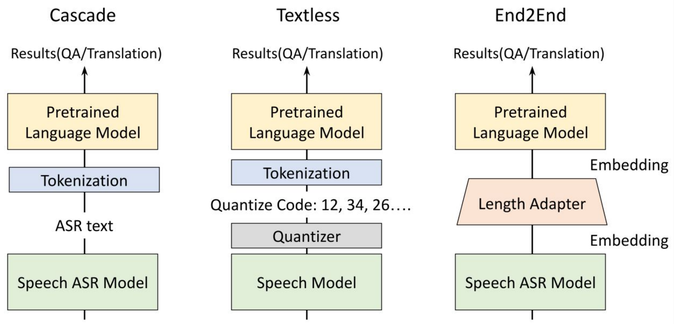
T5lephone: Bridging Speech and Text Self-supervised Models for Spoken Language Understanding via Phoneme level T5

Chan-Jan Hsu*, Ho-Lam Chung* Hung-Yi Lee, Yu Tsao

Background and Motivation

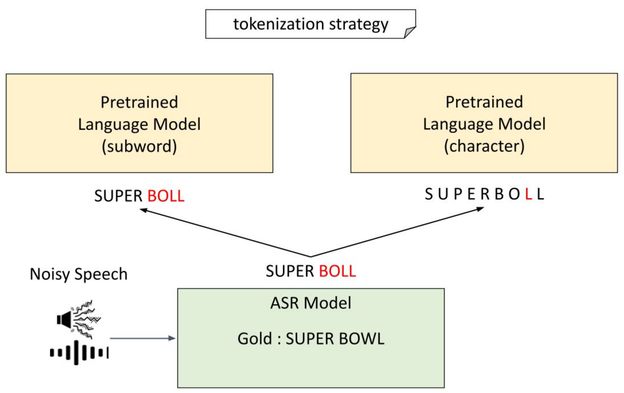
- Spoken Language Understanding often combines pretrained speech models and PLMs like T5.
- How PLMs granularity Effect Spoken Language Understanding



Why is granularity an important factor?

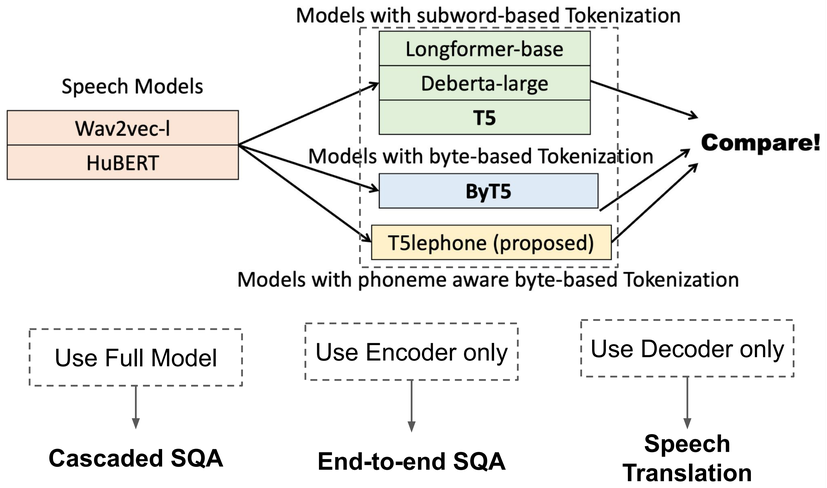
- Granularity is key for precise error spotting.
- Minor errors are caught in a cascaded approach.

Better alignment results from similar length/vocabulary size.

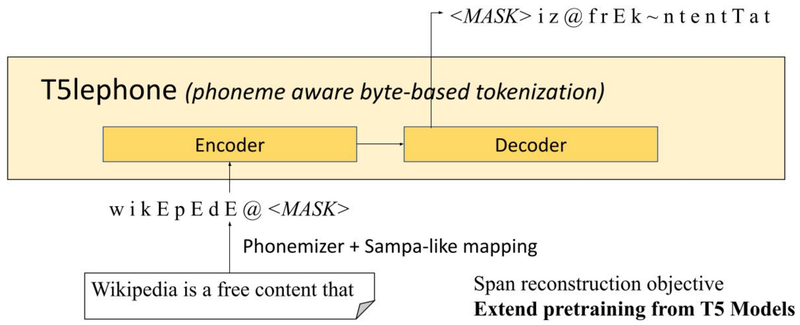


- Most systems use PLM with subword-based tokenization, PLM with character-based tokenization is underexplored
- Developed T5lephone, a variant of T5 using phonemicized text with self-supervised pretraining.

Method - How does input granularity affect the performance?



T5lephone - PLM with phonemcized inputs



- Developed T5lephone: accepts phoneme sequences, initialized with mT5/ByT5.
- Second-stage pretraining used phonemicized wiki text and original reconstruction objective.
- translate phonemes into ASCII characters, similar to SAMPA.
- This approach simplifies phoneme sequences, making them text-like.
- It's a more efficient method than using IPA symbols.

Experiments - Cascaded SQA

Corpuse: Natural Multi-speakers Spoken Question Answering (NMSQA) dataset.

