Google

Personal VAD: Speaker-Conditioned Voice Activity Detection

Authors: *Shaojin Ding, Quan Wang, Shuo-yiin Chang, Li Wan, Ignacio Lopez Moreno*

Presented by: *Quan Wang* <quanw@google.com>

おでっせい Odyssey 2020

Key messages

- What:
  - A system to detect the voice activity of a **target speaker**
- Why:
  - Reduces CPU, memory and battery consumption for **on-device speech recognition**
- How:
  - Frame-level streaming detection
  - **Speaker embedding** as side input

Part 1:

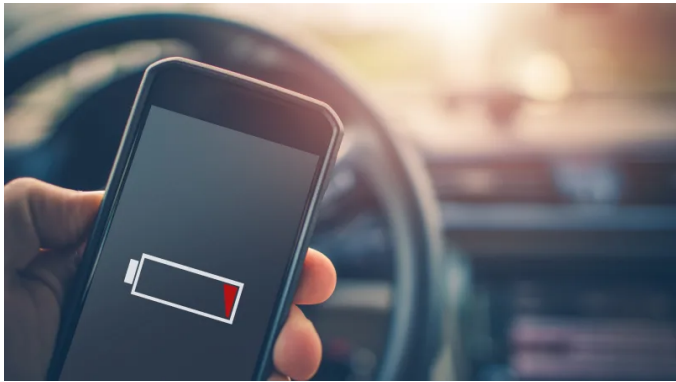
Background

On-device ASR

- Moving ASR from cloud to device is the trend
  - No requirement for **Internet**
  - Much less **latency** due to communications
  - Better security and **privacy** preservation
- Example use cases (smartphones, smart home devices):
  - “Turn on flashlight”
  - “Turn on bedroom lights”

Challenges for on-device ASR

- Limited CPU
- Limited memory
- Limited battery
- ASR is not alone – many other programs running on the device



On-device ASR: When to run it?

- On-device ASR can’t be always running
- Typical solution: Keyword detection (*a.k.a.* wakeword or hotword)
  - Keyword detection: Very cheap, always running
  - ASR: Expensive, only runs when keyword is detected
- Example:
  - “Hey Google, turn on flashlight”

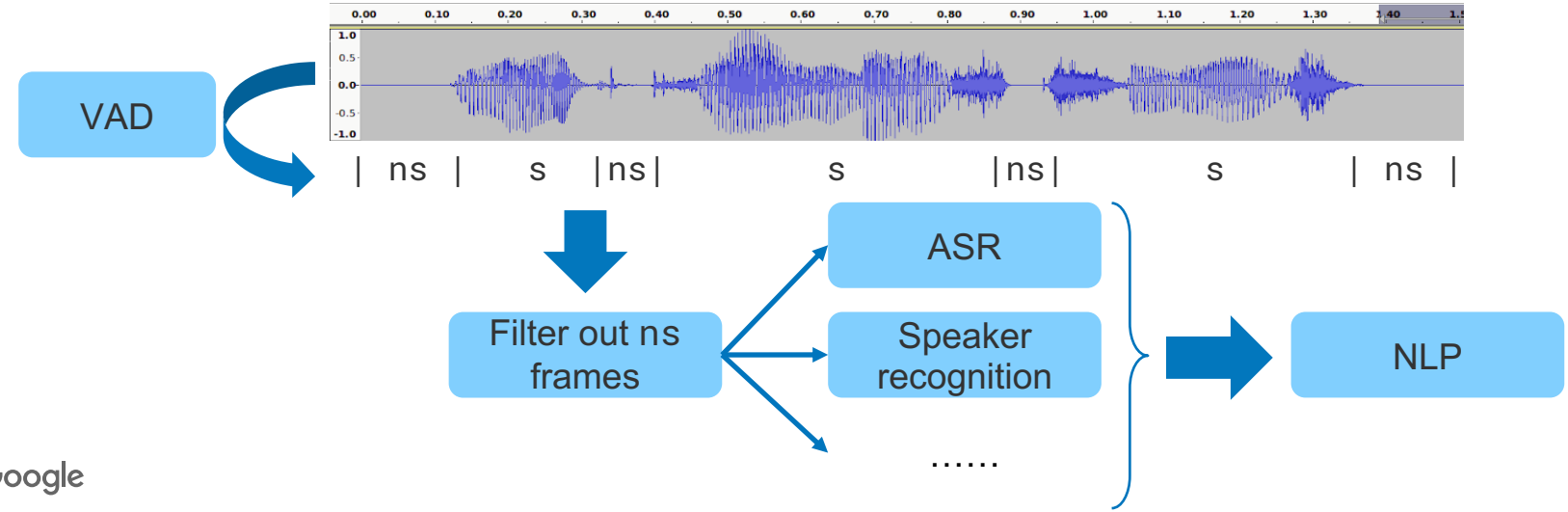
On-device ASR: When to run it?

