SpokeN-100

A Cross-Lingual Benchmarking Dataset for The Classification of Spoken Numbers in Different Languages

SpokeN

René Groh, Nina Goes and Andreas M. Kist
Friedrich-Alexander-Universität Erlangen-Nürnberg
{rene.groh, nina.goes, andreas.kist}@fau.de

tinyML summit
April 22–24, 2024
BURLINGAME/SAN FRANCISCO, CA

Problem

Speech to numerical input


Automatic recognition of a pronounced phone number

What syste capable of:

→ + 4 9 7 2 1 3 3 2 4 4 1

Reality:

→ + 49 72 11 78 33 24 41

We want a local system that is able to perform this task – is there a dataset for this?

- **Speech Commands/AudioMNIST: numbers 0 to 9**
- **Most datasets are monolingual**
- **Combining different data sets leads to inconsistent data quality**

SpokeN-100

Classification of spoken numbers

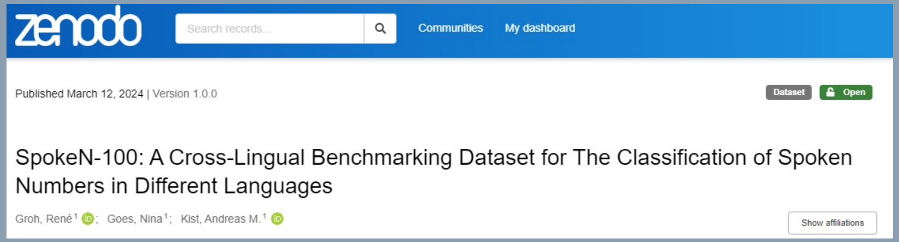
Spoken numbers from 0 to 99

32 different speaker

Four different languages: English, Mandarin, German and French

In total: 12,800 audio samples

Transparent, open-source, reproducible



→ **All data is artificially generated**

SpokeN-100 Dataset

Completely artificially generated data

LLM

	English	German	French	Mandarin
0	"Zero"	"Null"	"Zéro"	"零"
1	"One"	"Eins"	"Un"	"一"
2	"Two"	"Zwei"	"Deux"	"二"
…	…	…	…	…
N=99				