	PLM	tokenization	#Params	text dev EM	text dev F1	dev AOS	dev FF1	test-SQuAD AOS	test-SQuAD FF1	test-OOD AOS	test-OOD FF1
subword	longformer-base	subword	148M	85.0	91.9	47.7	58.6	50.1	62.5	43.9	53.6
	deberta-large	subword	405M	87.9	93.9	42.0	52.2	48.6	61.5	33.1	42.5
	T5-small	subword	61M	78.9	86.1	49.3	55.5	45.3	53.2	35.2	45.1
	T5-base	subword	222M	83.0	89.9	55.6	62.8	66.6	73.9	37.9	44.3
	T5-large	subword	770M	84.2	91.8	**65.2**	**70.0**	66.5	72.5	52.4	56.3
character	ByT5-small	byte	299M	78.4	83.9	60.0	64.7	69.9	74.1	53.9	57.7
	ByT5-base	byte	581M	80.6	87.0	64.0	68.8	68.6	73.5	**60.9**	**66.1**
phoneme	**ByT5lephone-small**	byte	299M	76.7	83.7	59.2	64.4	**70.5**	**75.5**	58.3	63.3

- Generally, ByT5lephone(ours) > Byte models > Subword models
- When ASR noise ↑, gap between ByT5lephone and other models is larger

Experiments - End-to-end SQA

Corpuse: NMSQA dataset.

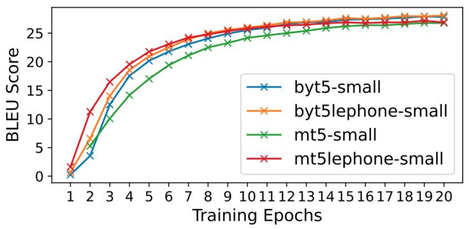
End-to-end Spoken Question Answering					
Text Model (enc. only)	Len	test-SQuAD AOS	test-SQuAD FF1	test-OOD AOS	test-OOD FF1
longformer[11]	4096	49.1	55.9	-	-
longformer	4096	46.0	53.9	32.2	36.9
longformer	1024	40.0	46.0	22.8	26.4
ByT5-small	1024	48.4	54.9	27.1	31.0
ByT5lephone-small	1024	**53.3**	**61.1**	**32.3**	**37.3**

- ByT5lephone(ours) significantly outperforms longformer (even 4096)
- Long context length > Short context length
- T5/mT5 -> context span too short to solve this task

Experiments - Speech Translation

Corpuse: Covost2

Speech to Text Translation En− >De		
Text Model (dec. only)	#Params	BLEU
mT5-small	153M	26.8
mT5lephone-small	153M	**27.2**
ByT5-small	82M	27.9
ByT5lephone-small	82M	**28.1**



- T5lephone(ours) slightly outperform their respective base model.
- T5lephone improves much faster than their respective T5 baseline models in the early epochs.

Conclusion

- Validated byte-level models in cascaded SQA.
- Extended concept with second-phase pretraining on phonemicized text, boosting performance.

T5lephone: Bridging Speech and Text Self-supervised Models for Spoken Language Understanding via Phoneme level T5

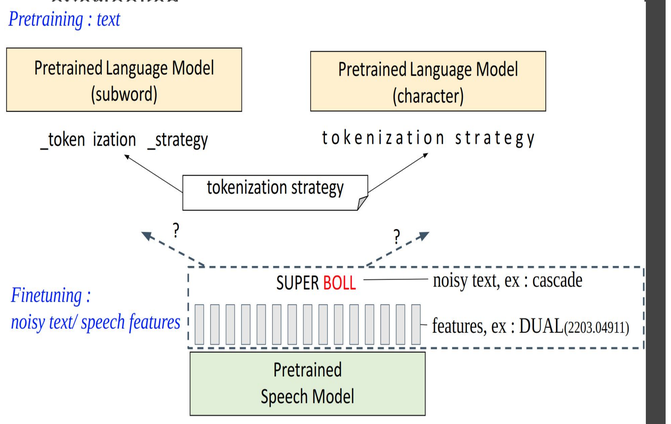
Chan-Jan Hsu*, Ho-Lam Chung* Hung-Yi Lee, Yu Tsao

Goal

- Conducted a study comparing T5/mT5 and ByT5 on SQA/ST datasets like NMSQA and Covost2.
- Developed T5lephone, a variant of T5 using phonemicized text with self-supervised pretraining.
- Achieved state-of-the-art, +12% NMSQA gain, surpassing previous methods with fewer parameters.

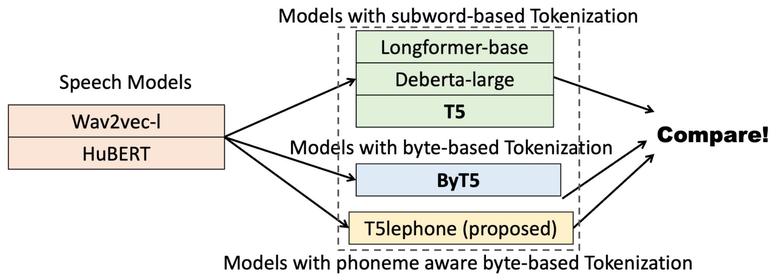
Background and Motivation

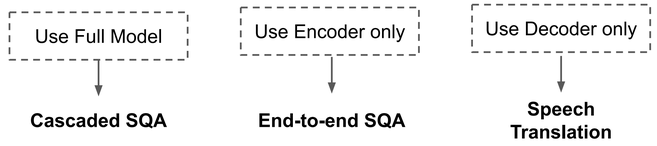
- Spoken Language Understanding often combines pretrained speech models and PLMs like T5.
- Most PLMs are subword based -> Effects of input granularity on speech-language model alignment and character tokenization overlooked



