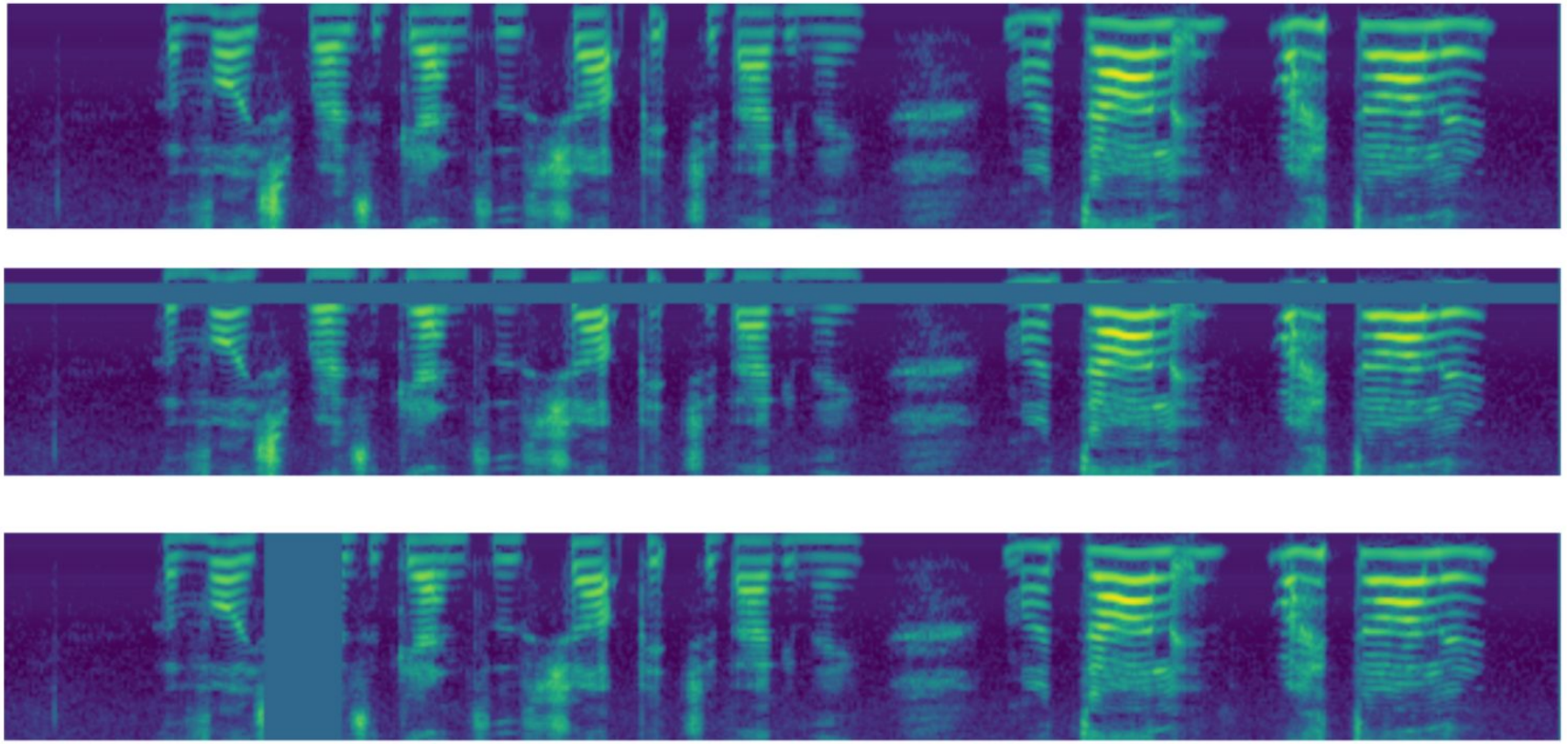
S-Transformer with:

- Many training sets
- Deep encoder
- Multitask learning with CTC loss
- **Data augmentation**
- Knowledge distillation

