Samsung Research

INTERSPEECH 2023

Online Continual Learning in Keyword Spotting for Low-Resource Devices via Pooling High-Order Temporal Statistics

Umberto Michieli, Pablo Peso Parada, Mete Ozay

Samsung Research UK

Samsung Research

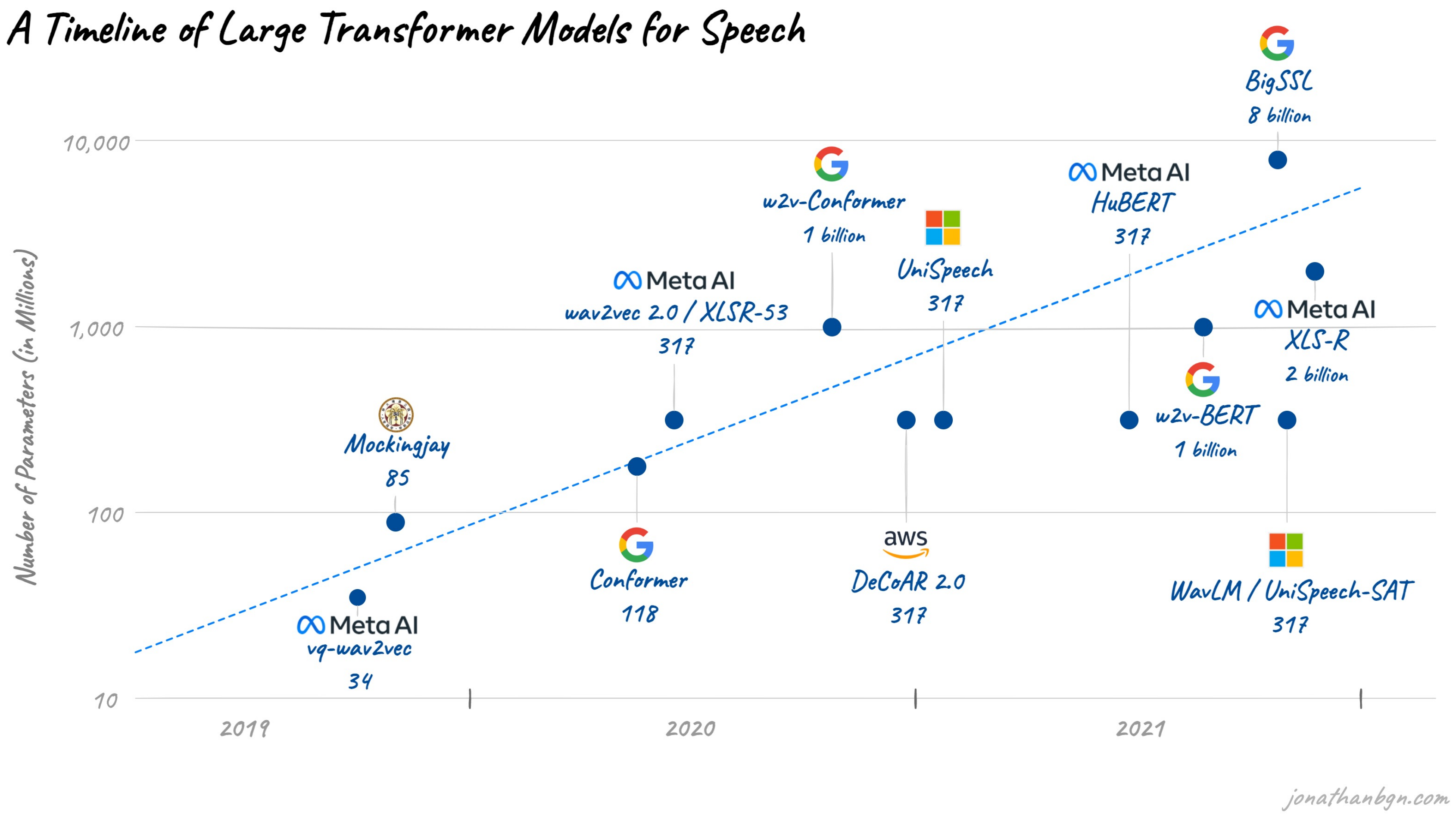
SUMMARY

1) Motivation
2) Setup
3) Our Method: TAP-SLDA
4) Main Results
5) Conclusion

SUMMARY

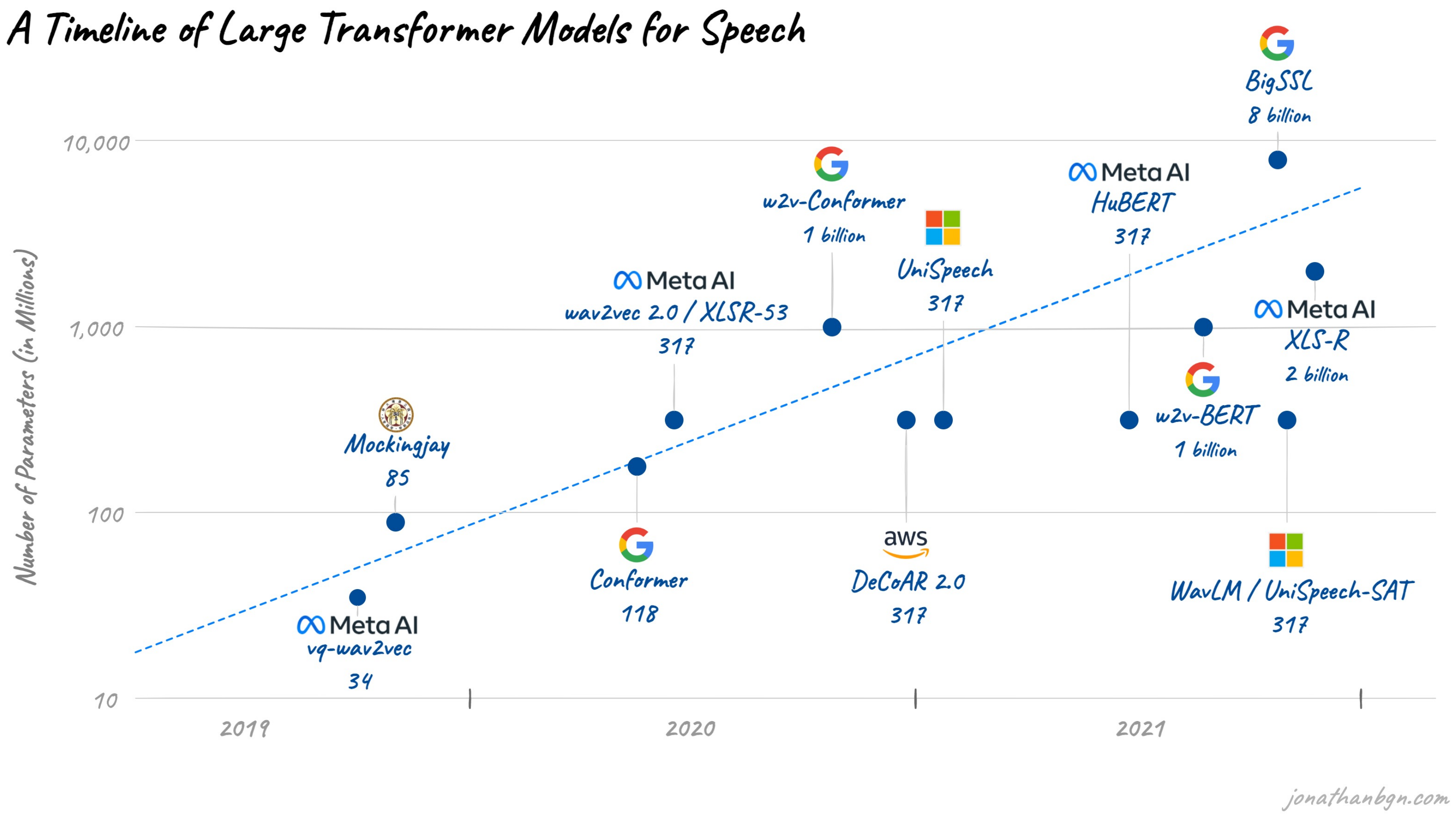
1) Motivation

Speech models are getting more powerful but also much larger!



Source: https://jonathanbgn.com/2021/12/31/timeline-transformers-speech.html

Speech models are getting more powerful but also much larger!



Cannot fine-tune these models on low-resource devices

Source: https://jonathanbgn.com/2021/12/31/timeline-transformers-speech.html

1) Motivation

Users of smart devices want on-device personalization

Example use-cases:

- Add custom commands to virtual assistants
- Add custom wake-up words to virtual assistants

Without sharing any data with the server

→ *Need for efficient and on-device personalization of keywords spotting (KWS) models*