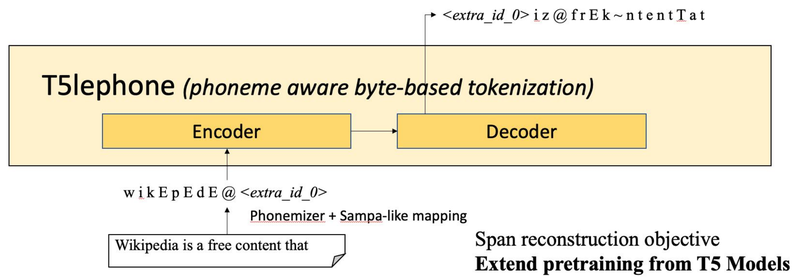
- Why is granularity an important factor?
  - Cascaded: Minor Errors
  - End-to-end (without speech-text pair): Better Alignment
    - Alignment : Similar in Length/ Similar in vocabulary.

Method - How does input granularity affect the performance?





T5lephone - PLM with phonemcized inputs



- Developed T5lephone: accepts phoneme sequences, initialized with mT5/ByT5.
- Second-stage pretraining used phonemicized wiki text and original reconstruction objective.
- Represent phonemes with ASCII characters (similar to SAMPA mapping), to make sequence closer to normal text, beats using IPA symbols.

Experiments - Cascaded SQA

	PLM	tokenization	#Params	text dev		dev		test-SQuAD		test-OOD	
				EM	F1	AOS	FF1	AOS	FF1	AOS	FF1
subword	longformer-base	subword	148M	85.0	91.9	47.7	58.6	50.1	62.5	43.9	53.6
	deberta-large	subword	405M	87.9	93.9	42.0	52.2	48.6	61.5	33.1	42.5
	T5-small	subword	61M	78.9	86.1	49.3	55.5	45.3	53.2	35.2	45.1
	T5-base	subword	222M	83.0	89.9	55.6	62.8	66.6	73.9	37.9	44.3
	T5-large	subword	770M	84.2	91.8	**65.2**	**70.0**	66.5	72.5	52.4	56.3
character	ByT5-small	byte	299M	78.4	83.9	60.0	64.7	69.9	74.1	53.9	57.7
	ByT5-base	byte	581M	80.6	87.0	64.0	68.8	68.6	73.5	**60.9**	**66.1**
phoneme	**ByT5lephone-small**	byte	299M	76.7	83.7	59.2	64.4	**70.5**	**75.5**	58.3	63.3

- Generally, ByT5lephone(ours) > Byte models > Subword models
- When ASR noise ↑, gap between ByT5lephone and other models is larger

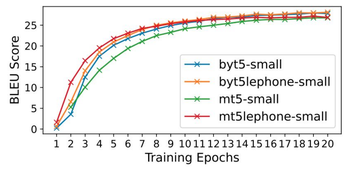
Experiments - End-to-end SQA

End-to-end Spoken Question Answering					
Text Model (enc. only)	Len	test-SQuAD		test-OOD	
		AOS	FF1	AOS	FF1
longformer[11]	4096	49.1	55.9	-	-
longformer	4096	46.0	53.9	32.2	36.9
longformer	1024	40.0	46.0	22.8	26.4
ByT5-small	1024	48.4	54.9	27.1	31.0
ByT5lephone-small	1024	**53.3**	**61.1**	**32.3**	**37.3**

- ByT5lephone(ours) significantly outperforms longformer (even 4096)
- Long context length > Short context length
- T5/mT5 -> context span too short to solve this task

Experiments - Speech Translation

Speech to Text Translation En−>De		
Text Model (dec. only)	#Params	BLEU
mT5-small	153M	26.8
mT5lephone-small	153M	**27.2**
ByT5-small	82M	27.9
ByT5lephone-small	82M	**28.1**



- T5lephone(ours) slightly outperform their respective base model.
- T5lephone improves much faster than their respective T5 baseline models in the early epochs.

Conclusion

- Validated byte-level models in cascaded SQA.
- Extended concept with second-phase pretraining on phonemicized text, boosting performance.

T5lephone: Bridging Speech and Text Self-supervised Models for Spoken Language Understanding via Phoneme level T5

Chan-Jan Hsu*, Ho-Lam Chung* Hung-Yi Lee, Yu Tsao

Goal

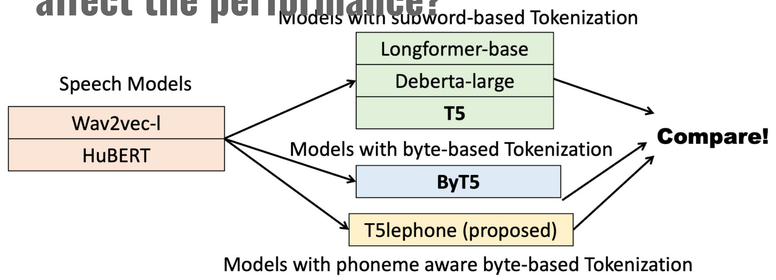
- Conducted a study comparing T5/mT5 and ByT5 on SQA/ST datasets like NMSQA and Covost2.
- Developed T5lephone, a variant of T5 using phonemicized text with self-supervised pretraining.
- Achieved state-of-the-art, +12% NMSQA gain, surpassing previous methods with fewer parameters.

Background and Motivation

- Spoken Language Understanding often combines pretrained speech models and PLMs like T5.
- Overlooked: effects of input granularity on speech-language model alignment and character tokenization.

