Exploring Phoneme-Level Representations for End-to-End Speech Translation

Elizabeth Salesky[ə], Matthias Sperber[ɣ], Alan W Black[ə]

[ə]CMU, [ɣ]KIT

Carnegie Mellon University
Language Technologies Institute

Speech-to-Text Translation Overview

speech

Speech Recognition (ASR)

ASR

Eres un mago, Harry

transcription

Machine Translation (MT)

MT

Yer a wizard, Harry

translation

Speech-to-Text Translation Overview

speech

Speech Recognition (ASR)

ASR

Eres un mago, Harry

transcription

Machine Translation (MT)

MT

Yer a wizard, Harry

translation

CASCADE

ASR + MT

Speech-to-Text Translation Overview

speech

Speech Recognition (ASR)

ASR

Eres un mago, Harry

transcription

Machine Translation (MT)

MT

Yer a wizard, Harry

translation

CASCADE

ASR + MT

END-TO-END

SLT

Speech-to-Text Translation Overview

speech

Speech Recognition (ASR)

ASR

Eres un mago, Harry

transcription

Machine Translation (MT)

MT

Yer a wizard, Harry

translation

CASCADE

ASR + MT

- Models trained separately, can propagate errors
- Can be more data efficient at low-resource

END-TO-END

SLT

Speech-to-Text Translation Overview

speech

Speech Recognition (ASR)

ASR

Eres un mago, Harry

transcription

Machine Translation (MT)

MT

Yer a wizard, Harry

translation

CASCADE

ASR + MT

- Models trained separately, can propagate errors
- Can be more data efficient at low-resource

END-TO-END

SLT

- May use multiple tasks, but models jointly trained, fewer parameters

Speech-to-Text Translation Overview

speech

Speech Recognition (ASR)

ASR

Eres un mago, Harry

transcription

Machine Translation (MT)

MT

Yer a wizard, Harry

translation

CASCADE

ASR + MT

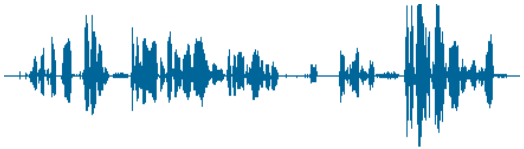
- Models trained separately, can propagate errors
- Can be more data efficient at low-resource

END-TO-END

SLT

- May use multiple tasks, but models jointly trained, fewer parameters

Source speech features

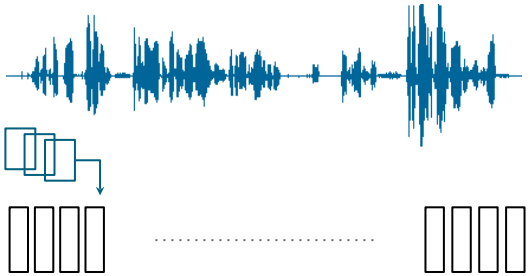


speech

windowing

frame-level
feature vectors

Source speech features

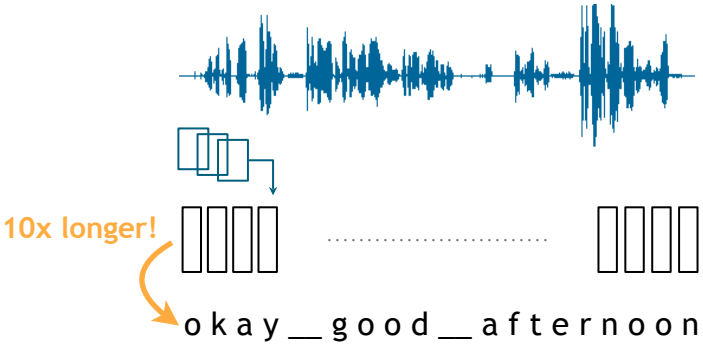


speech

windowing

frame-level
feature vectors

Source speech features



speech

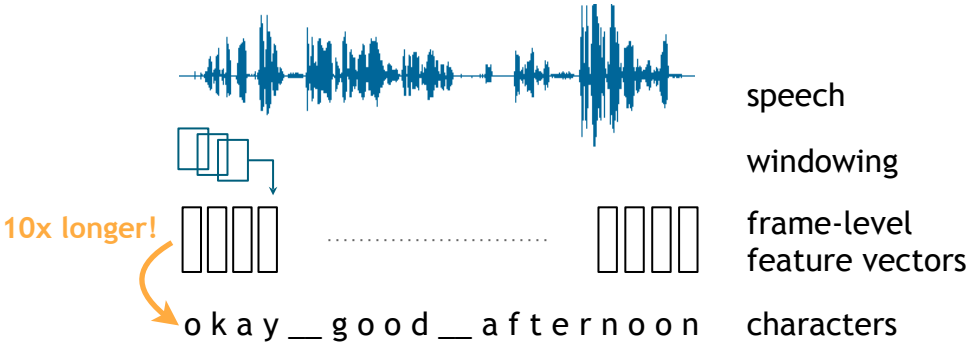
windowing

frame-level feature vectors

characters

- Speech feature sequences (e.g. MFCC, filterbank) are ~10x longer than equivalent character sequences