SUMMARY

2) Setup

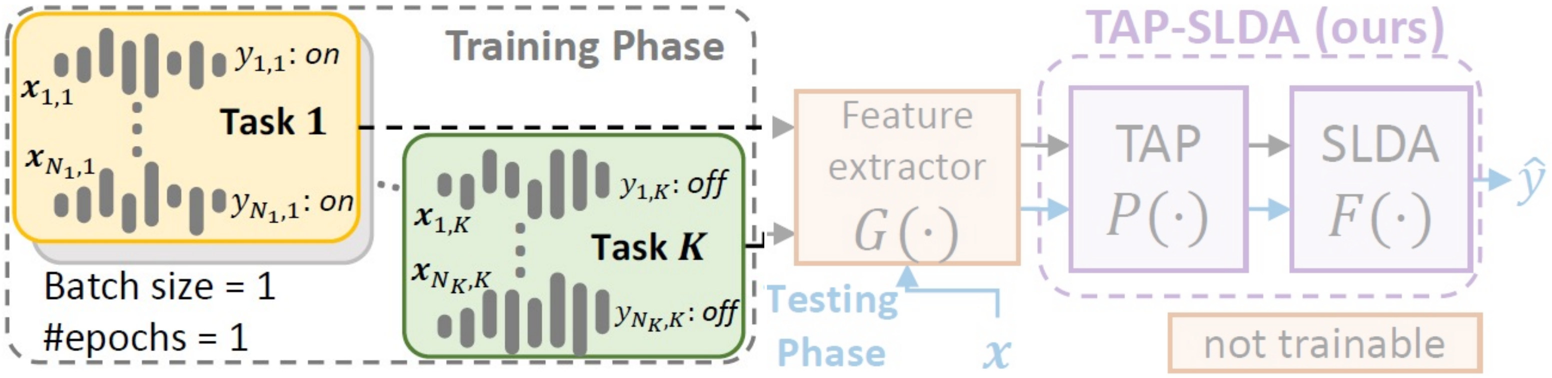
APPLICATION: Personalized Keywords Spotting

TASK: Class-Incremental **O**nline **C**ontinual **L**earning for **E**mbedded devices (EOCL)

DESIDERATA:

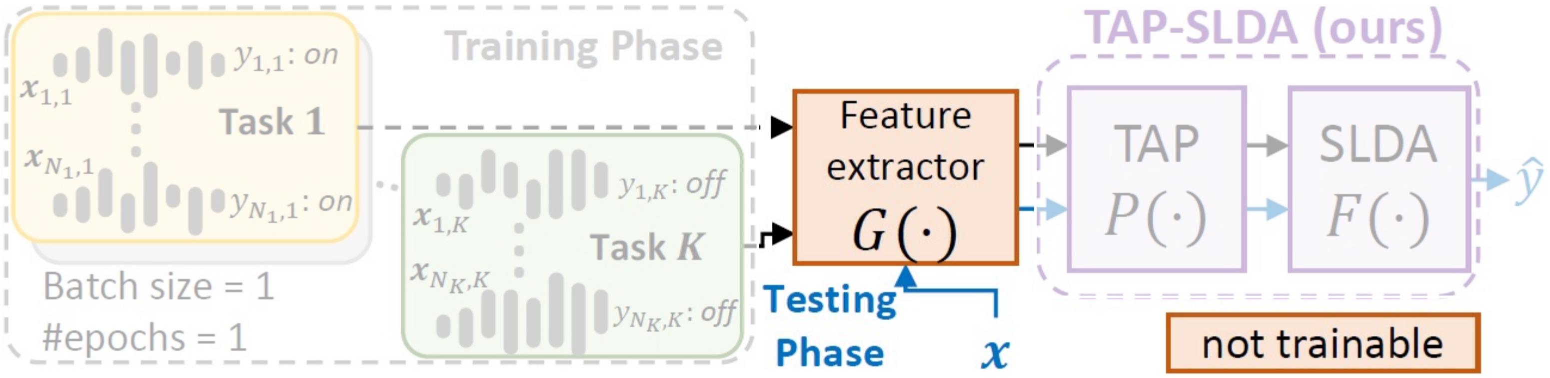
- Learn new concepts without forgetting old ones → Continual Learning
- Learn from a stream of data (samples are not stored on device) → Online
- Efficient update: targeting limited-resource devices → Embedded
  - Frozen backbone
  - Update via small batch size, as data is collected by the user
  - Limited number of training parameters

SUMMARY



Training Phase
$y_{1,1}$: on
$x_{1,1}$
Task 1
$x_{N_1,1}$
$y_{N_1,1}$: on
Batch size = 1
#epochs = 1
$y_{1,K}$: off
$x_{1,K}$
Task K
$x_{N_K,K}$
$y_{N_K,K}$: off
Feature extractor $G(\cdot)$
TAP-SLDA (ours)
TAP $P(\cdot)$
SLDA $F(\cdot)$
$\hat{y}$
Testing Phase
x
not trainable