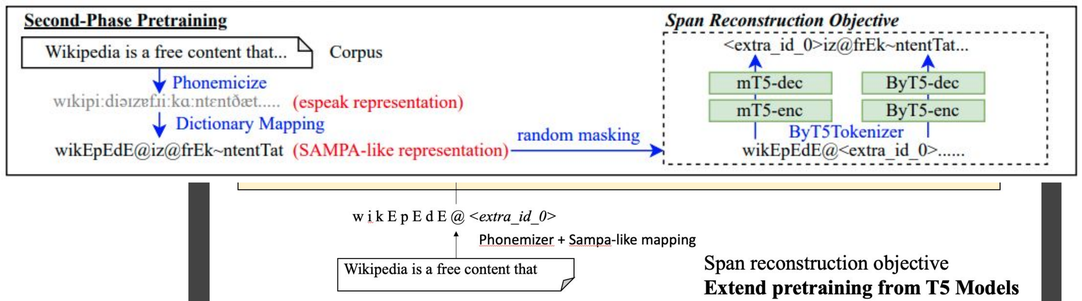
Most systems use PLM with subword-based tokenization, PLM with character-based tokenization is underexplored

Method - How does input granularity affect the performance?



- Benchmarked SQA (NMSQA) and ST (Covost2) datasets.
- Evaluated subword-level and character-level models.

T5lephone - PLM with phonemcized



- Developed T5lephone: accepts phoneme sequences, initialized with mT5/ByT5.
- Second-stage pretraining used phonemicized wiki text and original reconstruction objective.

Experiments - Cascaded SQA

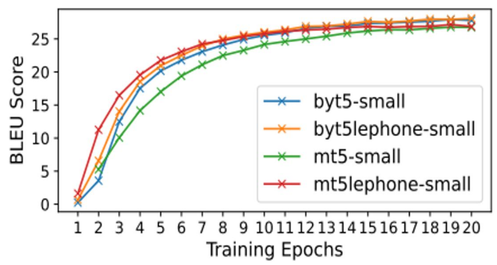
	PLM	#Params	text dev		dev		test-SQuAD		test-OOD	
			EM	F1	AOS	FF1	AOS	FF1	AOS	FF1
subword	longformer-base	148M	85.0	91.9	44.8	55.3	50.7	63.4	38.4	47.3
subword	deberta-large	405M	87.9	93.9	38.3	49.1	47.5	60.3	28.7	37.6
subword	T5-small	61M	78.9	86.1	44.7	51.3	44.6	51.5	29.0	35.1
subword	T5-base	222M	83.0	89.9	52.8	60.4	58.2	66.7	37.6	43.7
subword	T5-large	770M	84.2	91.8	57.6	62.8	58.8	65.5	48.4	52.4
character	ByT5-small	299M	78.4	83.9	55.4	60.6	62.8	67.7	41.3	45.4
phoneme	**ByT5lephone-small**	299M	76.7	83.7	55.0	60.8	**70.1**	**76.3**	**48.6**	**53.2**
character	ByT5-base	581M	80.6	87.0	**59.7**	**65.3**	65.2	69.7	**48.4**	**53.3**

Experiments - End-to-end SQA

End-to-end Spoken Question Answering					
Text Model (enc. only)	**Len**	**test-SQuAD**		**test-OOD**	
		AOS	FF1	AOS	FF1
longformer[11]	4096	49.1	55.9	-	-
longformer	4096	46.0	53.9	32.2	36.9
longformer	1024	40.0	46.0	22.8	26.4
ByT5-small	1024	48.4	54.9	27.1	31.0
ByT5lephone-small	1024	**53.3**	**61.1**	**32.3**	**37.3**

Experiments - Speech Translation

Text Model (dec. only)	**#Params**	BLEU
mT5-small	153M	26.8
mT5lephone-small	153M	27.2
ByT5-small	82M	27.9
ByT5lephone-small	82M	28.1



Conclusion

- Validated byte-level models in cascaded SQA.
- Extended concept with second-phase pretraining on phonemicized text, boosting performance.

T5lephone: Bridging Speech and Text Self-supervised Models for Spoken Language Understanding via Phoneme level T5

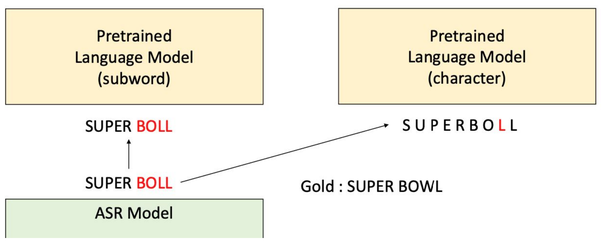
Chan-Jan Hsu*, Ho-Lam Chung* Hung-Yi Lee, Yu Tsao

Goal

- Conducted a study comparing T5/mT5 and ByT5 on SQA/ST datasets like NMSQA and Covost2.
- Developed T5lephone, a variant of T5 using phonemicized text with self-supervised pretraining.
- Achieved state-of-the-art, +12% NMSQA gain, surpassing previous methods with fewer parameters.

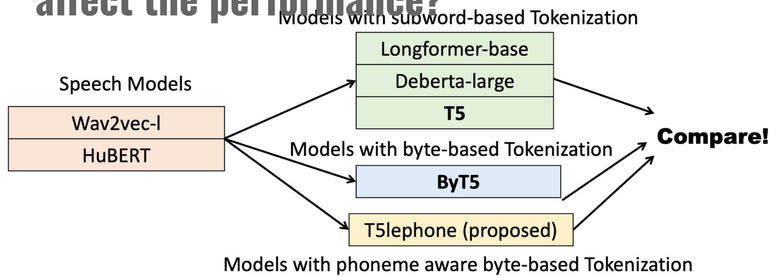
Background and Motivation

- Spoken Language Understanding often combines pretrained speech models and PLMs like T5.
- Overlooked: effects of input granularity on speech-language model alignment and character tokenization.