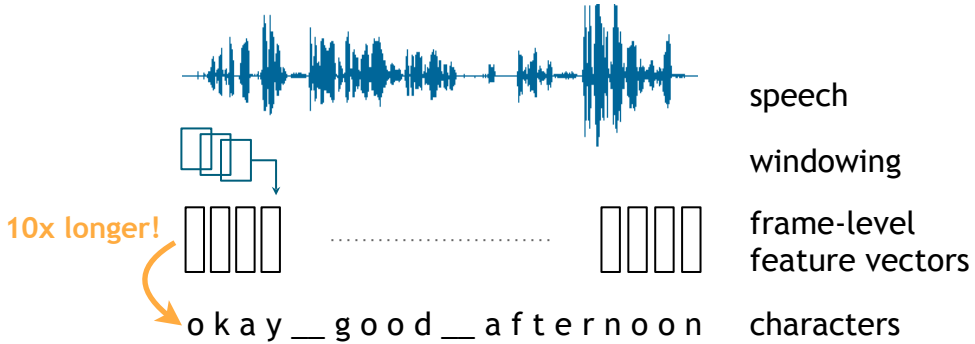
Source speech features



- Speech feature sequences (e.g. MFCC, filterbank) are ~10x longer than equivalent character sequences

  ⟶ *Impacts memory, time, & performance*

Source speech features

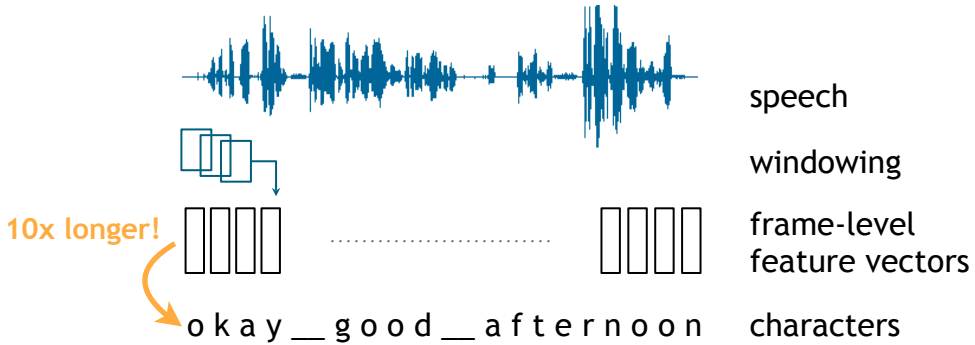


- Speech feature sequences (e.g. MFCC, filterbank) are ~10x longer than equivalent character sequences

  ⟶ *Impacts memory, time, & performance*

- Previous work used ***fixed-length*** downsampling to mitigate these issues
  (e.g. pyramidal encoder (Chan et al. 2015))

Source speech features

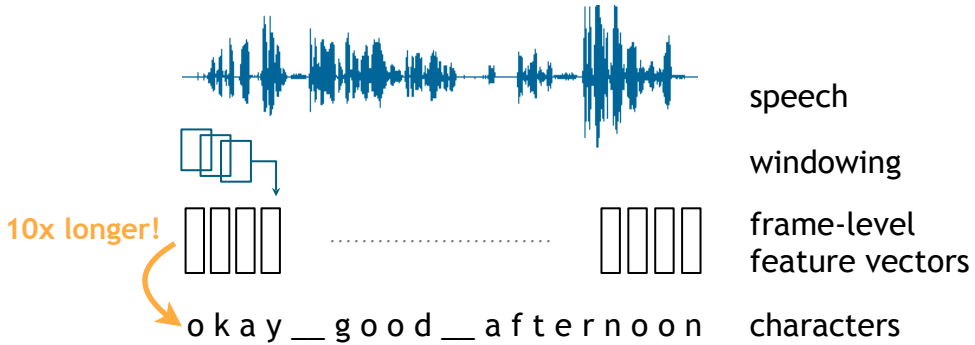


Our Question:
Does instead using linguistic information to do *variable-length* downsampling help?

- Speech feature sequences (e.g. MFCC, filterbank) are ~10x longer than equivalent character sequences
  → *Impacts memory, time, & performance*

- Previous work used ***fixed-length*** downsampling to mitigate these issues
  (e.g. pyramidal encoder (Chan et al. 2015))

Source speech features



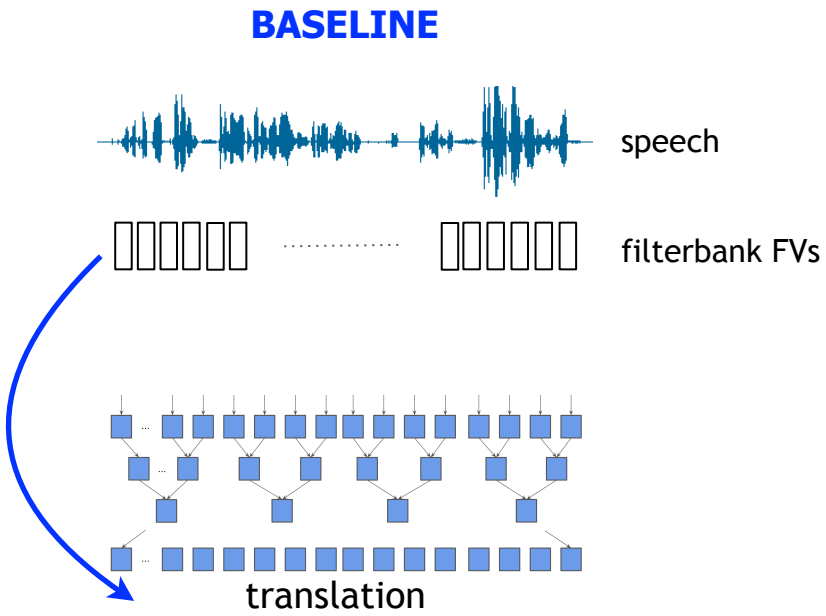
Our Question:
Does instead using linguistic information to do *variable-length* downsampling help?