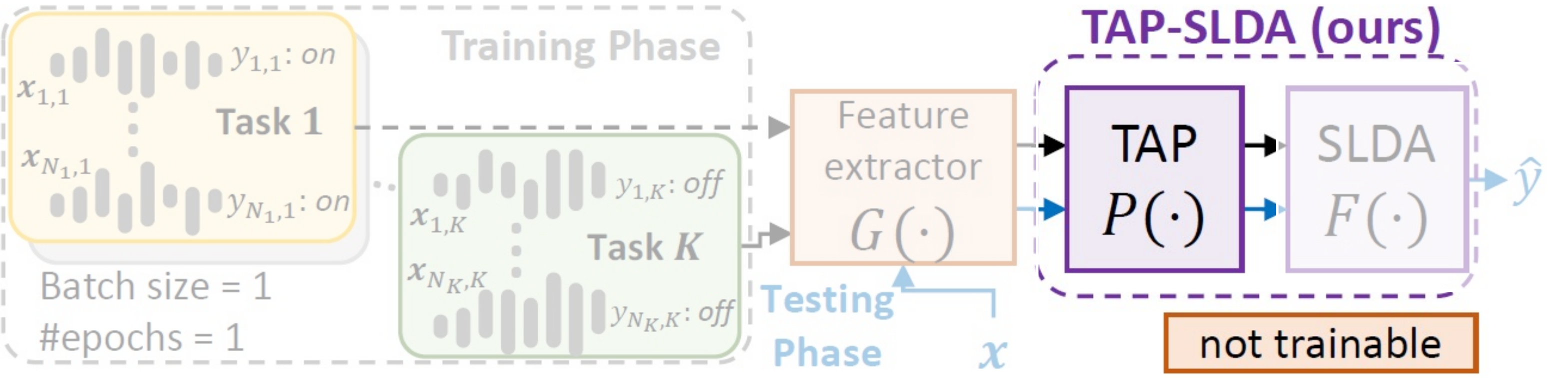
3) Our Method



Three main components:

1. **Feature Extractor**
   → pre-trained on server on public data and frozen

3) Our Method



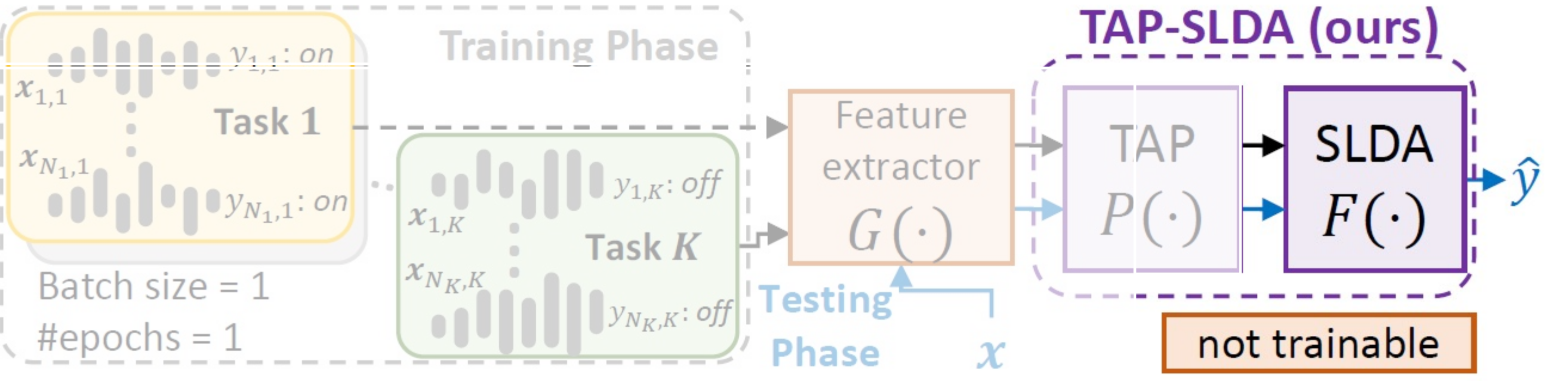
Three main components:

1. **Feature Extractor**
   → pre-trained on server on public data and frozen

2. **Temporal-Aware Pooling (TAP) →** concatenation of first R statistical moments (e.g., R=5)
   - Richer feature space to extract cues from a single-epoch training
   - Increased accuracy
   - Increased robustness

$$P(g) = \Big|\Big|\left(\mu, E_{\mathcal{G}}[(g-\mu)^2]^{\frac{1}{2}}, \Big|\Big|_{r=3}^{R} E_{\mathcal{G}}\left[\frac{g-\mu}{E_{\mathcal{G}}\,[(g-\mu)^2]^{\frac{1}{2}}}\right]^r\right)$$

3) Our Method



Three main components:

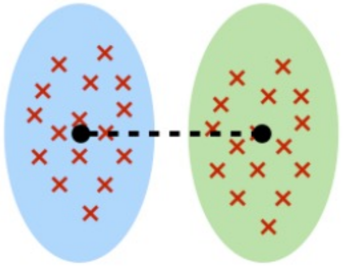
1. **Feature Extractor**
   → pre-trained on server on public data and frozen