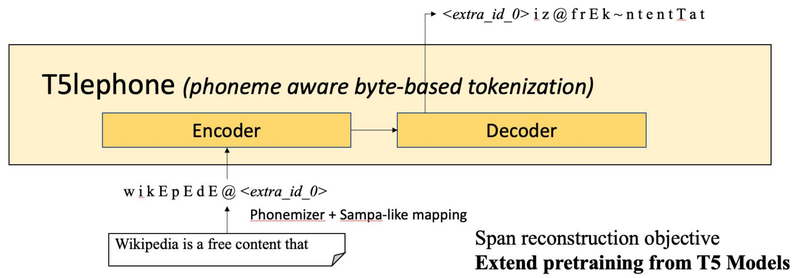
Most systems use PLM with subword-based tokenization, PLM with character-based tokenization is underexplored

Method - How does input granularity affect the performance?



- Benchmarked SQA (NMSQA) and ST (Covost2) datasets.
- Evaluated subword-level and character-level models.

T5lephone - PLM with phonemcized inputs



- Developed T5lephone: accepts phoneme sequences, initialized with mT5/ByT5.
- Second-stage pretraining used phonemicized wiki text and original reconstruction objective.

Experiments - Cascaded SQA

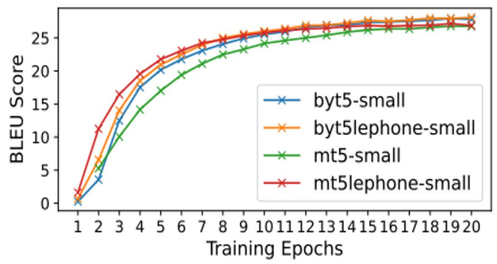
	PLM	#Params	text dev		dev		test-SQuAD		test-OOD	
			EM	F1	AOS	FF1	AOS	FF1	AOS	FF1
subword	longformer-base	148M	85.0	91.9	44.8	55.3	50.7	63.4	38.4	47.3
subword	deberta-large	405M	87.9	93.9	38.3	49.1	47.5	60.3	28.7	37.6
subword	T5-small	61M	78.9	86.1	44.7	51.3	44.6	51.5	29.0	35.1
subword	T5-base	222M	83.0	89.9	52.8	60.4	58.2	66.7	37.6	43.7
subword	T5-large	770M	84.2	91.8	57.6	62.8	58.8	65.5	48.4	52.4
character	ByT5-small	299M	78.4	83.9	55.4	60.6	62.8	67.7	41.3	45.4
phoneme	**ByT5lephone-small**	299M	76.7	83.7	55.0	60.8	**70.1**	**76.3**	**48.6**	**53.2**
character	ByT5-base	581M	80.6	87.0	**59.7**	**65.3**	65.2	69.7	**48.4**	**53.3**

Experiments - End-to-end SQA

End-to-end Spoken Question Answering					
Text Model (enc. only)	**Len**	**test-SQuAD**		**test-OOD**	
		AOS	FF1	AOS	FF1
longformer[11]	4096	49.1	55.9	-	-
longformer	4096	46.0	53.9	32.2	36.9
longformer	1024	40.0	46.0	22.8	26.4
ByT5-small	1024	48.4	54.9	27.1	31.0
ByT5lephone-small	1024	**53.3**	**61.1**	**32.3**	**37.3**

Experiments - Speech Translation

Text Model (dec. only)	**#Params**	BLEU
mT5-small	153M	26.8
mT5lephone-small	153M	27.2
ByT5-small	82M	27.9
ByT5lephone-small	82M	28.1



Conclusion

- Validated byte-level models in cascaded SQA.
- Extended concept with second-phase pretraining on phonemicized text, boosting performance.

T5lephone: Bridging Speech and Text Self-supervised Models for Spoken Language Understanding via Phoneme level T5

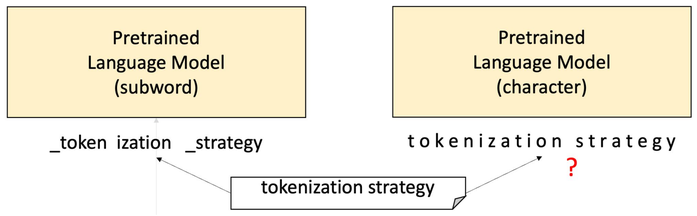
Chan-Jan Hsu*, Ho-Lam Chung* Hung-Yi Lee, Yu Tsao

Goal

We studied the impact of self-supervised Pretrained Language Models (PLMs) on SQA/ST performance, comparing T5/mT5 with ByT5, and introduced T5lephone, a phonemicized text-based T5 variant. Using unique re-representation of phonemicized text, we achieved state-of-the-art results, a +12% NMSQA performance gain with fewer parameters, surpassing previous methods.