spoiler alert: yes! in both BLEU and time

- Speech feature sequences (e.g. MFCC, filterbank) are ~10x longer than equivalent character sequences
  → *Impacts memory, time, & performance*

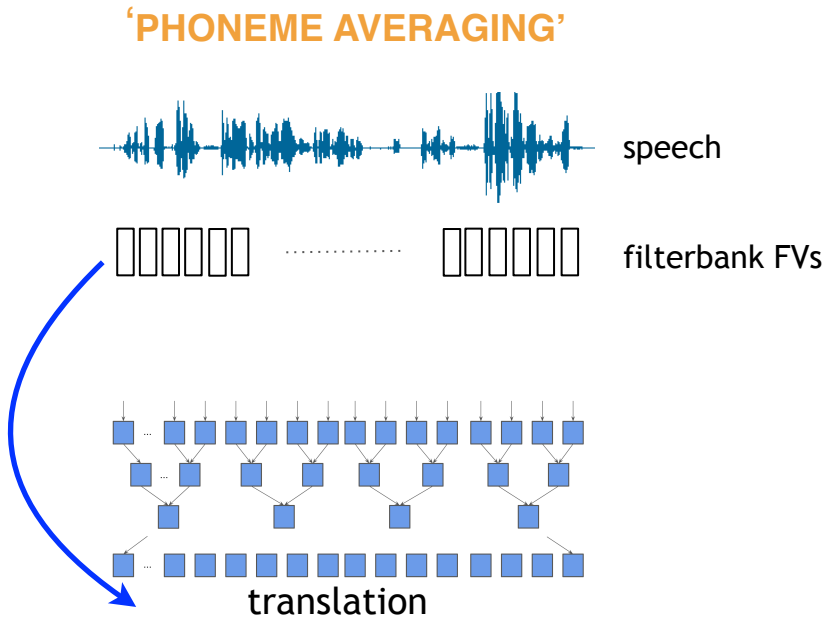
- Previous work used ***fixed-length*** downsampling to mitigate these issues
  (e.g. pyramidal encoder (Chan et al. 2015))

Our Method



Baseline: ‘Frames’

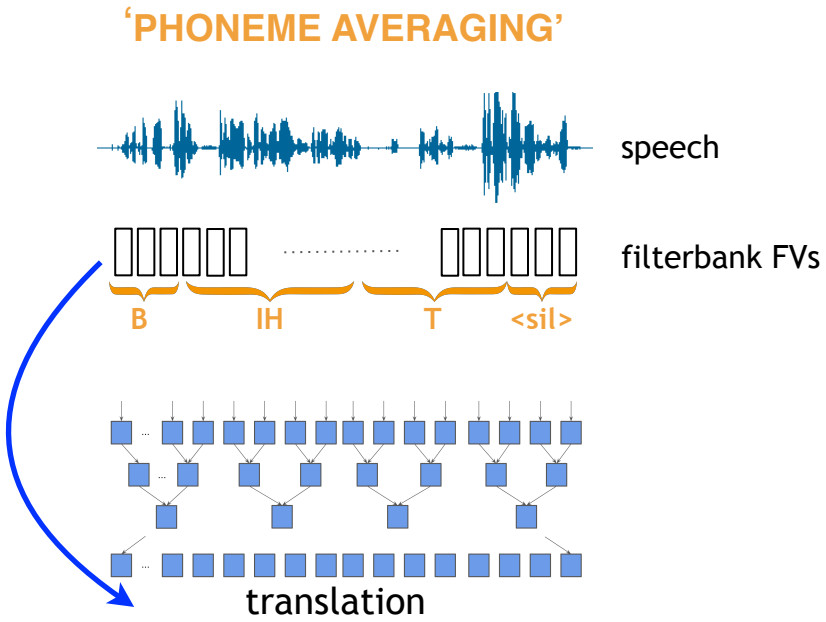
Our Method



- We translate *averaged* filterbank features in place of the full sequence

Baseline: 'Frames'