- However, many people prefer keyword-less interactions
  - More seamless, more natural, more intelligent

- Alternative solution: Voice Activity Detection (VAD)
  - VAD: Very cheap, always running
  - ASR: Expensive, only runs when VAD triggers

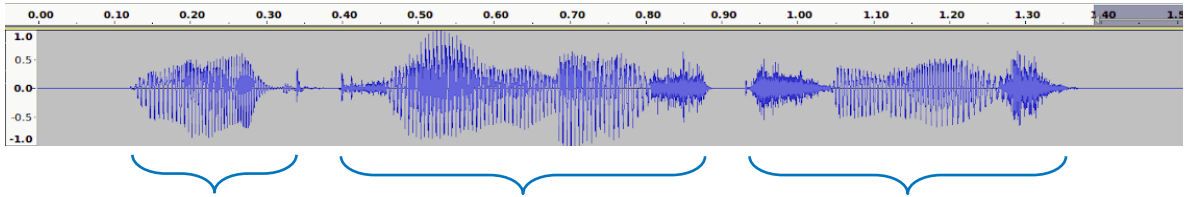
VAD: Rejects non-speech signals

- Each frame is categorized as non-speech (ns) or speech (s)
- Only run ASR on speech frames



VAD: Is it good enough?

- But what if:



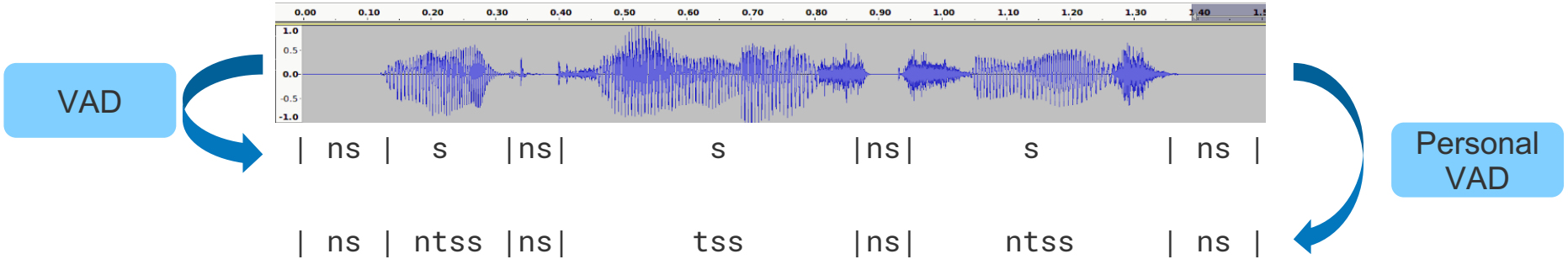
- These are all valid speech signals
- But ASR shouldn’t run on everything

Part 2:

Introducing Personal VAD

From VAD to personal VAD

- 3 categories instead of 2:
  - non-speech (ns), target speaker speech (tss), **non-target speaker speech** (ntss)



Benefits

- Only run ASR on target speaker speech (`tss`)
- Save lots of computational resources, *e.g.*:
  - When TV is on
  - When having multiple family members in the household
  - During social activities
- Key of success:
  - The personal VAD model needs to be **tiny and fast** (like keyword detector or standard VAD)
  - Very low FR, relative low FA

Personal VAD is NOT speaker recognition/diarization

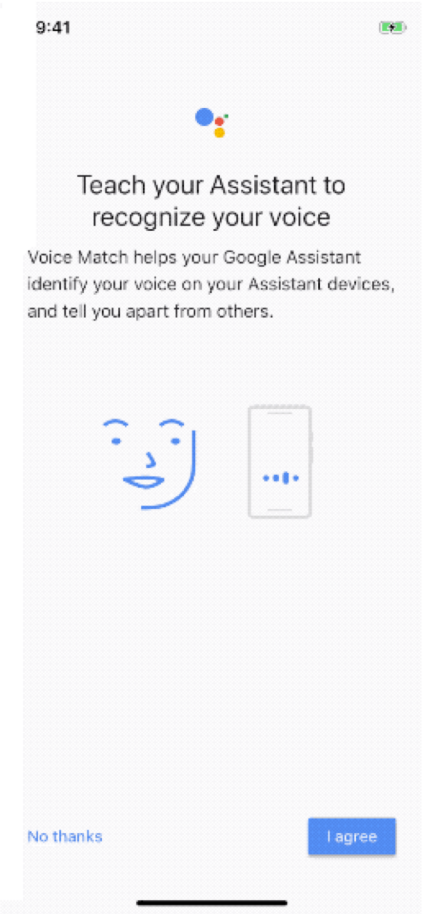
- Speaker recognition:
  - Typically utterance-level or window-level output
  - Models typically big (>5M parameters)
- Speaker diarization:
  - Needs to **cluster** unknown speakers
  - Number of speakers matters a lot
- Personal VAD:
  - Frame-level output
  - Only cares about target speaker; use non-target speaker to represent everyone else
  - Model must be tiny (<200K parameters), fast and streaming

Part 3:

Implementation

Whom to listen to?