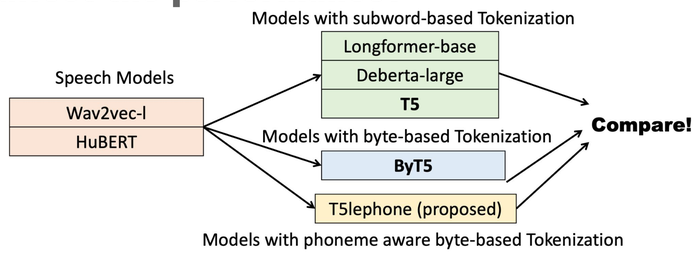
Background and Motivation

In Spoken Language Understanding (SLU), the combination of pretrained speech models (e.g., HuBERT) with PLMs like T5 is common. However, the effects of input granularity on speech and language model alignment are overlooked, particularly with character-based tokenization in PLMs.



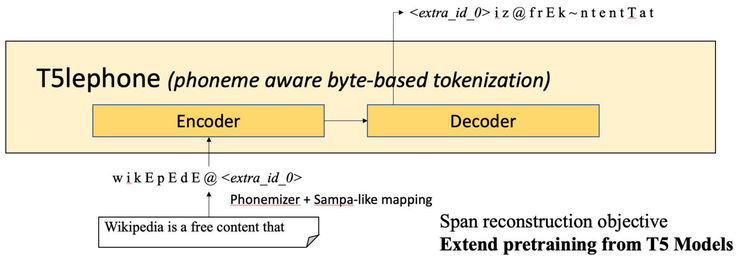
Most systems use PLM with subword-based tokenization, PLM with character-based tokenization is underexplored

Method - How does input granularity affect the performance?



Benchmark SQA Dataset (NMSQA) and ST Dataest (Covost2) and subword-level models and character-level models

T5lephone - PLM with phonemcized inputs



We developed T5lephone, a model that accepts phoneme sequences, by initializing with mT5/ByT5 and conducting a second-stage pretraining using phonemicized wiki text inputs with the original span reconstruction objective.

Experiments - Cascaded SQA

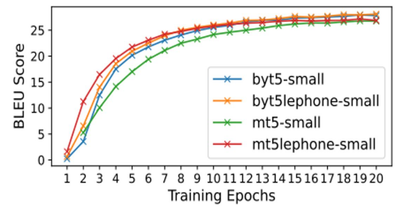
	PLM	#Params	text dev EM	text dev F1	dev AOS	dev FF1	test-SQuAD AOS	test-SQuAD FF1	test-OOD AOS	test-OOD FF1
subword	longformer-base	148M	85.0	91.9	44.8	55.3	50.7	63.4	38.4	47.3
	deberta-large	405M	87.9	93.9	38.3	49.1	47.5	60.3	28.7	37.6
	T5-small	61M	78.9	86.1	44.7	51.3	44.6	51.5	29.0	35.1
	T5-base	222M	83.0	89.9	52.8	60.4	58.2	66.7	37.6	43.7
	T5-large	770M	84.2	91.8	57.6	62.8	58.8	65.5	48.4	52.4
character phoneme character	ByT5-small	299M	78.4	83.9	55.4	60.6	62.8	67.7	41.3	45.4
	ByT5lephone-small	299M	76.7	83.7	55.0	60.8	**70.1**	**76.3**	**48.6**	**53.2**
	ByT5-base	581M	80.6	87.0	**59.7**	**65.3**	65.2	69.7	**48.4**	**53.3**

Experiments - End-to-end SQA

End-to-end Spoken Question Answering					
Text Model (enc. only)	**Len**	**test-SQuAD** AOS	**test-SQuAD** FF1	**test-OOD** AOS	**test-OOD** FF1
longformer[11]	4096	49.1	55.9	-	-
longformer	4096	46.0	53.9	32.2	36.9
longformer	1024	40.0	46.0	22.8	26.4
ByT5-small	1024	48.4	54.9	27.1	31.0
ByT5lephone-small	1024	**53.3**	**61.1**	**32.3**	**37.3**

Experiments - Speech Translation

Text Model (dec. only)	**#Params**	BLEU
mT5-small	153M	26.8
mT5lephone-small	153M	27.2
ByT5-small	82M	27.9
ByT5lephone-small	82M	28.1



Conclusion

- We initially validate the use of byte-level models in cascaded SQA. Then, by extending this concept, we undertake a second-phase pretraining of PLMs on phonemicized text, enhancing performance further.

T5lephone: Bridging Speech and Text Self-supervised Models for Spoken Language Understanding via Phoneme level T5

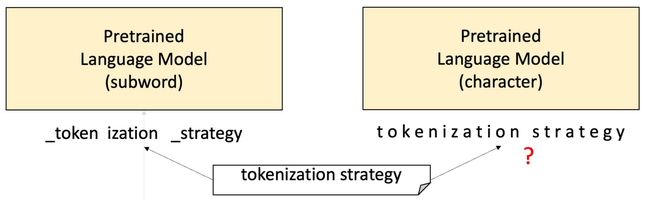
Chan-Jan Hsu*, Ho-Lam Chung* Hung-Yi Lee, Yu Tsao

Goal

We studied the impact of input granularity in self-supervised PLMs on SQA/ST performance, comparing T5/mT5 and ByT5 on datasets like NMSQA and Covost2. We developed T5lephone, a T5 variant using phonemicized text input, pretrained using Wikipedia text, with an initialization from mT5/ByT5. We innovatively re-represented the phonemicized text for maximum knowledge transfer from original ByT5 pretraining to phoneme pretraining. This led to state-of-the-art results and a 12% gain on NMSQA over previous models, with fewer parameters. T5lephone also outperformed previous methods on end-to-end NMSQA and ST.