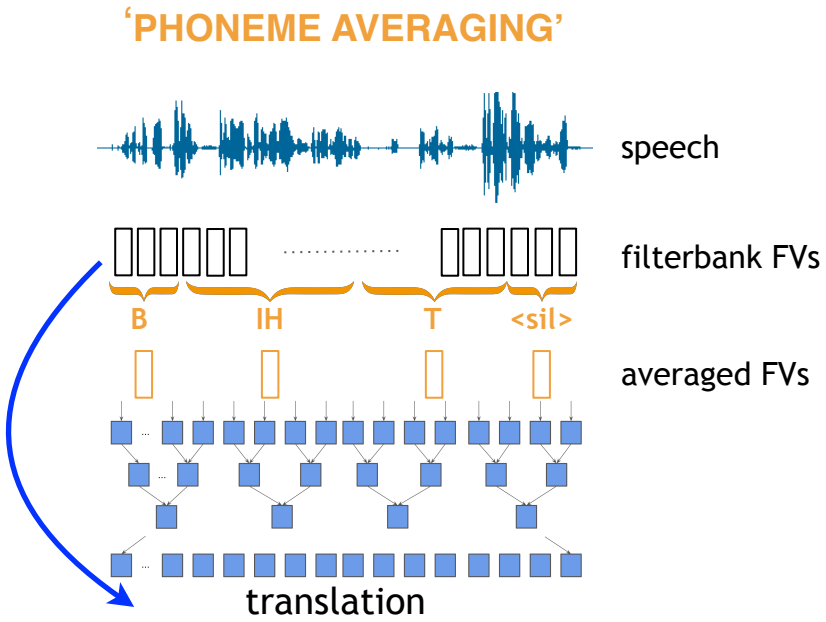
Our Method



- We translate *averaged* filterbank features in place of the full sequence

Baseline: 'Frames'

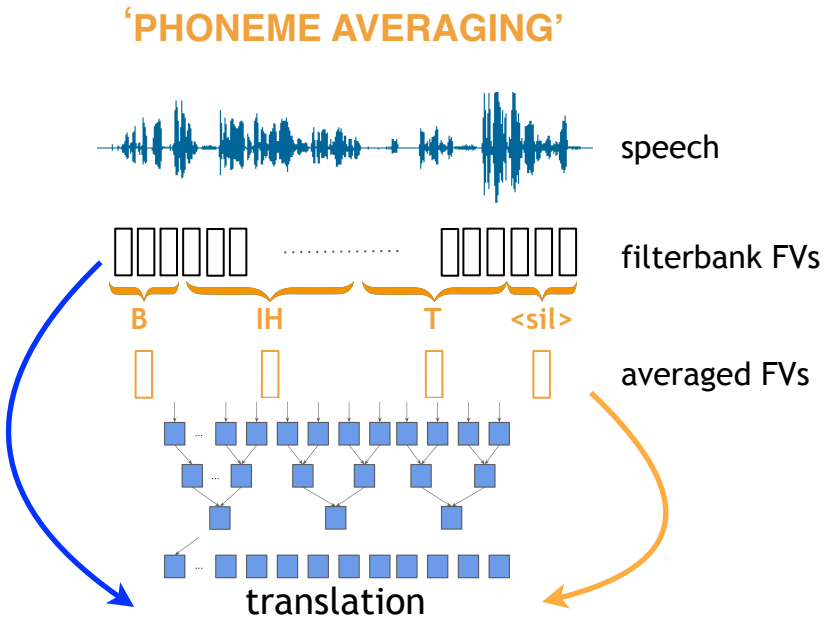
Our Method



- We translate *averaged* filterbank features in place of the full sequence

Baseline: 'Frames'

Our Method

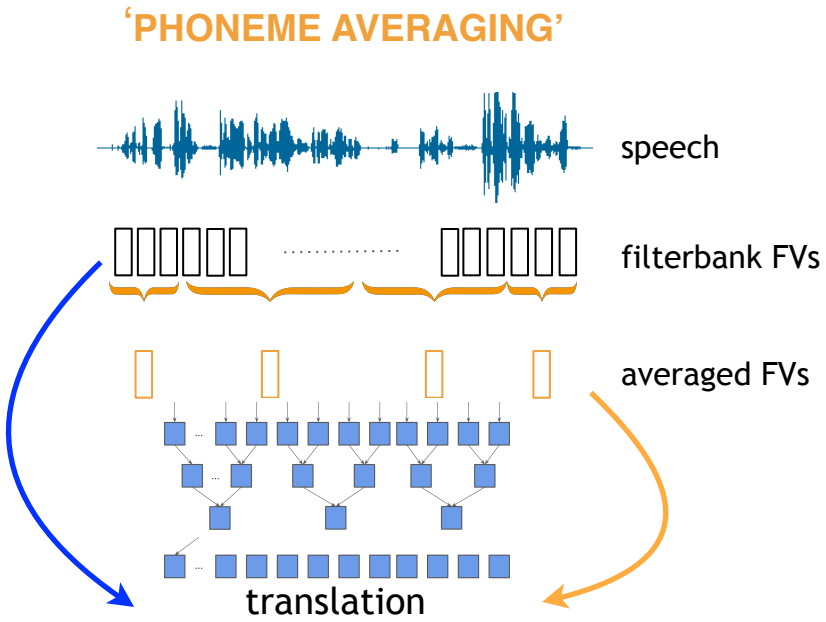


- We translate *averaged* filterbank features in place of the full sequence
- Yields variable-length downsampling: source length reduced by ~80%

Baseline: 'Frames'

Our method: 'Phonemes'