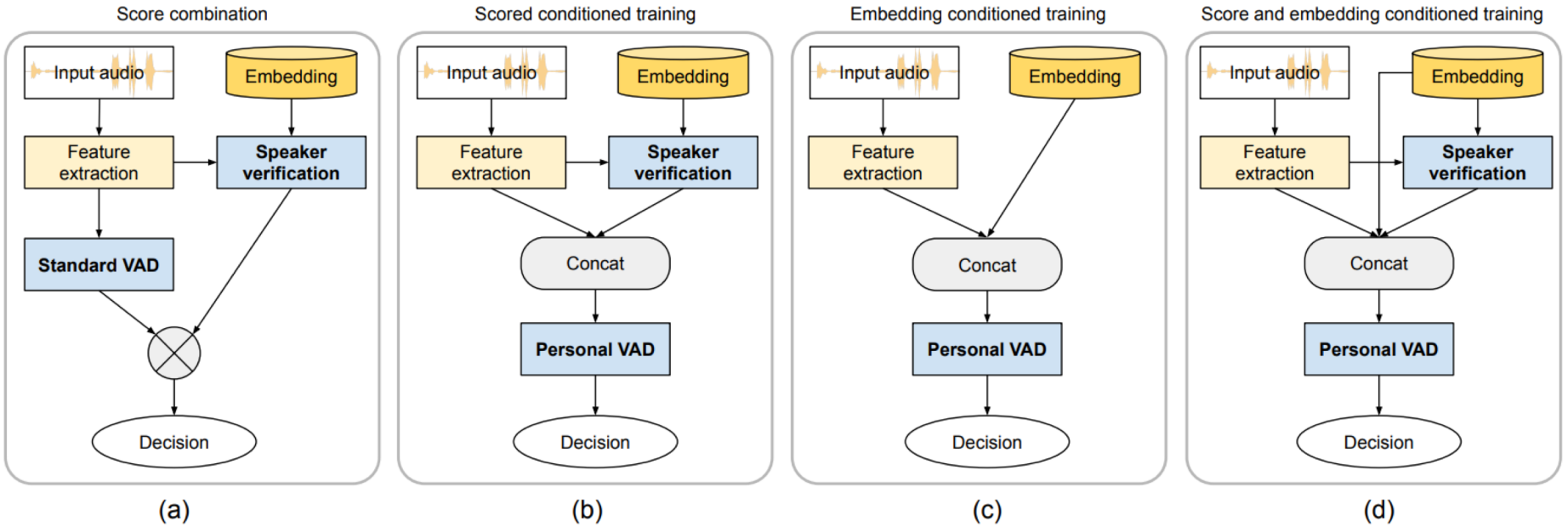
- Modern speech systems allow users to **enroll** their voice
- Enrollment is a **one-off** experience, thus the cost can be ignored at runtime
- After enrollment, **speaker embedding** (d-vector) will be stored on the device
- Usually used for speaker recognition or voice filtering



