Robust Wake-Up Word Detection by Two-stage Multi-resolution Ensembles

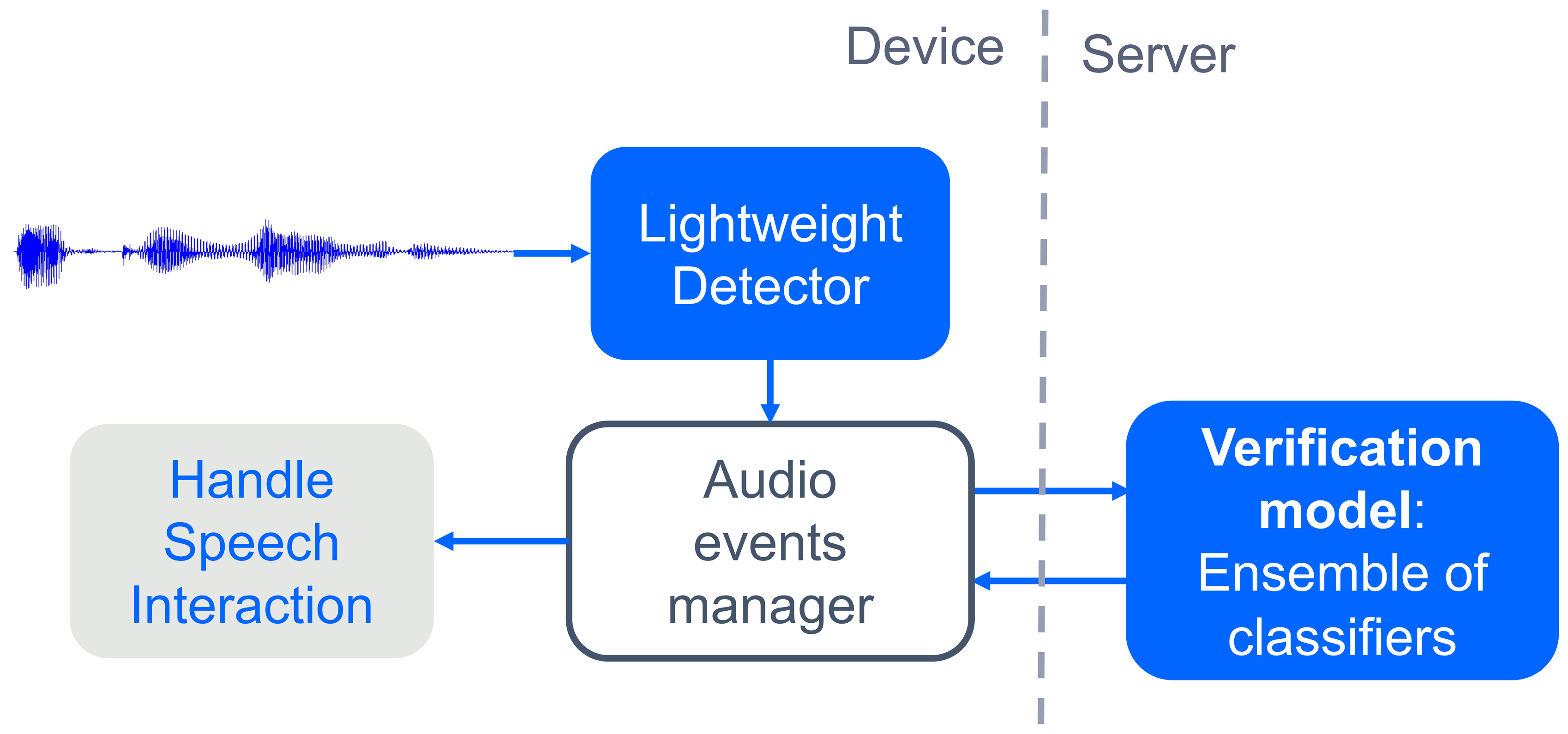
Fernando López[1,2], Jordi Luque[1], Carlos Segura[1], Pablo Gómez[1]

[1] Telefonica I+D, Spain
[2] Universidad Autónoma de Madrid

Introduction

Wake-up Word (WuW) in production scenarios priories **robustness, energy efficiency and minimizing communication delays**. Thus, this work proposes:

- Enhancing data with temporal alignments
- Parametric optimization of feature extraction
- Comparison of heterogeneous architectures in terms of performance and Real Time Factor (RTF)
- A two phases detection scheme with multi-resolution ensembles



Methodology

Database

- Augment “Okey Aura” database up to ~70 hours of audio
- + M-AILABS Spanish, SLR28, Valentini-Botinhao and new recordings

Audio processing

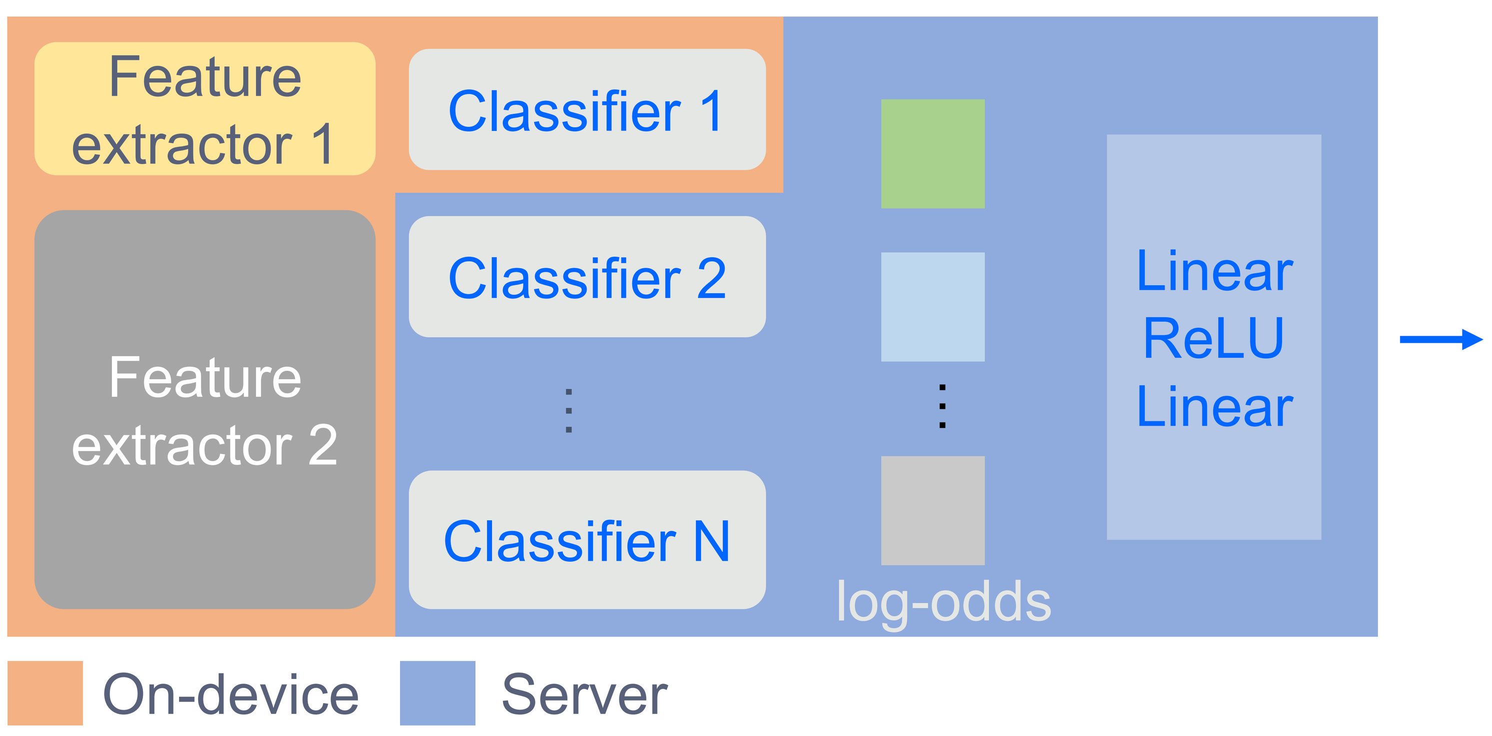
- CTC-based alignment of positive data
- MFCC’s parametric optimization:
  - Device: 13 coefficients, W=100ms, H=50ms
  - Server: 40 coefficients, W=30ms, H=10ms