Our Method

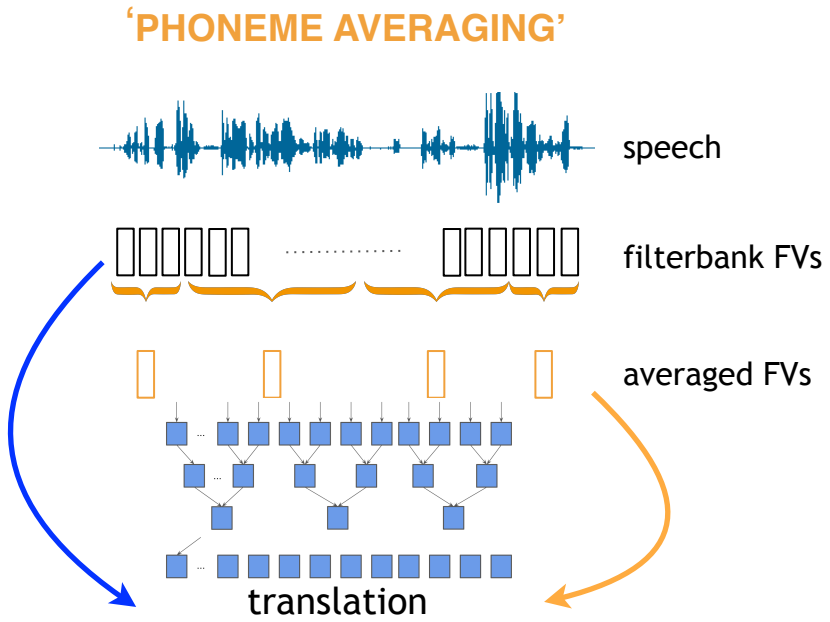


- We translate *averaged* filterbank features in place of the full sequence
- Yields variable-length downsampling: source length reduced by ~80%

Baseline: 'Frames'

Our method: 'Phonemes'

Our Method

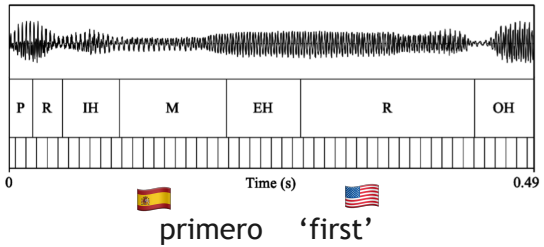


Baseline: 'Frames'

Our method: 'Phonemes'

- We translate *averaged* filterbank features in place of the full sequence

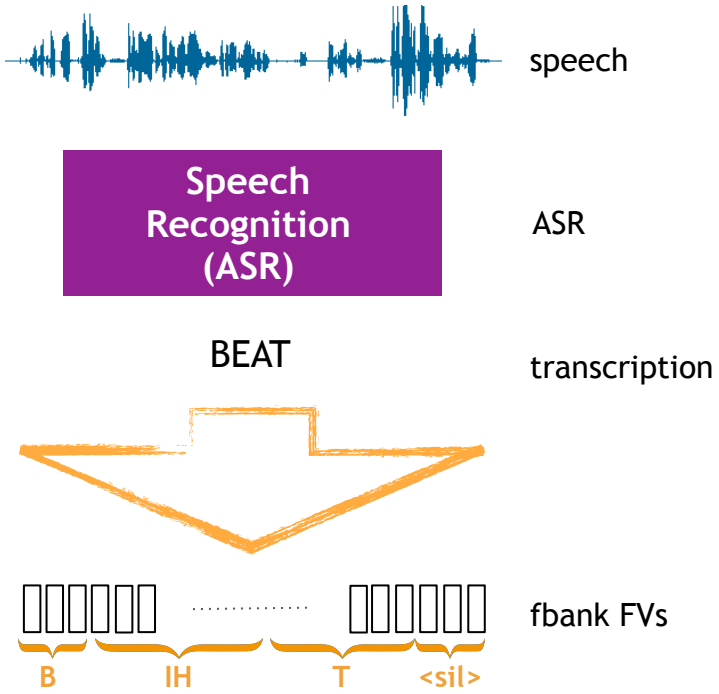
- Yields variable-length downsampling: source length reduced by ~80%



= 8 avg'd FVs

= 50 filterbank FVs

How do we get phoneme boundaries?



- Use a trained recognizer with *generated* transcripts to do alignment

- Requires a decent recognizer, but not necessarily a matched phoneme set:

  we care about the *boundaries* rather than the phoneme labels

Tasks

Fisher Spanish 🇪🇸 — English 🇺🇸

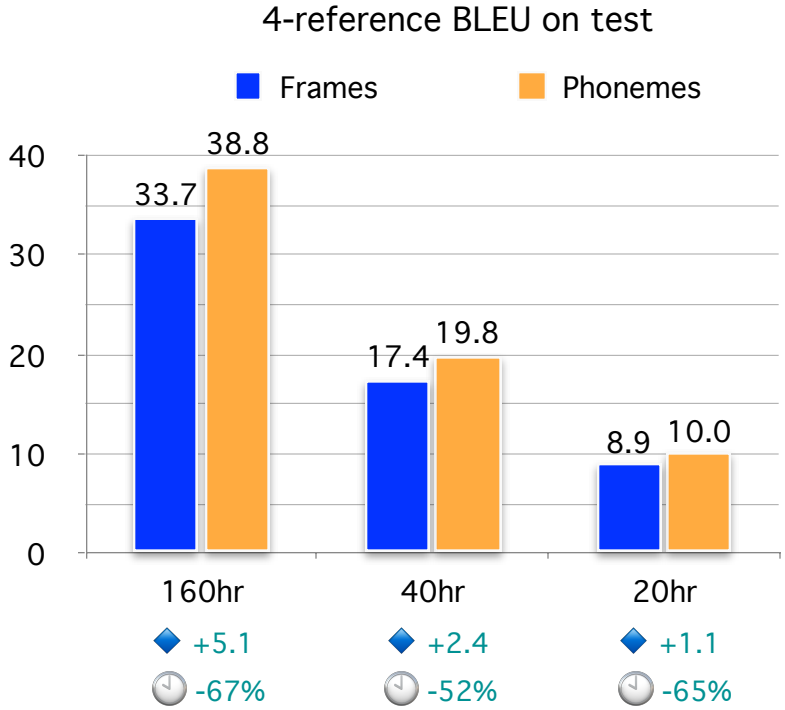
- Parallel speech, transcriptions, and translations
- Use 🇪🇸 Fisher ASR model for alignments
- Compare performance on 3 data sizes:
    1. ~160 hours (full dataset)
    2. 40 hours
    3. 20 hours

Mboshi 🇨🇬 — French 🇫🇷

- Parallel speech and translations
- Use 🇺🇸 ASR model for alignments
- Only 5 hours of speech:
  too little to train a good recognizer

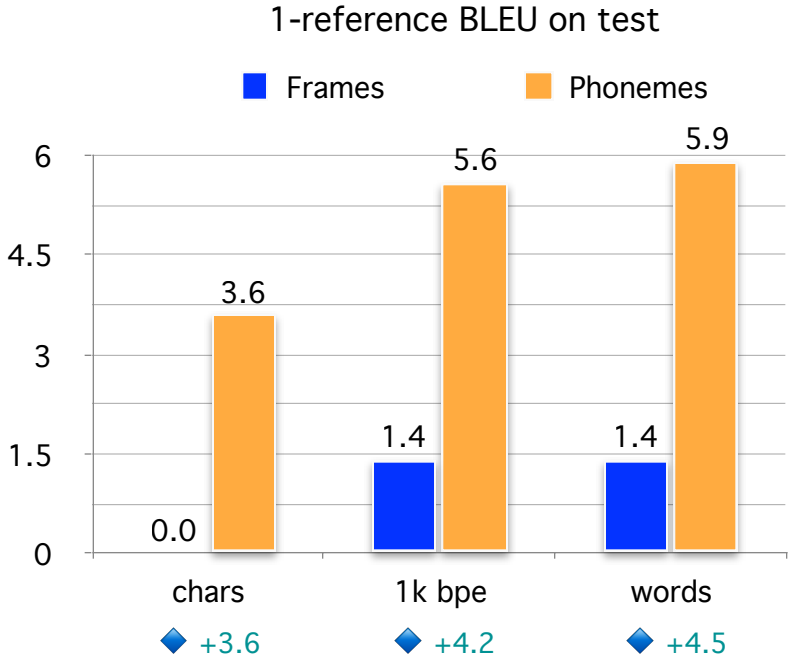
Results — Spanish-English

- Phoneme-representations help all data conditions, not just low-resource
  - Relative improvement of 13% on all data sizes
- Train and inference time is reduced significantly (by ~61%!)
  - We trained our full phoneme model in 39.5 hours on 1 GPU, rather than 118.2



Results — Mboshi-French

- We use an English recognizer to produce phoneme labels (Dalmia et al. 2018a)
- We see big improvements (*though note: this dataset is very small: 200 dev, 500 test*)
  - Noisy phoneme labels but useful phoneme boundaries
- **Main takeaway**: even with noisy phoneme alignments, we still see improvements



Additional Analysis

👀 [See the paper]

1. Does this method help ASR? → *Yes!*
   - Note: unlike translation ASR sequences are monotonic

2. Our method changes relative lengths of source and target
   - How much does this matter?
   - We compare with different target segmentations to see

3. We also compare with fixed source downsampling and removing encoder downsampling

Context with Other Work

- Sperber et al. 2019: Attention-Passing Models for Robust and Data-Efficient End-to-End Speech Translation (TACL)
  - Modeling approach with a similar motivation, using Fisher Spanish-English
  - *Will be presented Wednesday 16:20 in Session 8D: check it out!*

- Bansal et al. 2019: Pre-training on High-Resource Speech Recognition Improves Low-Resource Speech-to-Text Translation (NAACL)
  - Data approach using both Fisher Spanish-English and Mboshi-French corpora

Conclusions

- We used phoneme-boundary information to perform source downsampling for both seq2seq SLT and ASR
  - We showed variable-length downsampling can be more effective (***better performance***) and efficient (***faster training and decoding***) than fixed-length downsampling

- Exploratory work showing these types of features are useful!
  - The ideal level of linguistic or latent information, and the best way to incorporate it, remains to be seen: future work ☺

Thank you! Questions?