Implementation

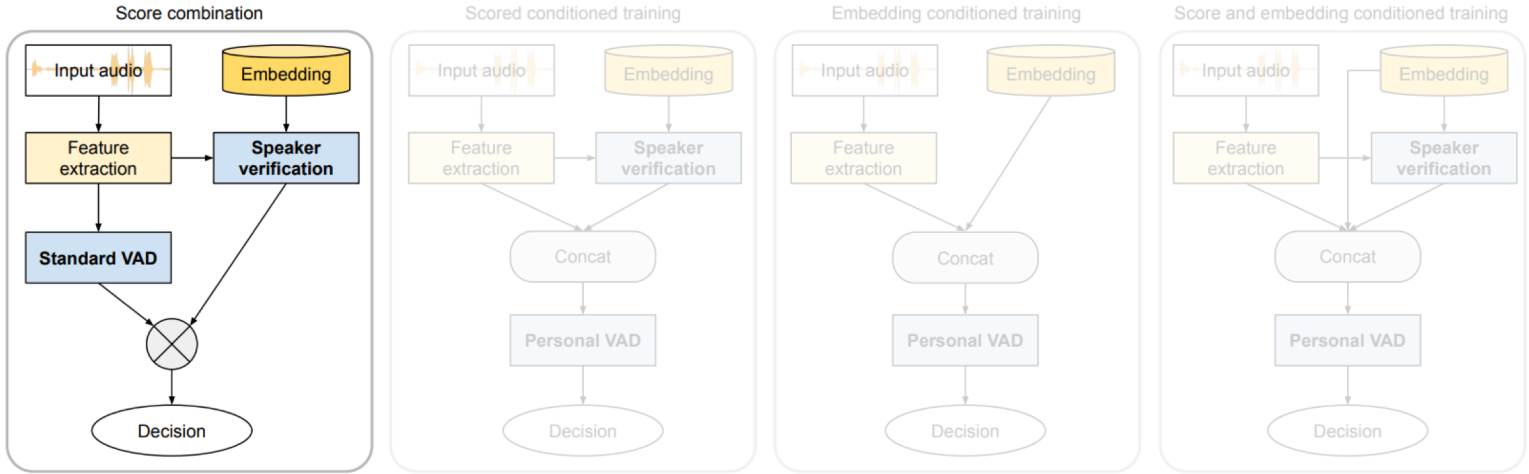
- Ways to implement personal VAD
  - [Baseline] Combine standard VAD and speaker verification
  - [Proposed] Train a new personal VAD model with:
    - Speaker verification score
    - Speaker embedding

Four architectures for personal VAD



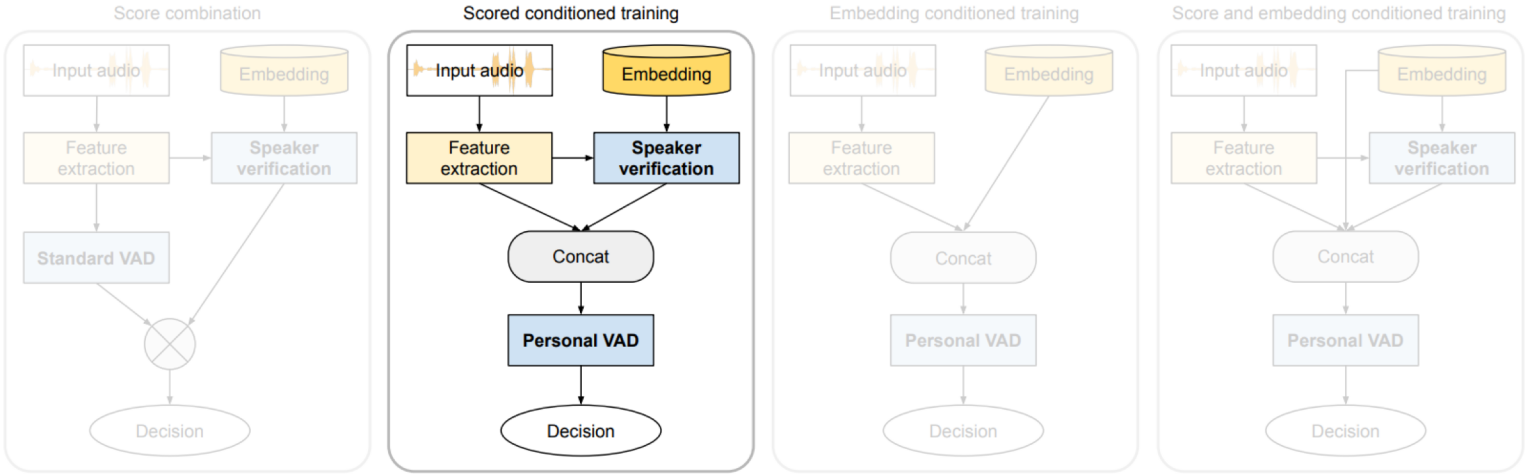
Score combination (SC)

- Use standard VAD and speaker verification (no training new model)
- We use it as a baseline
- Note: Running speaker verification is expensive