→ ChatGPT 3.5

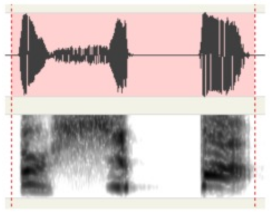
Generative Audio AI

IIElevenLabs

32 speaker

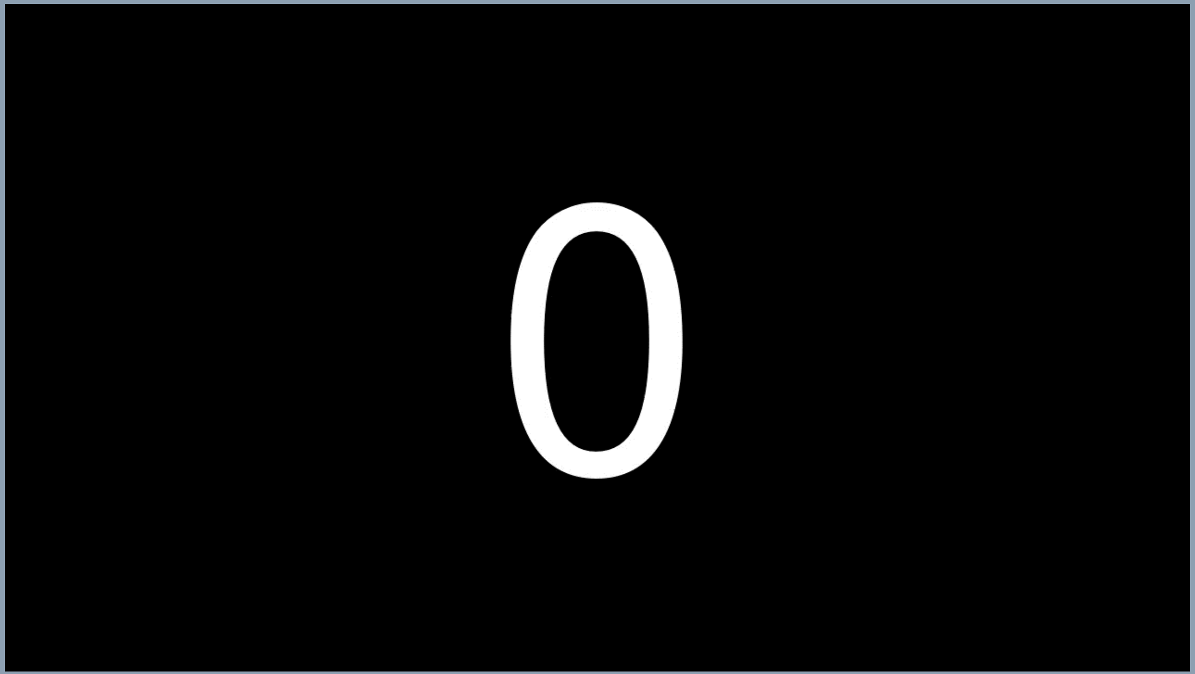
Audio Segmentation

WebMAUS



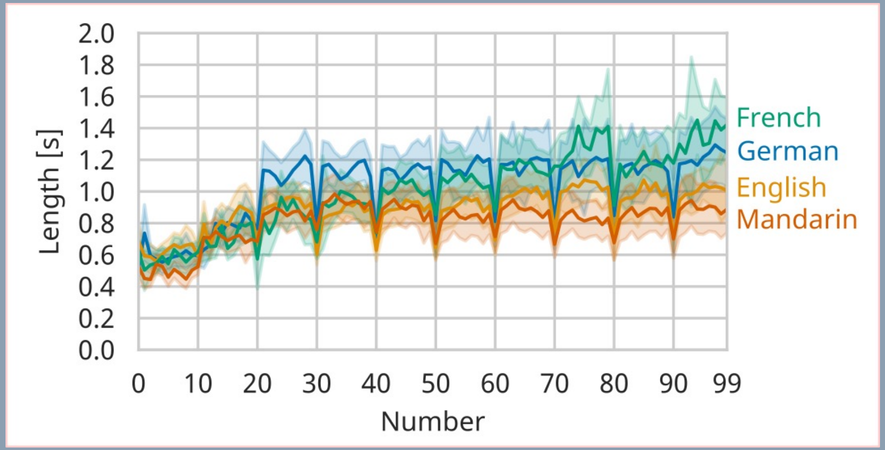
SpokeN-100 Dataset

Completely artificially generated data

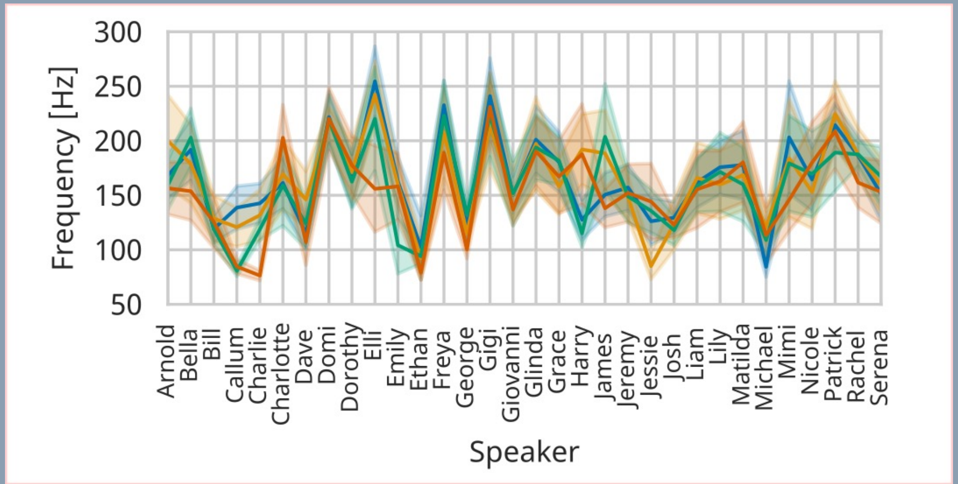


SpokeN-100 Dataset

Descriptive statistics of the generated data