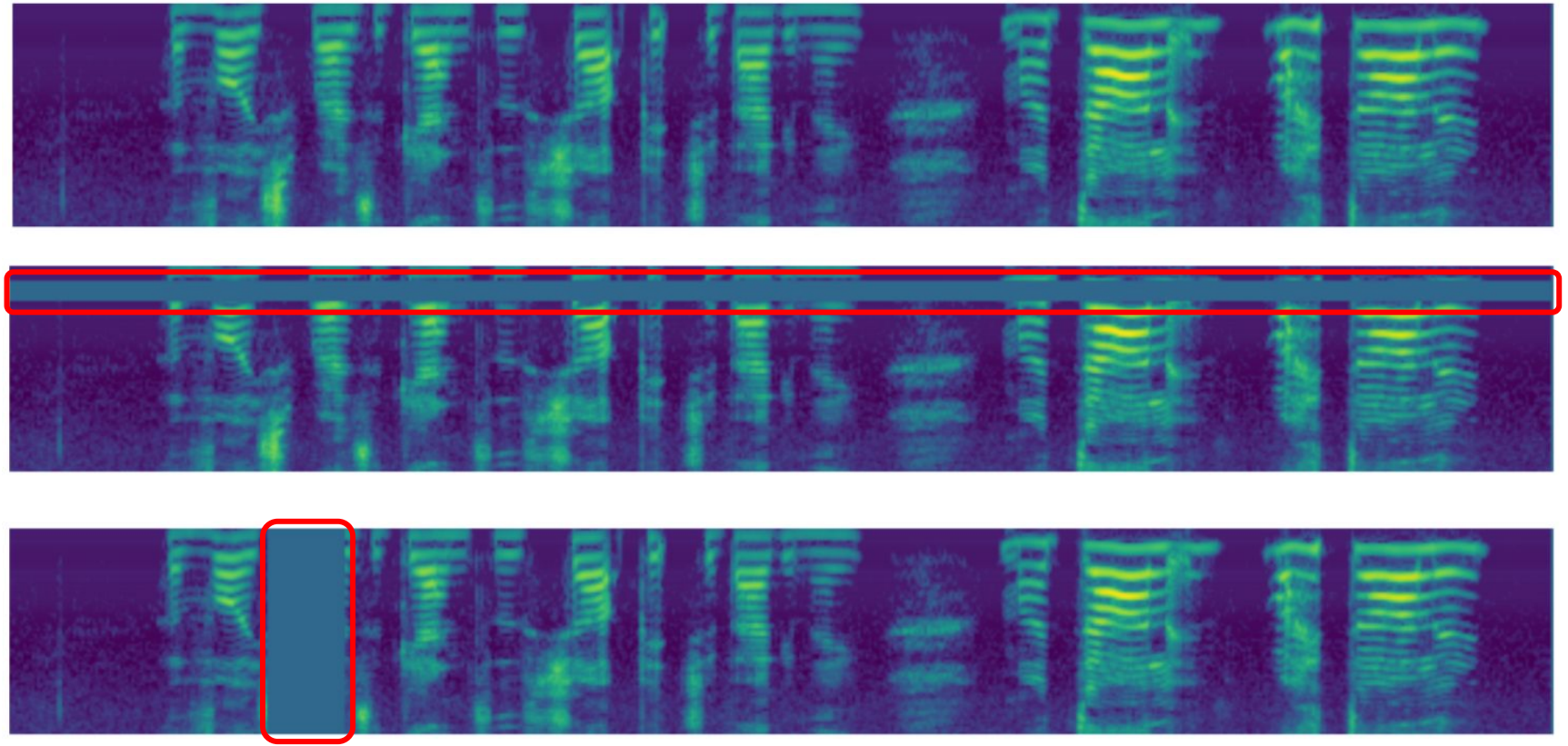
Data Augmentation

- **SpecAugment**: Zeroing out time or frequency bands (online)
- **Time Stretch:** Expanding or reducing time bands (online)
- **Synthetic Translations:** Use an MT model to translate transcripts of corpora with no translation (offline; as last year)