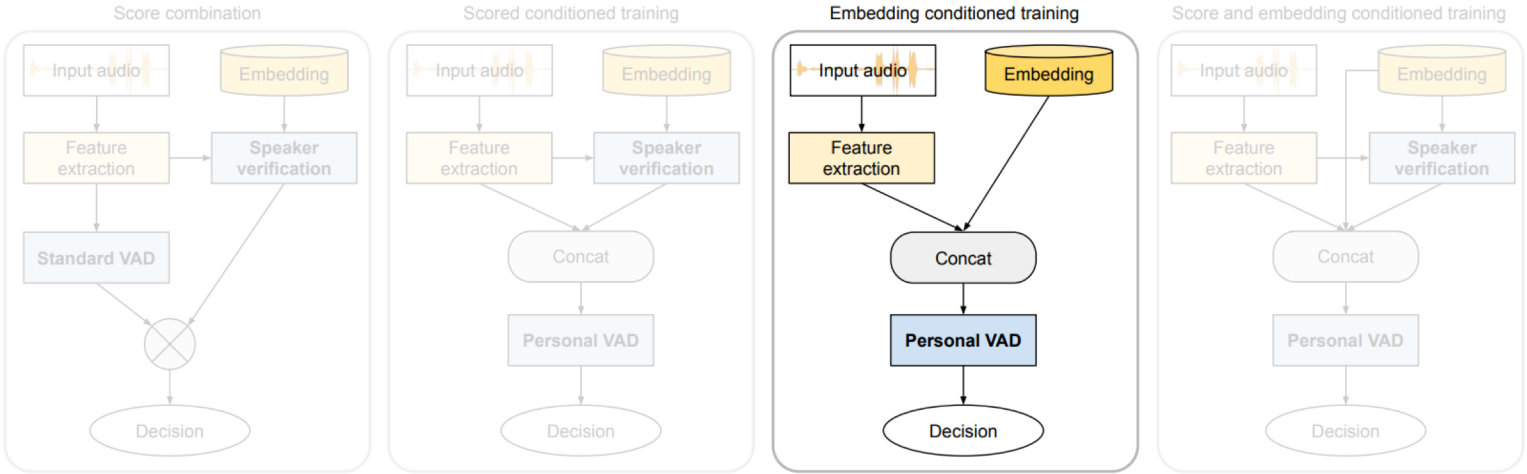
Score conditioned training (ST)

- Train a new model on feature + score
- Note: Running speaker verification is expensive



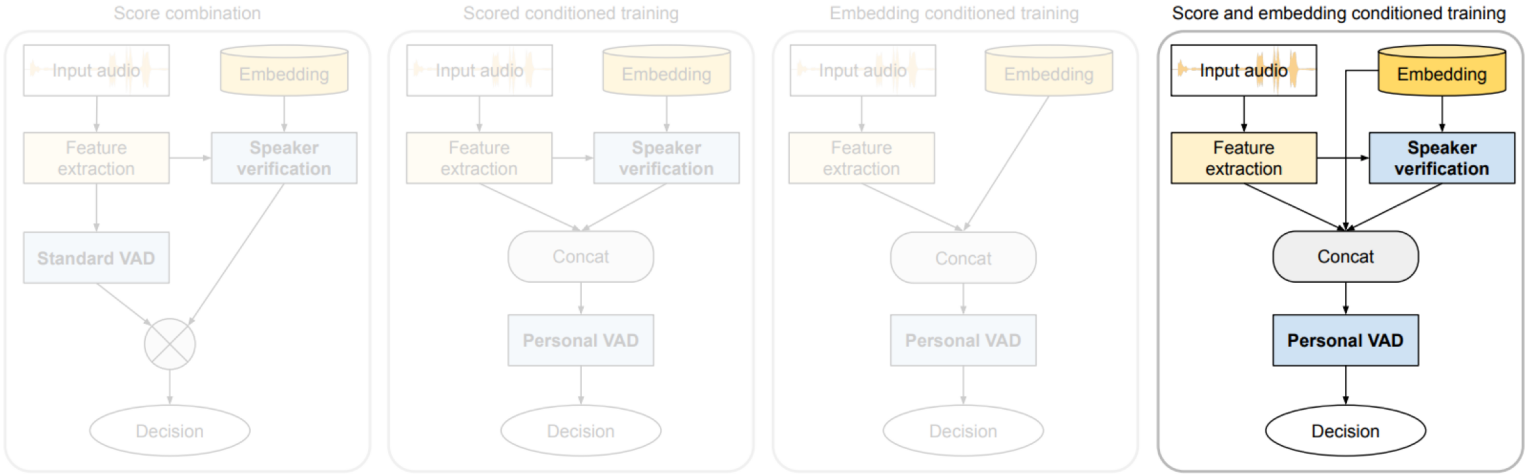
Embedding conditioned training (ET)

- Train a new model on feature + embedding
- Note: No need to run speaker verification at runtime; cheap and ideal for device



Score and embedding conditioned training (SET)

- Train a new model on feature + score + embedding
- Use most information from speaker recognition
- Note: Running speaker verification is expensive



Loss function

- Cross-entropy loss (CE)
  - Standard VAD usually uses **binary** CE loss
  - Personal VAD has 3 categories: can simply use **ternary** CE loss

$$L_{\mathrm{CE}}(y, \mathbf{z}) = -\log \frac{\exp(z_y)}{\sum_k \exp(z_k)}$$

- y is the index of ground truth label
- $\mathbf{z}$ is the unnormalized predicted probabilities
- z_k is the kth element of $\mathbf{z}$

Can we do better than cross-entropy?