EXTRAS

Effects on Training Process

- We see that phoneme-boundary informed embeddings are more **effective** in terms of performance and learn more **efficiently** over the course of training

- The most significant effect occurs at the beginning of training

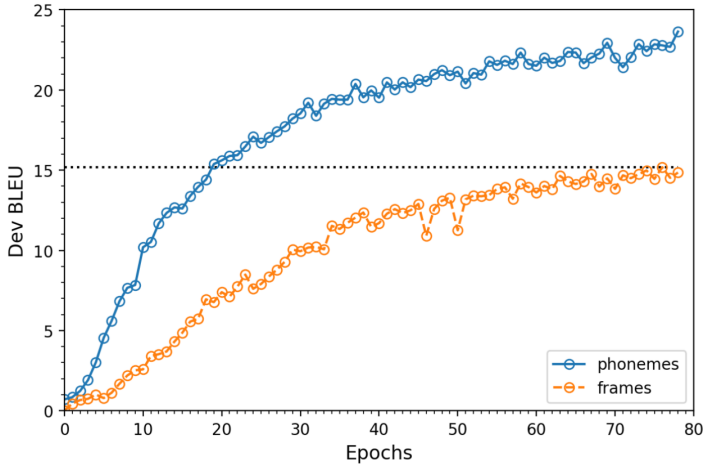
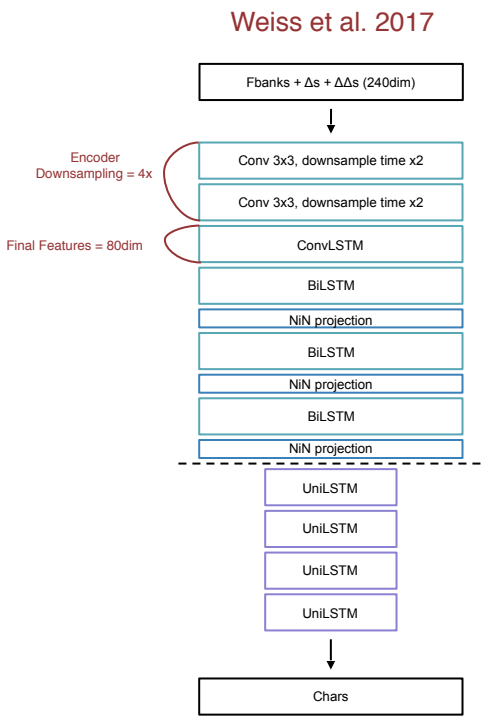
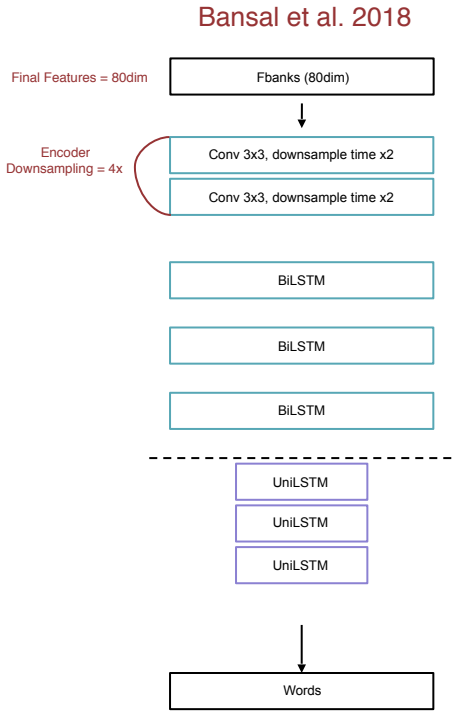
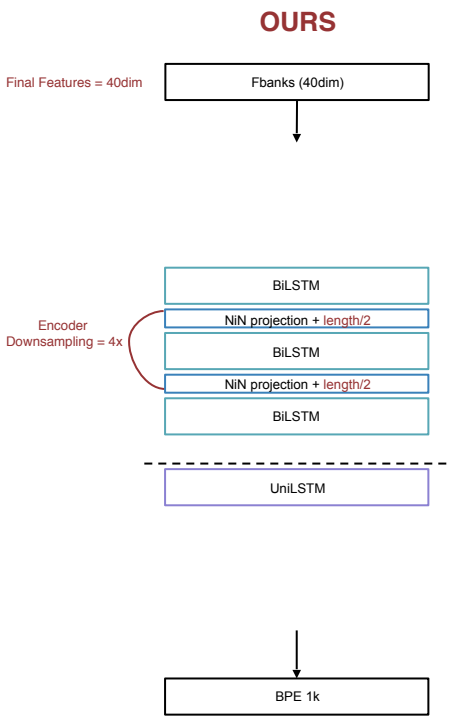


Figure 2: *Dev BLEU over training with frames vs phonemes. Single-reference BLEU on 1k lines of* `dev`.