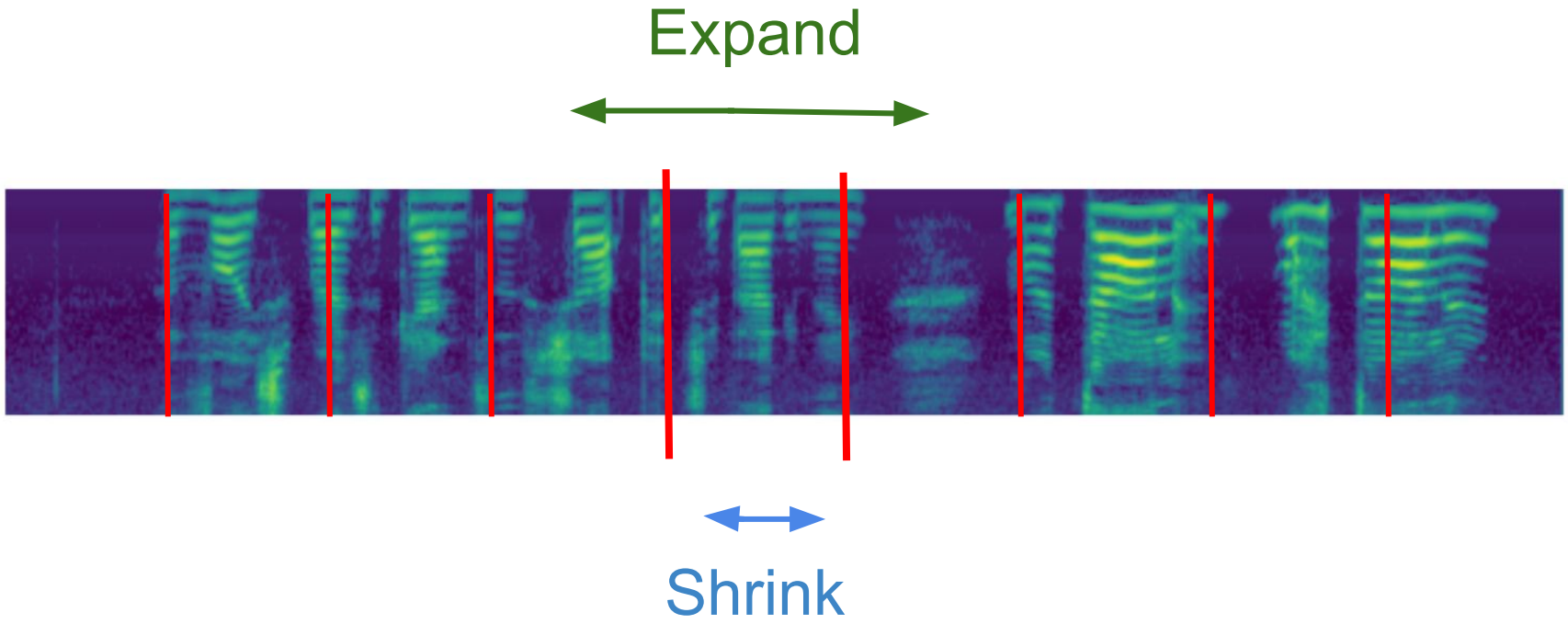
SpecAugment



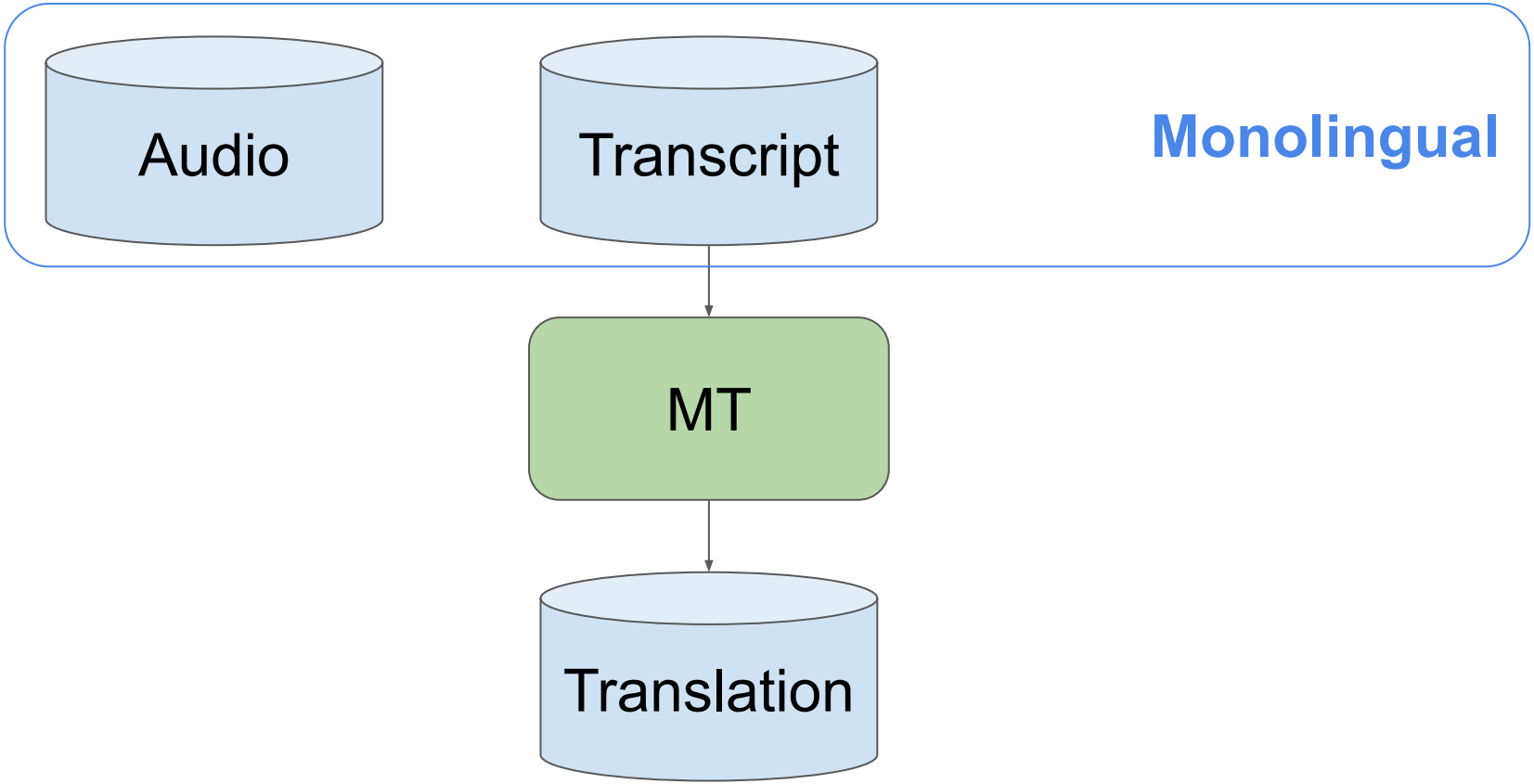
SpecAugment



Time Stretch



Synthetic Translations

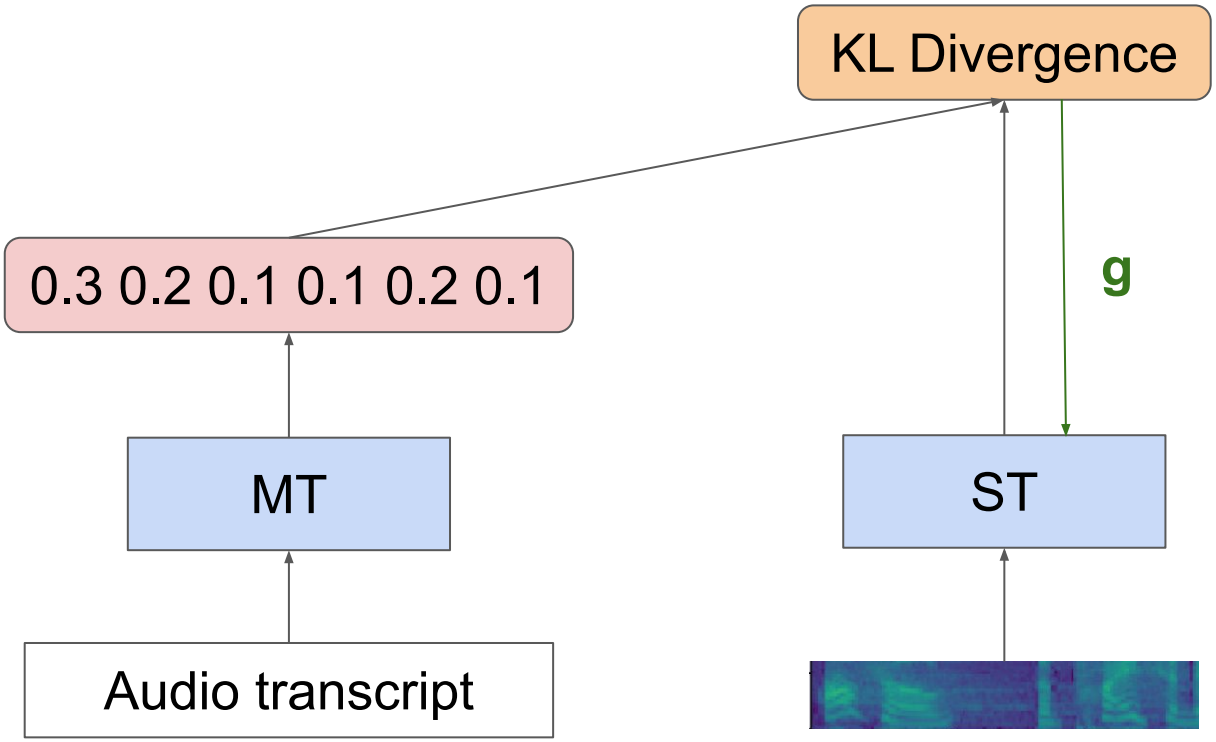


Goal: Approach Cascade's Results

S-Transformer with:

- Many training sets
- Deep encoder
- Multitask learning with CTC loss
- Data augmentation
- **Knowledge distillation**

Knowledge Distillation



Findings

- CTC loss improves results significantly
- Time stretch prevents overfitting
- SpecAugment always provides visible improvements (already known)
- Synthetic translations are fundamental for high scores
- Knowledge distillation + Fine-tuning w/o KD gives enormous boosts to the score
- Individual contributions are positively combined!