- For personal VAD, both **non-speech** (`ns`) and **non-target speaker speech** (`ntss`) will be discarded by downstream components
- ASR only triggers on **target speaker speech** (`tss`)
- Thus the costs for different confusion errors are different

Weighted pairwise loss

- We propose **weighted pairwise loss** (WPL)
- Set different weights between different pairs of classes

$$L_{\mathrm{WPL}}(y, \mathbf{z}) = -\mathbb{E}_{k \neq y}\left[w_{<k,y>} \cdot \log \frac{\exp(z_y)}{\exp(z_y) + \exp(z_k)}\right]$$

- y is the index of ground truth label
- $\mathbf{z}$ is the unnormalized predicted probabilities
- z_k is the kth element of $\mathbf{z}$
- $w_{<k,y>}$ is the weight between class k and class y

For example, $w_{<tss,ns>} = w_{<tss,ntss>} = 1, w_{<ntss,ns>} = 0.1$

Part 4:

Experiment Setup

Ideal dataset

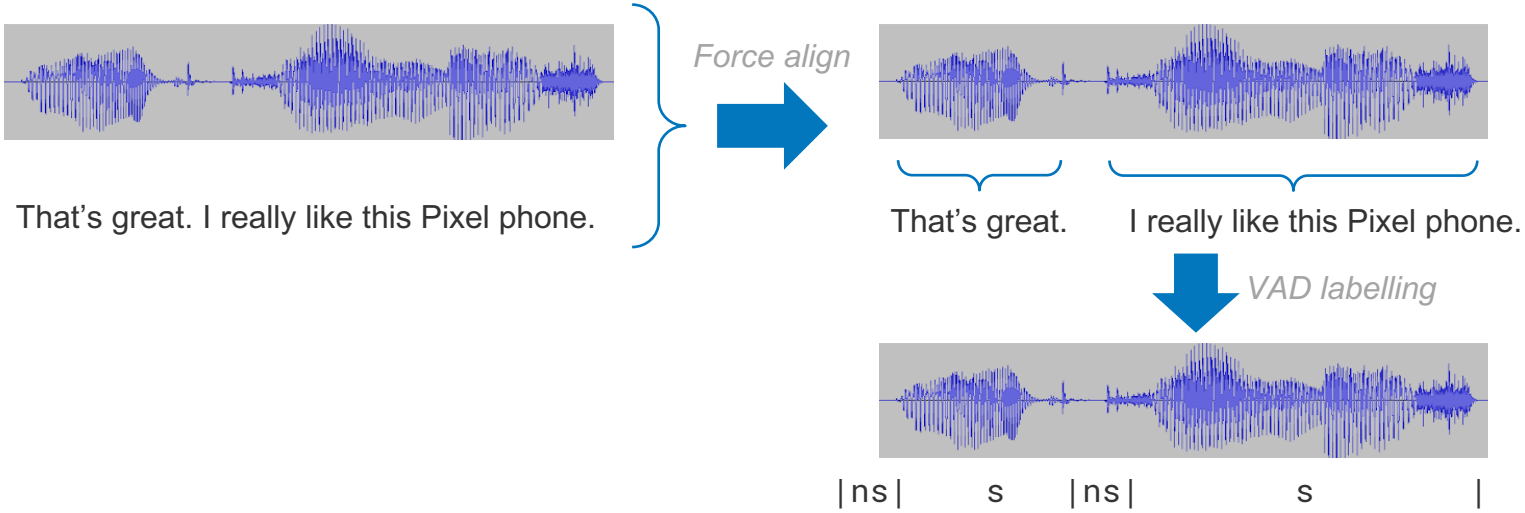
- An “ideal” dataset for personal VAD should have:
  - Realistic and natural speaker turns
  - Diverse noise conditions
  - Frame-level speaker labels
  - Enrollment utterances for each speaker
- Unfortunately, we cannot find a good candidate dataset

Dataset

- As a proof-of-concept, we make artificial “conversational” speech
- Our experiments are based on LibriSpeech:
  - 16kHz read English speech
  - Derived from read audiobooks
  - 960 hours of clean and noisy speech for training
  - Clean and noisy speech for testing