Model Architecture







Training time: 2.5 weeks

Training time: 5 days

Training time: 5 days

lon University
echnologies Institute

Target Preprocessing

- Previous work on seq2seq ASR and SLT models used encoder downsampling of 4x or 8x, to reduce parameters and create more 1-to-1 final encoder and decoder states
  - We use encoder downsampling of 4x, concatenating adjacent states after each layer
- With our method, source sequence lengths are further reduced by ~79%, yielding final encoder state lengths of 22, closest in length to 1k BPE targets (14) rather than chars (50)
- Given that the 1k BPE model performs best, it does appear that more similar source and target lengths boost performance

Target Preproc.	Target Length	Frames		Phonemes	
		dev	test	dev	test
chars	50.2	18.8	**17.3**	20.0	18.4
1k bpe	13.7	**19.5**	**17.4**	**21.0**	**19.8**
10k bpe	10.6	16.2	14.7	18.4	17.5
words	10.4	16.4	14.6	18.2	17.4

Table 3: *Comparing effects of target preprocessing with different sources on BLEU, Spanish-English 40hr*

Carnegie Mellon University
Language Technologies Institute

Encoder Downsampling

- Our alignment and averaging method creates variable-length source downsampling
- We compare our variable-length downsampling to fixed-stride downsampling
- On 40 hours, we see fixed-stride downsampling hurts, while our variable-length downsampling helps:

Stride	dev	test
1	19.5	17.4
2	17.0	15.6
3	13.7	11.8
Variable	21.0	19.8

- Phoneme-informed reduction is clearly more effective than fixed schedule downsampling

Attention Passing

- A multi-stage end-to-end trainable model where the input to the MT model is the attention context vectors from ASR (Sperber et al. 2019)
- Our method achieves comparable results to this model when it is extended with additional data (parallel text from OpenSubtitles):
  - 37.6 on dev, 38.8 on test

Model	BLEU
Cascade	32.45
Direct	35.30
Basic two-stage	34.36
APM	35.31
APM + cross connections	36.51
APM + cross conn. + additional loss	**36.70**
Best APM w/o block dropout	36.04

Table 2: Results for cascaded and multi-task models under full training data conditions.

Model	Fisher	Fisher+OpenSub
Cascade	32.45	34.58 (+6.2% rel.)
Direct model	35.30	36.45 (+3.2% rel.)
Basic two-stage	34.36	36.91 (+6.9% rel.)
Best APM	36.70	38.81 (+5.4% rel.)

Table 3: Adding auxiliary OpenSubtitles MT data to the training. The two-stage models benefit much more strongly than the direct model, with our proposed model yielding the strongest overall results.

Carnegie Mellon University
Language Technologies Institute

Baseline Numbers

- Weiss et al. is a significantly deeper network taking >2 weeks to train
- Bansal et al. is more similar to ours, with reductions in network size to also train in ~5 days
- Shows reasonable results before approaching our method

	Weiss et al. (2017)		Bansal et al. (2018a)		Ours	
	dev	**test**	**dev**	**test**	**dev**	**test**
BLEU	46.5	47.3	29.5	29.4	32.4	33.7

Table 1: *Single task end-to-end speech translation BLEU scores on full dataset.*

Encoder Downsampling

- Our alignment and averaging method creates variable-length source downsampling
- We compare our variable-length downsampling to fixed-stride downsampling
- On 40 hours, we see fixed-stride downsampling hurts, while our variable-length downsampling helps:

Stride	dev	test
1	19.5	17.4
2	17.0	15.6
3	13.7	11.8
Variable	21.0	19.8

- Phoneme-informed reduction is clearly more effective than fixed schedule downsampling

ASR: Attention

- To test whether our approach yields generally more effective input representations, or chiefly helps for SLT where reducing the distance between inputs and outputs which may be reordered, we apply our method to ASR, where are alignments are monotonic

- We see ~18% relative improvement on all three data conditions, similar to SLT

	Frames		Phonemes		WER	Time
Data	**dev**	**test**	**dev**	**test**	Δ	Δ
Full	33.4	30.0	28.0 **–5.4**	23.4 **–6.6**	–6.0	–43%
40hr	44.8	46.7	36.6 **–8.2**	36.6 **–10.1**	–9.2	–40%
20hr	56.3	59.1	48.2 **–8.1**	49.1 **–10.0**	–9.1	–50%

Table 5: *Comparison of frame vs phoneme input on Spanish ASR, with average reduction in WER and average reduction in training time.*

(single reference for ASR)

Spanish 5 hour

- To test the Spanish data in more similar conditions to Mboshi-French, we also tested 5 hours of Spanish, with single reference scores

- They were in a similar range to the Mboshi scores, with some improvement with targets of 1k bpe and words, but not so marked as other conditions

source	target	Fisher, 5hr dev	Fisher, 5hr test
frames	chars	1.4	1.4
frames	1k bpe	1.7	1.7
frames	words	1.1	1.1
phonemes	chars	0.9	1.1
phonemes	1k bpe	1.7	2.1
phonemes	words	1.6	1.9

Table 6: *BLEU, Fisher Spanish-English, 5 hours*

Datasets

Spanish-English.

We use the common Fisher Spanish-English dataset which consists of ~160 hours of Spanish telephone speech, translated via crowdsourcing.

The training data is split into 138K utterances.

We use the standard dev and test sets, each with ~4k utterances. We do not use dev2.

4 reference translations used to score outputs

Mboshi-French.

We use the Mboshi-French parallel corpus for our low-resource setting, which has <5 hours of speech split into training and development sets of 4616 and 500 utterances respectively.

No designated test set, so as in Bansal et al. (2018b) we removed 200 randomly sampled utterances from training for dev and use the designated dev set as test.

1 reference translation used to score outputs