IWSLT 2020 Official Results

Speech Translation : TED English-German tst 2020 (given segmentation)

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
AppTek/RWTH	22.49	65.20	51.40	52.75	23.73	62.93
KIT	22.06	65.38	51.22	51.26	23.24	63.10
SRPOL	21.49	65.74	49.81	56.20	22.7	63.82
FBK	20.75	68.11	49.87	55.31	21.88	66.04
AppTek/RWTH	20.5	70.08	49.65	54.85	21.84	67.95
KIT	19.82	70.51	48.62	56.91	22	67.36
BHANSS	18.09	71.78	47.09	60.96	19	70.06
HY	17.02	76.37	47.03	58.32	18.07	74.23
DiDi Labs	10.14	101.56	41.95	62.60	10.83	99.60
HY	6.77	86.31	36.81	76.30	7.26	84.91

IWSLT 2020 Official Results

Speech Translation : TED English-German tst 2020 (given segmentation)

System	BLEU	TER	BEER	characTER	BLEU(CI)	TER(CI)
AppTek/RWTH	22.49	65.20	51.40	52.75	23.73	62.93
KIT	22.06	65.38	51.22	51.26	23.24	63.10
SRPOL	21.49	65.74	49.81	56.20	22.7	63.82
FBK	20.75	68.11	49.87	55.31	21.88	66.04
AppTek/RWTH	20.5	70.08	49.65	54.85	21.84	67.95
KIT	19.82	70.51	48.62	56.91	22	67.36
BHANSS	18.09	71.78	47.09	60.96	19	70.06
HY	17.02	76.37	47.03	58.32	18.07	74.23
DiDi Labs	10.14	101.56	41.95	62.60	10.83	99.60
HY	6.77	86.31	36.81	76.30	7.26	84.91

MuST-SHE

TEST SET

EVALUATION

EXPERIMENTS

MuST-SHE

Research questions:

1. How to build a test set to evaluate gender bias of **speech** translation systems? (English = Gender agnostic source)
2. How to perform the evaluation? Is BLEU score really sufficient?
3. Can the access to audio reduce the bias? (Cascade vs Direct)