2. **Temporal-Aware Pooling (TAP)** → concatenation of first R statistical moments (e.g., R=5)
   - Richer feature space to extract cues from a single-epoch training
   - Increased accuracy
   - Increased robustness

$$P(g) = \Big\Vert \left(\mu, E_{\mathcal{G}}[(g-\mu)^2]^{\frac{1}{2}}, \Big\Vert_{r=3}^{R} E_{\mathcal{G}}\left[\frac{g-\mu}{E_{\mathcal{G}}[(g-\mu)^2]^{\frac{1}{2}}}\right]^r \right)$$

3. **Classifier** → lightweight online continual learning method
   → We use SLDA [1] on the enriched feature space

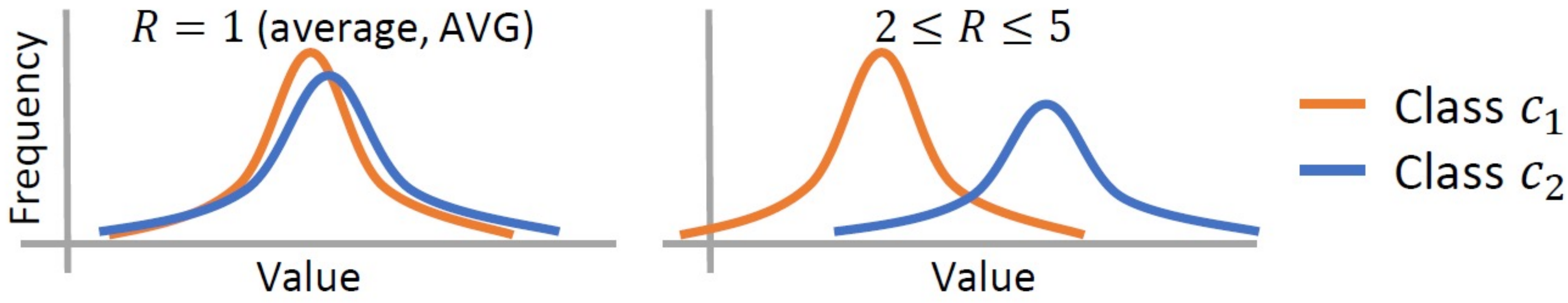
   SLDA estimates a Gaussian model for each class over the feature space with a class-wise mean (prototype) and shared-across-classes variance
   - Online estimate of covariance
   - Shared covariance across classes



[1] Hayes, Tyler L., and Christopher Kanan. "Lifelong machine learning with deep streaming linear discriminant analysis." CVPRW 2020

3) Our Method: Intuition

We plot the distribution of statistical moments of features extracted from class1 or class2 :



→ Features of different classes have similar distribution of 1st moments, while…

→ … higher moments capture the difference

Samsung Research

SUMMARY

1) Motivation
2) Setup
3) Our Method: TAP-SLDA
4) **Main Results**
5) Conclusion

4) Results: Preliminaries

Datasets

- **GSC-V2:** 35 English words
- **MSWC:** we picked the 5 most represented languages (*en*, *de*, *fr*, *ca*, *rw*) and create 3 micro-sets with different number of keywords *N={25, 50, 100}*. Splits are available at [2]

Metrics

- Acc: assesses the final model performance on all classes
- CL metrics: BwT (↑), Forgetting (↓), Plasticity (↑),
- Relative comparison: $(Acc_2 - Acc_1)/(100 - Acc_1)$

[2] MSWC splits: https://github.com/umbertomichieli/tap-slda

4) Results: GSC-V2

TAP-SLDA (ours) outperforms competing approaches on 6 architectures:

	W2V-B				W2V-L				Emf-B				HB-B				HB-L				HB-XL			
	Acc	BwT	Forg	Pla	Acc	BwT	Forg	Pla	Acc	BwT	Forg	Pla	Acc	BwT	Forg	Pla	Acc	BwT	Forg	Pla	Acc	BwT	Forg	Pla
FT	2.5$_{\pm1.0}$	1.2	98.5	**100**	2.5$_{\pm1.0}$	1.3	98.5	**100**	2.7$_{\pm0.8}$	1.5	97.6	**99.8**	2.5$_{\pm1.0}$	1.2	98.5	**100**	2.5$_{\pm1.0}$	1.2	98.5	**100**	2.5$_{\pm1.0}$	1.2	98.5	**100**
PRCP [10]	2.5$_{\pm1.0}$	1.2	98.4	**100**	2.5$_{\pm1.0}$	1.3	98.4	**100**	2.9$_{\pm0.7}$	2.2	95.8	97.5	2.5$_{\pm1.0}$	1.4	98.4	**100**	2.5$_{\pm1.0}$	1.2	98.4	**100**	2.6$_{\pm1.0}$	1.3	98.4	**100**
SNB [27]	3.7$_{\pm0.0}$	6.5	38.6	15.7	9.4$_{\pm2.4}$	9.2	36.1	16.6	7.2$_{\pm0.0}$	13.1	24.8	22.8	77.4$_{\pm0.0}$	80.9	18.1	84.2	6.5$_{\pm0.0}$	12.0	37.1	23.7	57.9$_{\pm0.0}$	60.3	23.8	65.2
SOvR [10]	1.8$_{\pm0.0}$	6.2	39.6	14.5	1.8$_{\pm0.0}$	6.2	31.6	15.8	4.9$_{\pm0.0}$	7.3	24.4	14.5	19.1$_{\pm0.0}$	29.8	41.2	42.5	14.0$_{\pm0.0}$	21.8	34.8	32.9	15.7$_{\pm0.0}$	22.6	35.3	33.8
NCM [23]	67.5$_{\pm6.0}$	74.0	13.2	77.3	69.5$_{\pm7.0}$	76.0	12.1	80.0	8.8$_{\pm0.0}$	14.1	23.6	22.0	83.9$_{\pm0.0}$	86.0	8.2	89.1	46.7$_{\pm0.0}$	55.4	22.8	62.4	62.5$_{\pm0.0}$	66.7	17.3	72.2
SLDA [18]	82.4$_{\pm0.1}$	84.2	8.0	87.0	81.6$_{\pm0.1}$	83.8	8.3	86.8	23.2$_{\pm0.0}$	32.9	23.7	41.9	94.2$_{\pm0.0}$	94.9	3.8	96.2	85.5$_{\pm0.0}$	88.2	8.3	90.8	93.3$_{\pm0.0}$	94.1	5.1	95.4
SQDA [26]	80.6$_{\pm2.5}$	78.2	5.7	81.2	80.5$_{\pm2.4}$	76.9	**5.1**	80.6	24.3$_{\pm0.7}$	21.9	**17.3**	31.5	90.0$_{\pm3.4}$	87.8	**2.8**	90.0	67.4$_{\pm4.4}$	59.4	**4.4**	64.8	83.0$_{\pm2.1}$	73.9	**0.0**	76.8
TAP-SLDA (ours)	**89.9**$_{\pm0.0}$	**91.8**	**5.6**	93.7	**90.0**$_{\pm0.0}$	**91.7**	5.4	93.4	**50.8**$_{\pm0.3}$	**58.8**	20.3	65.8	**95.7**$_{\pm0.0}$	**96.0**	3.0	96.9	**90.8**$_{\pm0.0}$	**91.8**	6.1	93.6	**95.5**$_{\pm0.0}$	**95.8**	3.4	96.6
iCaRL [6]	76.9$_{\pm1.0}$	79.1	14.7	83.6	73.6$_{\pm1.8}$	78.0	17.4	83.5	18.2$_{\pm0.3}$	26.9	28.8	44.7	93.7$_{\pm0.1}$	94.6	4.2	96.7	78.5$_{\pm0.5}$	83.1	12.8	85.1	92.9$_{\pm0.3}$	93.8	5.3	95.5
Avg	47.5	49.6	33.5	73.0	48.1	50.1	32.5	73.7	15.1	19.3	38.0	46.2	64.3	65.9	28.6	88.5	44.1	46.9	34.6	71.6	56.8	57.7	30.4	80.8

→ The temporal-aware enriched feature space provides useful temporal characteristics to the Gaussian modelling

4) Results: GSC-V2

TAP improves every OCL method:

TAP+	W2V-B	W2V-L	Emf-B	HB-B	HB-L	HB-XL	Avg
FT	5.4	6.1	2.9	2.7	2.7	2.7	3.8 (+1.2)
PRCP	3.5	4.6	3.0	2.8	2.8	2.9	3.3 (+0.7)
SNB	3.9	7.1	9.3	84.1	6.9	59.9	28.5 (+2.3)
SOvR	51.3	60.9	5.8	54.3	14.9	49.6	39.5 (+29.9)
NCM	78.2	79.8	12.1	87.2	44.5	84.9	64.5 (+8.0)
CBCL	75.9	77.3	12.0	88.7	48.0	86.1	64.7 (+8.2)
SLDA	**89.9**	**90.0**	**50.8**	**95.7**	**90.8**	**95.5**	**85.5** (+8.8)
SQDA	85.5	84.0	48.7	88.8	67.1	82.7	76.1 (+5.1)
iCaRL	82.9	85.7	31.0	90.9	76.9	90.8	76.4 (+4.1)
Avg	52.9	55.1	19.5	66.1	39.4	61.7	

4) Results: GSC-V2

TAP improves every OCL method:

TAP+	W2V-B	W2V-L	Emf-B	HB-B	HB-L	HB-XL	Avg
FT	5.4	6.1	2.9	2.7	2.7	2.7	3.8 (+1.2)
PRCP	3.5	4.6	3.0	2.8	2.8	2.9	3.3 (+0.7)
SNB	3.9	7.1	9.3	84.1	6.9	59.9	28.5 (+2.3)
SOvR	51.3	60.9	5.8	54.3	14.9	49.6	39.5 (+29.9)
NCM	78.2	79.8	12.1	87.2	44.5	84.9	64.5 (+8.0)
CBCL	75.9	77.3	12.0	88.7	48.0	86.1	64.7 (+8.2)
SLDA	**89.9**	**90.0**	**50.8**	**95.7**	**90.8**	**95.5**	**85.5** (+8.8)
SQDA	85.5	84.0	48.7	88.8	67.1	82.7	76.1 (+5.1)
iCaRL	82.9	85.7	31.0	90.9	76.9	90.8	76.4 (+4.1)
Avg	52.9	55.1	19.5	66.1	39.4	61.7	

4) Results: GSC-V2

Samsung Research

TAP outperforms other pooling schemes:

	W2V-B	W2V-L	Emf-B	HB-B	HB-L	HB-XL	**Avg**
AVG [29]	82.4	81.6	23.2	94.2	85.5	93.3	76.7
MAX [30]	87.7	88.3	34.9	94.8	87.2	94.1	81.2
MIX (50%) [31]	87.8	87.3	31.1	94.7	87.3	94.0	80.4
STOCH [32]	80.9	77.6	24.6	85.3	64.5	77.9	68.4
L2 [36]	79.0	79.7	17.4	92.9	74.4	89.7	72.2
L3 [36]	79.3	81.2	15.9	92.4	70.4	89.1	71.4
RAP (10%) [33]	86.5	86.9	36.3	94.8	87.5	93.8	81.0
AVGMAX [34]	<u>89.1</u>	<u>89.6</u>	44.8	<u>95.2</u>	<u>89.1</u>	<u>94.7</u>	<u>83.8</u>
iSQRT-COV [37]	80.3	80.3	**55.1**	92.4	83.8	90.3	80.4
TSDP [35]	83.9	83.6	32.4	94.4	84.9	93.9	78.9
TSTP [35]	87.4	87.6	39.1	95.1	88.0	94.5	82.0
TAP (ours)	**90.0**	**90.0**	<u>50.8</u>	**95.7**	**90.8**	**95.5**	**85.5**
Avg	84.5	84.5	33.8	93.5	82.8	91.7	

Larger feature space is *not* all we need:

Δ_{fs}: increase of pooled feature size

	W2V-B	W2V-L	Emf-B	HB-B	HB-L	HB-XL	**Avg**	Δ_{fs}
AVG	82.4	81.6	23.2	94.2	85.5	93.3	76.7	1
MAX	87.7	88.3	34.9	94.8	87.2	94.1	81.2	1
RAP 5%	85.7	85.8	35.1	94.5	86.8	93.8	80.3	26.8
RAP 10%	86.5	86.9	36.3	94.8	87.5	93.8	81.0	53.5
RAP 20%	85.7	85.9	36.3	94.5	86.8	93.8	80.5	107
$MAXW_2$	87.7	88.3	34.9	94.8	87.2	94.1	81.2	5
$MAXW_5$	85.8	86.5	34.9	94.7	87.1	93.9	80.5	11
$MAXW_{10}$	85.6	86.1	35.3	94.7	86.7	93.7	80.3	21
FLAT	85.1	86.2	24.6	94.3	85.7	93.5	78.2	535
TAP (R=2)	87.4	87.6	39.1	95.1	88.0	94.5	82.0	2
TAP (R=3)	89.3	89.2	47.3	95.5	89.8	95.4	84.4	3
TAP (R=4)	90.0	90.0	49.7	95.6	90.4	**95.5**	85.2	4
TAP (R=5)	90.0	90.0	**50.8**	**95.7**	**90.8**	**95.5**	**85.5**	5
TAP (R=6)	**90.1**	**90.2**	47.8	**95.7**	89.9	**95.5**	84.9	6

4) Results: GSC-V2

TAP only adds minimal overhead:

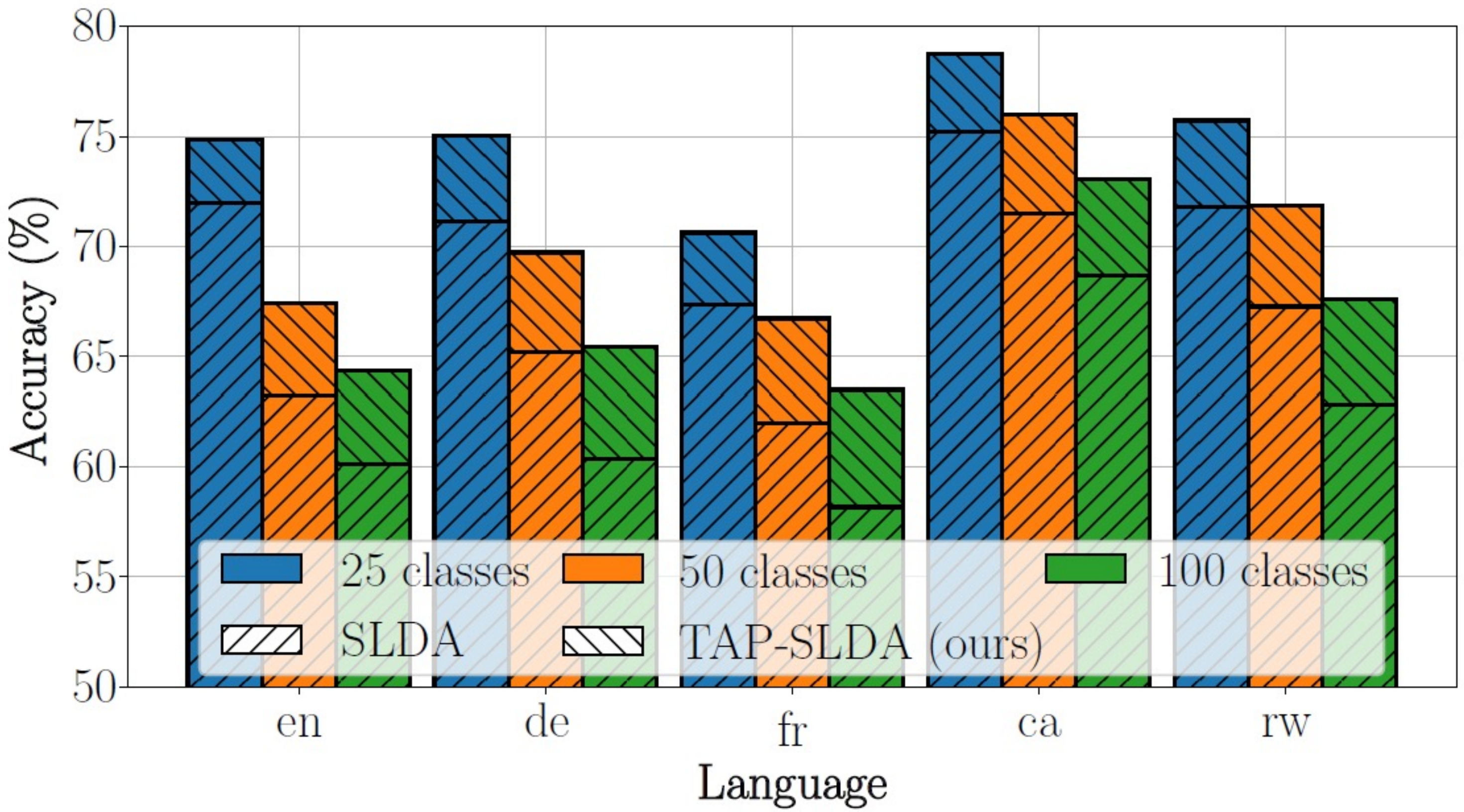
TAP+	$R=1$		$R=2$		$R=3$		$R=4$		$R=5$		$R=6$	
	Acc	Δ_p	Acc	Δ_p	Acc	Δ_p	Acc	Δ_p	Acc	Δ_p	Acc	Δ_p
FT	1.6	0.01	1.5	0.01	1.5	0.02	1.5	0.02	**3.8**	0.03	2.6	0.04
NCM	52.1	0.01	62.8	0.01	63.5	0.02	64.0	0.02	**64.5**	0.03	56.7	0.04
SLDA	76.7	0.10	82.0	0.10	84.4	0.11	85.2	0.12	**85.5**	0.12	84.9	0.13

Δ_p: increase of parameters percentage over the backbone

4) Results: MSWC

TAP enhances personalization to other languages:

HuBERT-Base model pre-trained on English data only and adapted to recognize keywords in different languages



SUMMARY

5) Conclusion

New task: online continual learning for KWS models targeting low-resource devices with limited computational and storage capability

New method: TAP-SLDA, a parameter-efficient online continual learning method

TAP-SLDA features:

- New **temporal-aware pooling** scheme based on the first 5 moments of extracted features
- **Lightweight** solution: frozen feature extractor + class-conditional Gaussian modelling of feature space
- Extraction of **high-order statistical** moments of the embedded features of input samples
- **Robust** results in a variety of scenarios on several backbones

Samsung Research

Thank you!