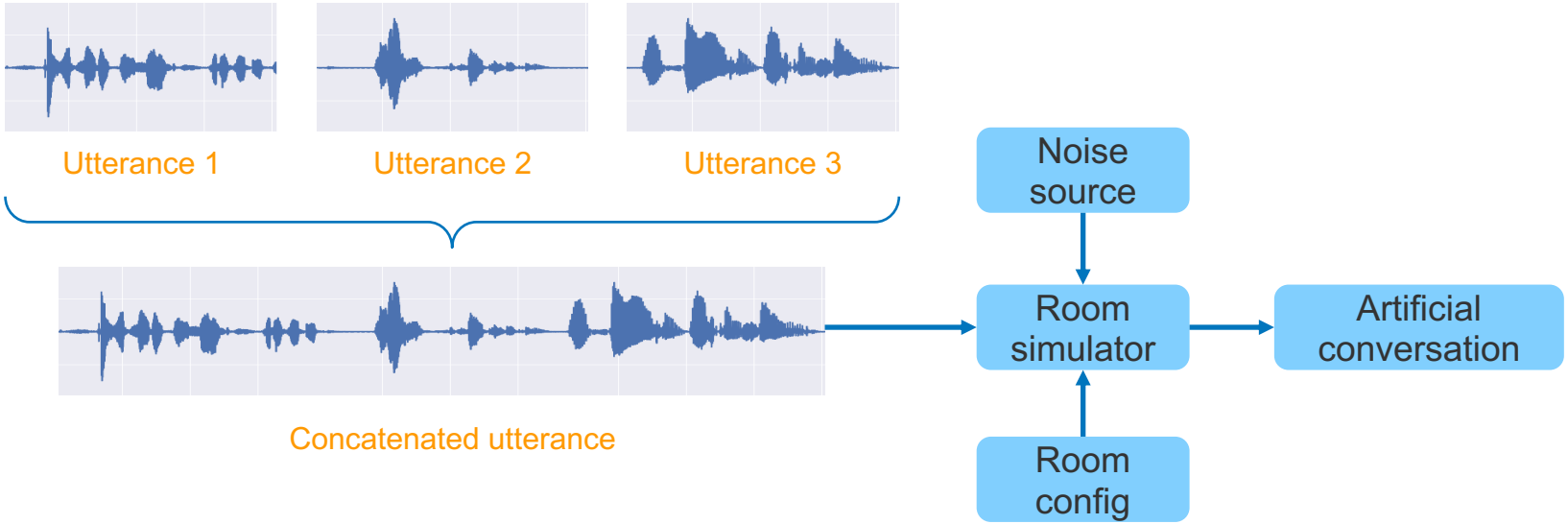
Force alignment

- VAD frame-level ground truth labels:
  - Use a pre-trained ASR model
  - Force align with ground truth ASR transcript



Artificial “conversational” speech

- Concatenation: Simulate speaker turns, with ground truth speaker labels
- Multi-style training (MTR): Avoid domain overfitting; mitigate concatenation artifacts



Model configuration

	Standard VAD	Personal VAD	Speaker verification
Configuration	• 2-layer LSTM, each layer has 64 nodes • 1 FC layer with 64 nodes	• 2-layer LSTM, each layer has 64 nodes • 1 FC layer with 64 nodes	• 3-layer LSTM, each layer has 768 nodes and 256-dim projections • 1 FC layer with 256 nodes
Number of parameters	0.13 million	0.13 million	4.82 million

Evaluation metrics

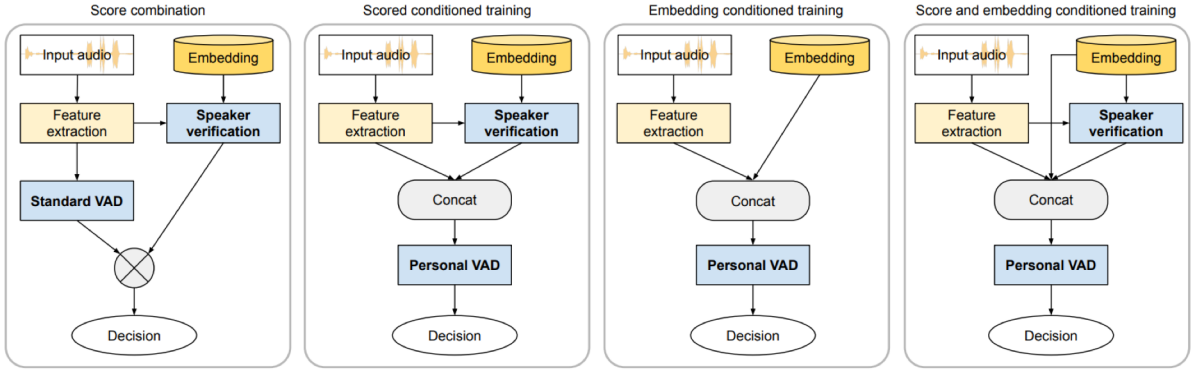
- Personal VAD is a classification problem
- We use Average Precision (AP):
  - For each class: `ns`, `tss`, and `ntss`
  - Mean Average Precision over all classes
- We also apply MTR on testing data
  - Explore personal VAD performance on noisy speech
  - Compare AP w/ and w/o MTR