Models

- Thirteen heterogeneous architectures
  - CNNs, RNNs, ResNets, Lambda Networks, Performers, Conformers and Broadcasted Residual Learning

Two-stages detection scheme:

- A lightweight on-device model for real-time processing
- A verification model on the server-side, which is an ensemble of heterogeneous architectures. The strength of different architectures is leveraged by using the stacking method



Experiments and Results

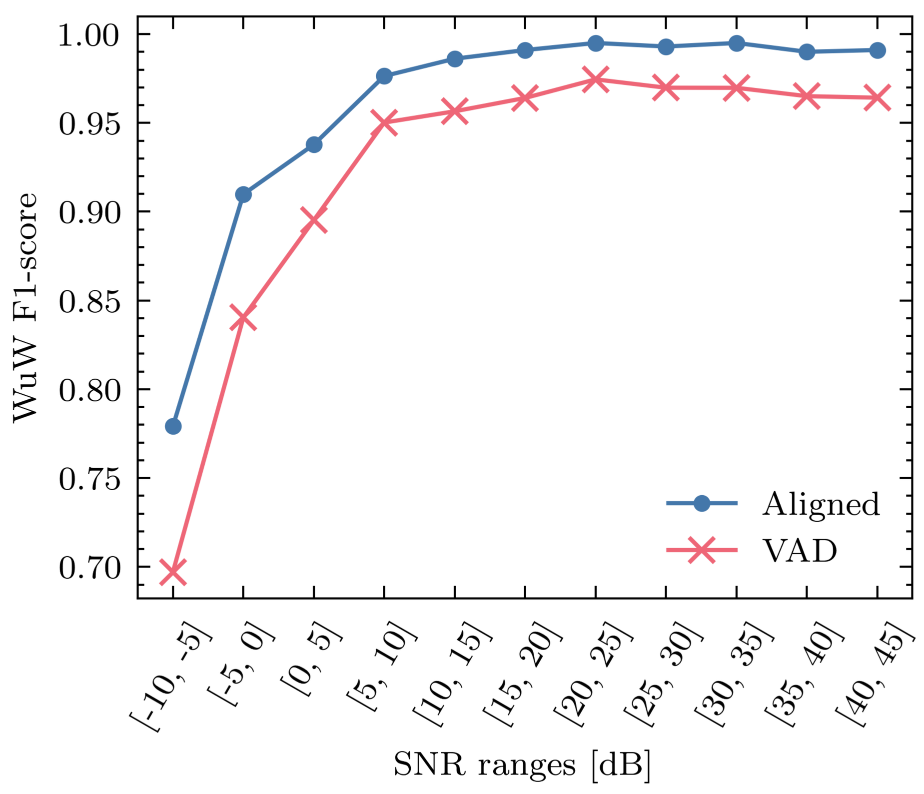
- Training and evaluation with a fixed-length window: 1.5 seconds
- Audio corrupted with background noise, SNR range: [-10, 50] dB
- Training from scratch during at most 700 epochs minimizing a Cross Entropy Loss
- Batch size of 128, and Adam optimizer with an initial Learning Rate (LR) of 0.001
- The LR is scheduled with on plateau reduction

Alignment Impact

sgru architecture trained with temporal annotations from:

- Voice Activity Detector (VAD)
- CTC-based aligner

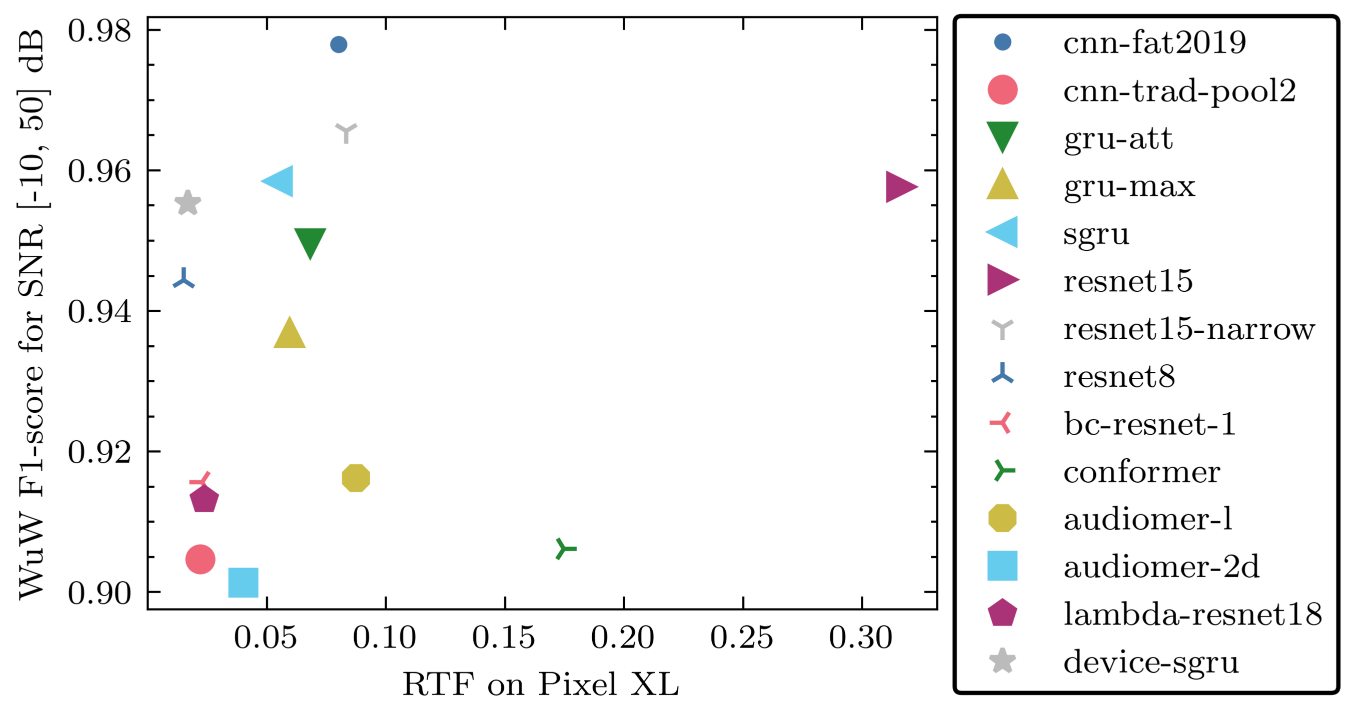
CTC-aligned data obtains an average **+4,175%** relative improvement in all SNR ranges



Audio classifiers

Thirteen audio classifiers compared in terms of performance and RTF on the Pixel XL

- Best trade-off for device is **device-sgru**
- Best classifier is **cnn-fat2019**



Ensemble

We experimented combinations of the best heterogeneous architectures using the stacking method

- The **ensemble-3** is the best performing ensemble. It combines three classifiers in the server-side with the on-device model
- The scheme allows to configure two operational points

	Models	F1-score	Improvement
	device-sgru	0.955	-
	cnn-fat2019	0.978	+2.408 %
ensemble-1	+ device-sgru	0.972	+1.780 %
ensemble-2	+ device-sgru + resnet15-narrow	0.977	+2.304 %
ensemble-3	+ device-sgru + resnet15-narrow + bc-resnet-1	**0.981**	**+2.723 %**
ensemble-4	+ device-sgru + resnet15-narrow + bc-resnet-1 + lambda-resnet18	0.958	+0.314 %

Conclusions

- Improvements in all SNR ranges thanks to use CTC-based speech-to-text alignments
- Selected different feature extraction for on-device and server detection (multi-resolution)
- Compared heterogeneous audio classifiers in terms of RTF and performance
- We propose a robust detection scheme with two-phases. Using two models, multi-resolution and ensembles we achieve a ~25ms delay in the first detection, an overall WuW F1-score of 0.981, and a WuW verification in ~293ms.

Telefónica
DeepLearning Barcelona Symposium