Bentivogli, Luisa, et al. "Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus." *in Proceedings of ACL 2020*

MuST-SHE - Categories

Extract segments from MuST-C EN → IT,FR containing:

1. Gender phenomena related to the speaker

   Nice to meet you! → Encantado/a de conocerte

2. Gender phenomena recognizable from the context

   my mother is the doctor → mi madre es la doctora

MuST-SHE - Categories

Extract segments from MuST-C ing:

Use the speaker's voice to decide

1. Gender phenomena related

   Nice to meet you! → Encantado/a de conocerte

2. Gender phenomena recognizable from the context

   my mother is the doctor → mi madre es la doctora

MuST-SHE - Categories

Extract segments from MuST-C EN → IT,FR containing:

1. Gender phenomena related to the speaker

   Nice to meet you! → Encantado/a de

2. Gender phenomena recognizable from

   my mother is the doctor → mi madre es la doctora

Decide regardless of speaker's voice

MuST-SHE - Balancing

In MuST-C, male speakers are the majority (70%)

To measure bias, we need a balanced test set

Balanced for speakers number **AND** gender phenomena

MuST-SHE

TEST SET

EVALUATION

EXPERIMENTS

Previous Evaluation

BLEU was the main metric to compute improvements in terms of gender bias reduction.

Assumption: Translation differences **only** due to gender phenomena.

MuST-SHE evaluation

1. BLEU
2. Accuracy of gendered words
3. With “Wrong references”!

MuST-SHE evaluation

1. BLEU
2. Accuracy of gendered words
3. With “Wrong references”!

SRC	I was **born** and **brought up** in Mumbai.
C-REF$_{It}$	Sono **nata** e **cresciuta** a Mumbai.
W-REF$_{It}$	Sono **nato** e **cresciuto** a Mumbai.

MuST-SHE Evaluation

- Wrong references are manually produced and provided with MuST-SHE
- Score difference between correct and wrong reference measures the impact of gender bias
- BLEU difference is still difficult to interpret because the impact on higher-order n-grams is unknown
- Accuracy is more focused but relies on presence of exact word

MuST-SHE

TEST SET

EVALUATION

EXPERIMENTS

3 Systems

- **Cascade**: Strong ASR + MT, lost audio information
- **Direct**: Trained on less data, can use audio information
- **Cascade + Tag**: Cascade + correct speaker’s gender information provided.

Results - BLEU

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	21.5	19.7	1.8	20.2	19.3	0.9	22.7	20.0	2.7
	Cascade	24.1	22.4	1.8	22.8	21.9	0.8	25.5	22.8	2.7
	Cascade+Tag	23.8	20.9	2.9	23.0	20.4	2.6	24.5	21.3	3.2

Results - BLEU

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	21.5	19.7	1.8	20.2	19.3	0.9	22.7	20.0	2.7
	Cascade	24.1	22.4	1.8	22.8	21.9	0.8	25.5	22.8	2.7
	Cascade+Tag	23.8	20.9	2.9	23.0	20.4	2.6	24.5	21.3	3.2

Results - BLEU

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	21.5	19.7	1.8	20.2	19.3	0.9	22.7	20.0	2.7
	Cascade	24.1	22.4	1.8	22.8	21.9	0.8	25.5	22.8	2.7
	Cascade+Tag	23.8	20.9	2.9	23.0	20.4	2.6	24.5	21.3	3.2

Results - Accuracy

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	43.3	16.4	26.9	34.2	24.0	10.2	51.3	9.6	41.7
	Cascade	41.1	17.5	23.6	33.7	24.5	9.2	47.6	11.2	36.4
	Cascade+Tag	48.0	10.4	37.6	44.7	14.0	30.7	51.0	7.2	43.8

Results - Accuracy

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	43.3	16.4	26.9	34.2	24.0	10.2	51.3	9.6	41.7
	Cascade	41.1	17.5	23.6	33.7	24.5	9.2	47.6	11.2	36.4
	Cascade+Tag	48.0	10.4	37.6	44.7	14.0	30.7	51.0	7.2	43.8