Part 5:

Results and Conclusions

Architecture comparison

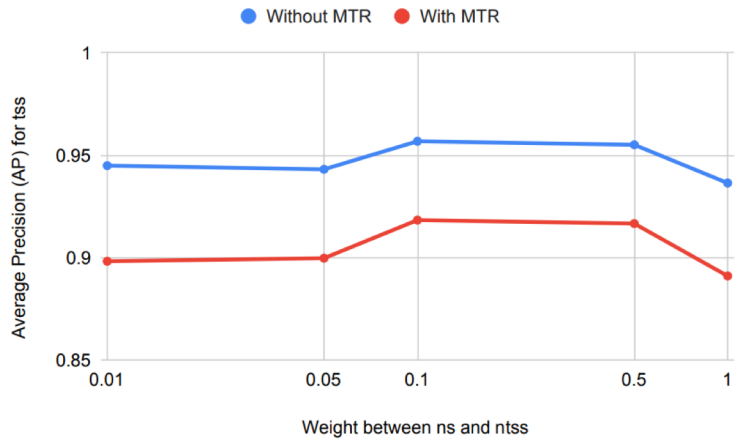
Method	Loss	Without MTR				With MTR				Network params
		tss	**ns**	**ntss**	mean	**tss**	**ns**	**ntss**	mean	
SC	CE	0.886	0.970	0.872	0.900	0.777	0.908	0.768	0.801	4.95M
ST		0.956	0.968	0.956	0.957	0.905	0.885	0.905	0.901	4.95M
ET		0.932	0.962	0.946	0.946	0.878	0.873	0.890	0.883	0.13M
SET		**0.970**	0.969	0.972	0.969	**0.938**	0.888	0.938	0.928	4.95M



- The proposed systems (ST, ET, SET) significantly outperform baseline (SC)
- ET achieve near-optimal performance with 2.6% parameters comparing to SET

Loss comparison

Method	Loss	Without MTR				With MTR				Network params
		tss	**ns**	**ntss**	mean	**tss**	**ns**	**ntss**	mean	
ET	CE	0.932	0.962	0.946	0.946	0.878	0.873	0.890	0.883	0.13M
	WPL	0.955	0.965	0.961	0.959	0.916	0.883	0.920	0.912	0.13M



- WLP achieves optimal performance when setting weight between **ns** and **ntss** to 0.1
- WLP (weight=0.1) outperforms CE

Personal VAD for standard VAD task

- Ultimate goal is to replace standard VAD by personal VAD
- Didn’t observe significant performance difference

Method	Loss	Without MTR		With MTR	
		speech	non-speech	speech	non-speech
Standard VAD	CE	0.992	0.975	0.975	0.918
Personal VAD (ET)	CE	0.991	0.965	0.979	0.893
Personal VAD (ET)	WPL	0.991	0.967	0.979	0.901

Conclusions

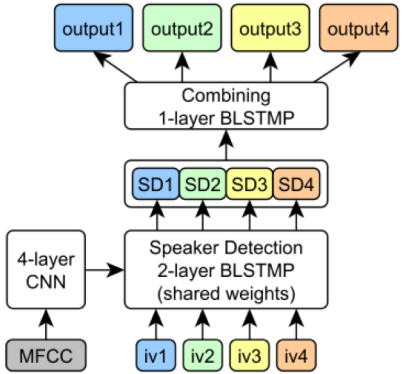
- Proposed personal VAD architectures outperform the VAD+SV baseline
  - SET architecture achieved best performance
  - ET architecture achieved near-optimal performance, with only 2.6% runtime parameters of SET
  - ET is ideal for on-device deployment
- Proposed weighted pairwise loss outperforms cross-entropy
- Personal VAD and standard VAD perform almost equally well on a standard VAD task

Part 6:

Future Work

Future work

- Training and evaluation on realistic (instead of artificial) conversations
  - This requires additional data collection and labelling efforts
- Personal VAD for speaker diarization (especially with overlapped speech)
  - Work already exists! “Target-Speaker VAD” system by Ivan Medennikov et al. [1]



[1] Medennikov, Ivan, et al. "Target-Speaker Voice Activity Detection: a Novel Approach for Multi-Speaker Diarization in a Dinner Party Scenario." *arXiv preprint arXiv:2005.07272* (2020).

Questions?


おでっせい
Odyssey
2020
The Speaker and Language Recognition Workshop
Tokyo, Japan