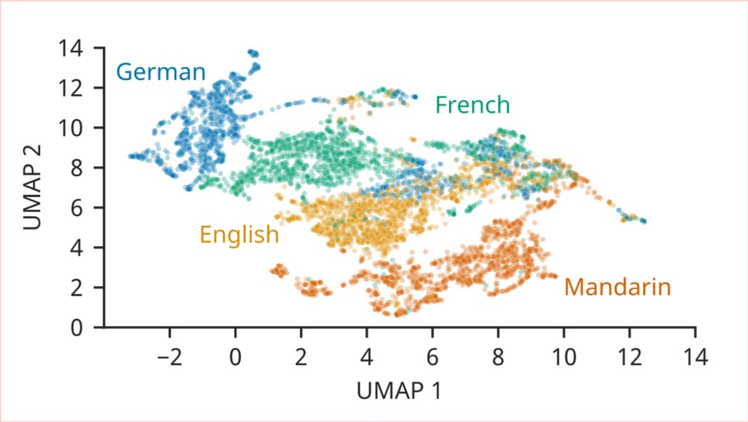
- Audio length increases as the numerical value increases
- Mandarin most efficient
- German numbers are the longest, except for higher numbers where French takes over

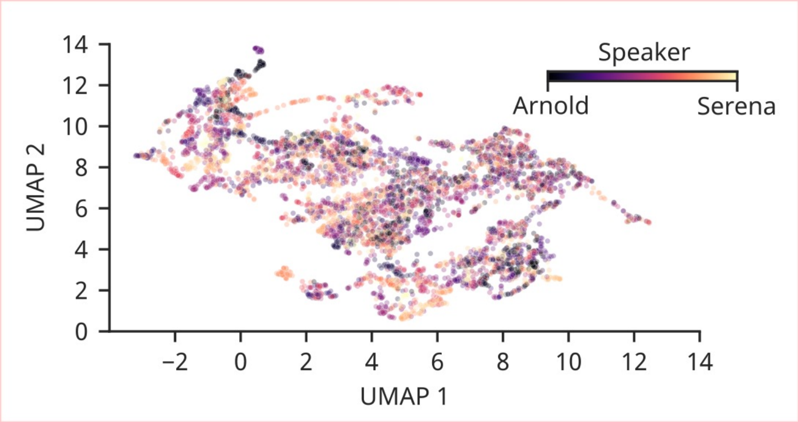


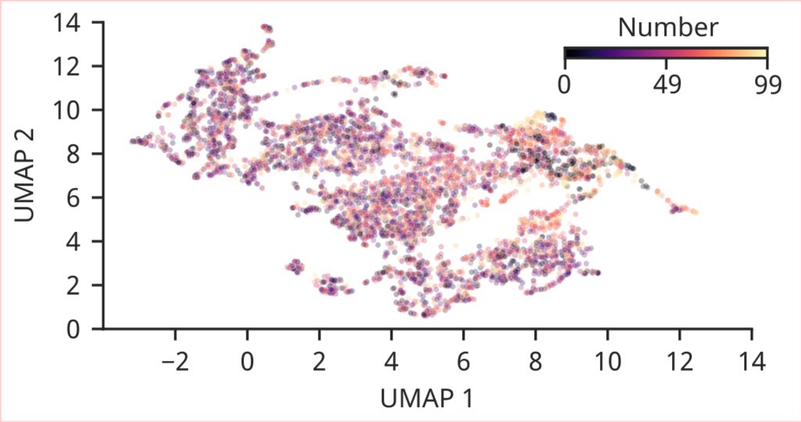
- Fundamental frequency similar for every speaker and every language
- Wide variety of fundamental frequencies