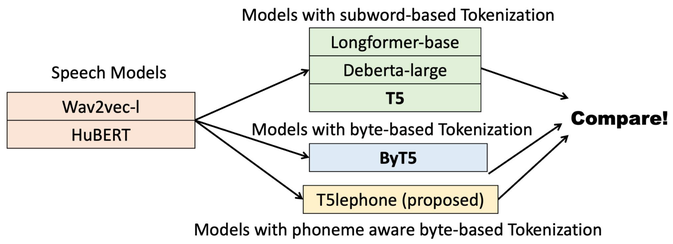
Background and Motivation

In Spoken Language Understanding (SLU), combining pretrained speech models (like HuBERT) with Pretrained Language Models (PLMs, e.g., T5) is common. Despite most studies utilizing subword-based tokenization in PLMs, the alignment between speech model outputs and language model inputs can be affected by input unit granularity. The exploration of character-based tokenization in PLMs remains insufficient.



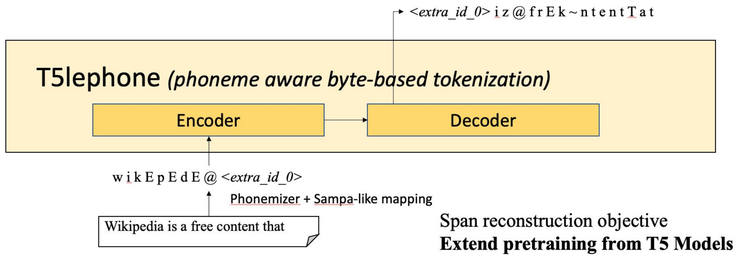
Most systems use PLM with subword-based tokenization, PLM with character-based tokenization is underexplored

Method - How does input granularity affect the performance?



Benchmark SQA Dataset (NMSQA) and ST Dataest (Covost2) and subword-level models and character-level models

T5lephone - PLM with phonemcized inputs



To create a model that takes in phoneme sequence as input, we conduct second-stage pretraining to the variants of T5. That is, we use mT5/ByT5 as initialization and train the model with the original span reconstruction objective using phonemicized inputs from the wiki text corpus. The resulting model is named T5lephone.

Experiments - Cascaded SQA

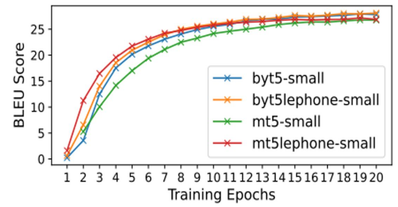
	PLM	#Params	text dev EM	text dev F1	dev AOS	dev FF1	test-SQuAD AOS	test-SQuAD FF1	test-OOD AOS	test-OOD FF1
subword	longformer-base	148M	85.0	91.9	44.8	55.3	50.7	63.4	38.4	47.3
	deberta-large	405M	87.9	93.9	38.3	49.1	47.5	60.3	28.7	37.6
	T5-small	61M	78.9	86.1	44.7	51.3	44.6	51.5	29.0	35.1
	T5-base	222M	83.0	89.9	52.8	60.4	58.2	66.7	37.6	43.7
	T5-large	770M	84.2	91.8	57.6	62.8	58.8	65.5	48.4	52.4
character / phoneme / character	ByT5-small	299M	78.4	83.9	55.4	60.6	62.8	67.7	41.3	45.4
	ByT5lephone-small	299M	76.7	83.7	55.0	60.8	**70.1**	**76.3**	**48.6**	**53.2**
	ByT5-base	581M	80.6	87.0	**59.7**	**65.3**	65.2	69.7	**48.4**	**53.3**

Experiments - End-to-end SQA

End-to-end Spoken Question Answering					
Text Model (enc. only)	**Len**	**test-SQuAD** AOS	**test-SQuAD** FF1	**test-OOD** AOS	**test-OOD** FF1
longformer[11]	4096	49.1	55.9	-	-
longformer	4096	46.0	53.9	32.2	36.9
longformer	1024	40.0	46.0	22.8	26.4
ByT5-small	1024	48.4	54.9	27.1	31.0
ByT5lephone-small	1024	**53.3**	**61.1**	**32.3**	**37.3**

Experiments - Speech Translation

Text Model (dec. only)	**#Params**	BLEU
mT5-small	153M	26.8
mT5lephone-small	153M	27.2
ByT5-small	82M	27.9
ByT5lephone-small	82M	28.1



Conclusion

- We initially validate the use of byte-level models in cascaded SQA. Then, by extending this concept, we undertake a second-phase pretraining of PLMs on phonemicized text, enhancing performance further.

T5lephone: Bridging Speech and Text Self-supervised Models for Spoken Language Understanding via Phoneme level T5

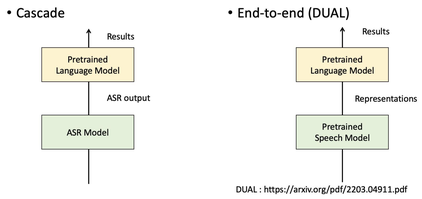
Chan-Jan Hsu*, Ho-Lam Chung* Hung-Yi Lee, Yu Tsao

Goal

- In this work, we conduct an extensive study on how self-supervised PLMs with different input granularity affect SQA/ST performance, by inferring on datasets such as NMSQA and Covost2. In particular, we compared T5/mT5 with ByT5, which has similar pretraining settings.
- We then further extend the idea to create T5lephone, a variant of T5 that takes phonemicized text as input. T5lephone is realized by self-supervised second-phase pretraining using phonemicized text from Wikipedia, with the model being initialized from mT5/ByT5.
- We devised a novel way to re-represent the phonemicized text to maximize the transferable knowledge from original text pretraining of ByT5 to our phoneme pretraining. We reached state-of-the-art and +12% performance gain on previous cascaded NMSQA results while using fewer parameters. The performance of our T5lephone model also exceeds previous methods on end-to-end NMSQA and ST.