Results - Accuracy

	Systems	All			Feminine			Masculine		
		Correct	Wrong	Diff	Correct	Wrong	Diff	Correct	Wrong	Diff
En-It	End2End	43.3	16.4	26.9	34.2	24.0	10.2	51.3	9.6	41.7
	Cascade	41.1	17.5	23.6	33.7	24.5	9.2	47.6	11.2	36.4
	Cascade+Tag	48.0	10.4	37.6	44.7	14.0	30.7	51.0	7.2	43.8

Results - Split by Category

	En-it					
	Feminine			Masculine		
	End2End					
	Corr.	Wrong	Diff.	Corr.	Wrong	Diff.
Cat. 1	26.7	27.2	-0.5	46.3	6.8	39.5
Cat. 2	40.6	20.5	20.1	53.9	10.9	43.0
	Cascade					
Cat. 1	15.9	34.5	-18.6	40.0	12.0	28.0
Cat. 2	48.9	15.7	33.2	51.2	10.8	40.4
	Cascade+Tag					
Cat. 1	43.0	10.4	32.6	48.5	2.9	45.6
Cat. 2	46.9	15.5	31.4	51.7	9.6	42.1

Results - Split by Category

	En-it					
	Feminine			Masculine		
	End2End					
	Corr.	Wrong	Diff.	Corr.	Wrong	Diff.
Cat. 1	26.7	27.2	-0.5	46.3	6.8	39.5
Cat. 2	40.6	20.5	20.1	53.9	10.9	43.0
	Cascade					
Cat. 1	15.9	34.5	-18.6	40.0	12.0	28.0
Cat. 2	48.9	15.7	33.2	51.2	10.8	40.4
	Cascade+Tag					
Cat. 1	43.0	10.4	32.6	48.5	2.9	45.6
Cat. 2	46.9	15.5	31.4	51.7	9.6	42.1

Results - Split by Category

	En-it					
	Feminine			Masculine		
	End2End					
	Corr.	Wrong	Diff.	Corr.	Wrong	Diff.
Cat. 1	26.7	27.2	-0.5	46.3	6.8	39.5
Cat. 2	40.6	20.5	20.1	53.9	10.9	43.0
	Cascade					
Cat. 1	15.9	34.5	-18.6	40.0	12.0	28.0
Cat. 2	48.9	15.7	33.2	51.2	10.8	40.4
	Cascade+Tag					
Cat. 1	43.0	10.4	32.6	48.5	2.9	45.6
Cat. 2	46.9	15.5	31.4	51.7	9.6	42.1

MuST-SHE Recap

1. Developed a new test set annotated with gender phenomena and their “wrong reference”
2. Proposed a new evaluation method for gender bias, but more research is needed in this sense
3. Gender information in input (from audio or provided) reduce the errors in the first category (speaker) but confuse the system for the second category (context)

Conclusions

Recap

&

Future Directions

FBK's Research on Direct Speech Translation /1

- Proposed MuST-C, the largest multilingual corpus for ST.
  - Each linguistic section is larger than any other available dataset.
  - Used as training set for IWSLT 2019-2020
- Proposed S-Transformer, an adaptation of Transformer for ST.
  - Improved results on MuST-C.
  - Further improvement from multilingual training and subword-level segmentation in the target side.

FBK’s Research on Direct Speech Translation /2

- Participated to the IWSLT shared tasks on offline speech translation
  - Combined many techniques proposed in literature
  - State-of-the-art system
- Gender bias in Speech Translation
  - Proposed MuST-SHE, a test set to evaluate gender bias
  - Improved evaluation over BLEU score
  - Direct speech translation is promising for using voice information during translation

Future Directions

- Novel data augmentation methods (self-supervised learning is the new trend)
- Gender bias: improve systems, ignore or use speaker info, better evaluation
- Domain coverage: now direct ST only possible for few domains covered by the few and small available datasets

FONDAZIONE BRUNO KESSLER

Recent Developments in Direct Speech Translation

Mattia Antonino Di Gangi, Marco Gaido, Matteo Negri,
Luisa Bentivogli, Beatrice Savoldi, Marco Turchi
surname@fbk.eu

UPC Seminars
July 15 2020

Medium

@fbk_mt @mdigangiPA Machine Translation @ FBK @mattiadigangi