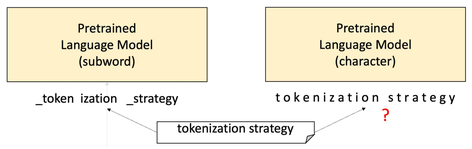
Background and Motivation

In Spoken language understanding (SLU), a natural solution is concatenating pre-trained speech models (e.g. HuBERT) and pretrained language models (PLM, e.g. T5). Most previous works use pre- trained language models with subword-based tokenization. How- ever, the granularity of input units affects the alignment of speech model outputs and language model inputs, and PLM with character- based tokenization is underexplored.



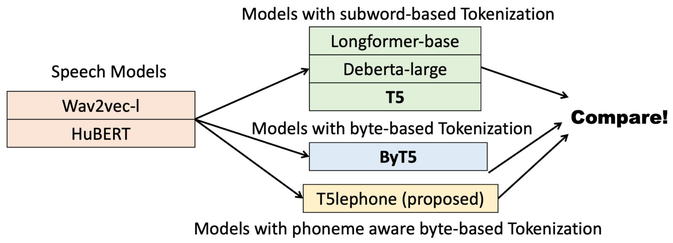
Speech language understanding is best solved incorporating a pretrained language model (PLM).

- Speech Translation : no-PLM (SUPERB-SG) < with PLM (Li, et al.)
- Spoken Question Answering : with PLM only



Most systems use PLM with subword-based tokenization, PLM with character-based tokenization is underexplored

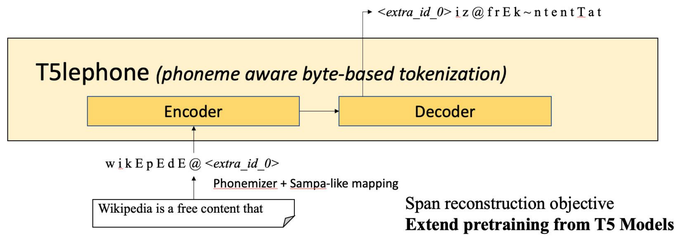
Method - How does input granularity affect the performance?



Benchmark SQA Dataset (NMSQA) and ST Dataest (Covost2) and subword-level models and character-level models

All of our experimented downstream tasks require a speech model followed by a language model. The speech model is responsible for extracting speech information, which is either vector representations or ASR outputs. The information is then forwarded into the language model for task-specific training.

T5lephone - PLM with phonemcized inputs



To create a model that takes in phoneme sequence as input, we conduct second-stage pretraining to the variants of T5. That is, we use mT5/ByT5 as initialization and train the model with the original span reconstruction objective using phonemicized inputs from the wiki text corpus. The resulting model is named T5lephone.

Experiments - Cascaded SQA

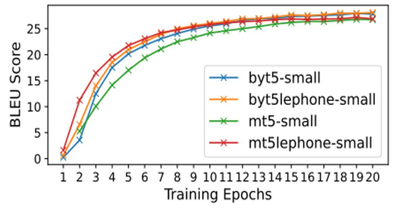
	PLM	#Params	text dev EM	text dev F1	dev AOS	dev FF1	test-SQuAD AOS	test-SQuAD FF1	test-OOD AOS	test-OOD FF1
subword ▲	longformer-base	148M	85.0	91.9	44.8	55.3	50.7	63.4	38.4	47.3
	deberta-large	405M	87.9	93.9	38.3	49.1	47.5	60.3	28.7	37.6
	T5-small	61M	78.9	86.1	44.7	51.3	44.6	51.5	29.0	35.1
	T5-base	222M	83.0	89.9	52.8	60.4	58.2	66.7	37.6	43.7
	T5-large	770M	84.2	91.8	57.6	62.8	58.8	65.5	48.4	52.4
character	ByT5-small	299M	78.4	83.9	55.4	60.6	62.8	67.7	41.3	45.4
phoneme	**ByT5lephone-small**	299M	76.7	83.7	55.0	60.8	**70.1**	**76.3**	**48.6**	**53.2**
character ▼	ByT5-base	581M	80.6	87.0	**59.7**	**65.3**	65.2	69.7	**48.4**	**53.3**

Experiments - End-to-end SQA

End-to-end Spoken Question Answering					
Text Model (enc. only)	Len	test-SQuAD AOS	test-SQuAD FF1	test-OOD AOS	test-OOD FF1
longformer[11]	4096	49.1	55.9	-	-
longformer	4096	46.0	53.9	32.2	36.9
longformer	1024	40.0	46.0	22.8	26.4
ByT5-small	1024	48.4	54.9	27.1	31.0
ByT5lephone-small	1024	**53.3**	**61.1**	**32.3**	**37.3**

Experiments - Speech Translation

Text Model (dec. only)	#Params	BLEU
mT5-small	153M	26.8
mT5lephone-small	153M	27.2
ByT5-small	82M	27.9
ByT5lephone-small	82M	28.1



Conclusion

- In this work, we first justify the use of byte-level models in cascaded SQA.
- We then extended this idea and conducted second-phase pretraining of pretrained language models on phonemicized text, which further improves performance.