SpokeN-100

Dimensionality reduction of Mel-Spectrograms using UMAP







- Languages cluster together in the 2D visualization
- No clear clusters emerge when color-coded based on <u>speakers</u>
- No clear clusters emerge when color-coded based on <u>spoken number</u>

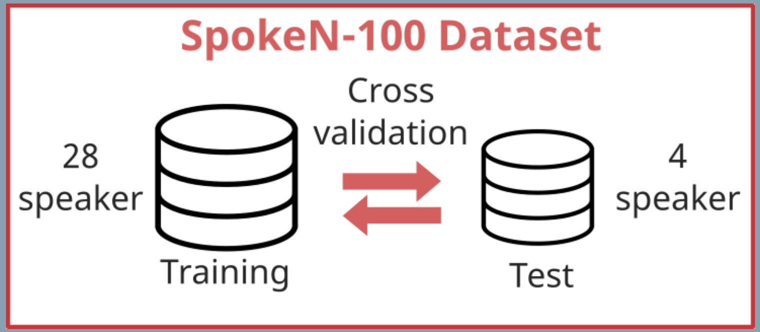
Benchmark task design

We introduce two classification tasks

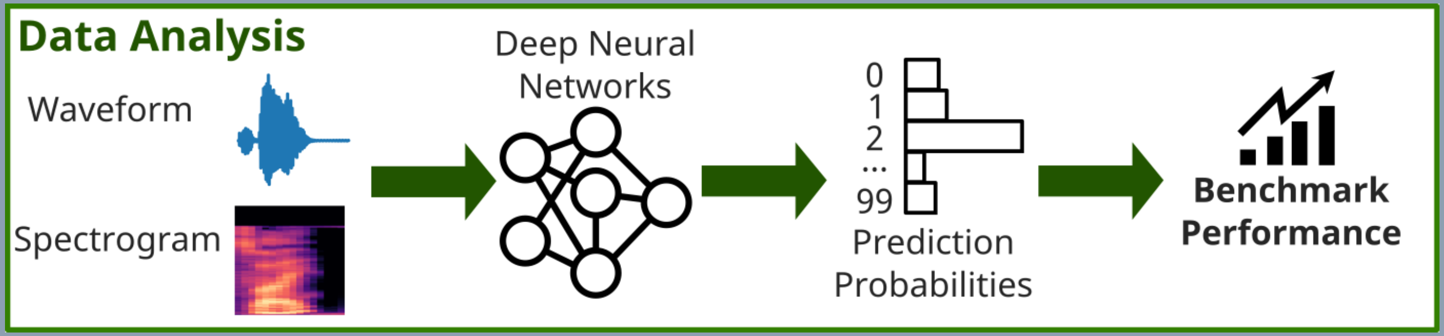
Classification of

a) **Languages** (Four classes)

b) **Spoken numbers** for <u>any language</u> (100 classes)

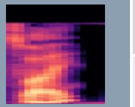


8-fold cross-validation



Results

Baseline performance



Architecture
1D CNN
2D CNN
RNN
EfficientNet-B0 CNN
Transformer

→ **Not deployable to a microcontroller**

Results

EvoNAS: Optimizing an architecture for MCU

Groh and Kist – „End-to-end evolutionary neural architecture search for microcontroller units“, 2023

nRF52840

- 64 MHz Cortex-M4
- 1 MB Flash
- 256 KB SRAM

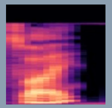
Source: https://www.nordicsemi.com/

Zephyr™ · TensorFlow

→ **Goal**: Maximize fitness which is a metric that combines multiple objectives (accuracy, inference time, energy consumption)

Results

Benchmark results for MCU-optimized architectures

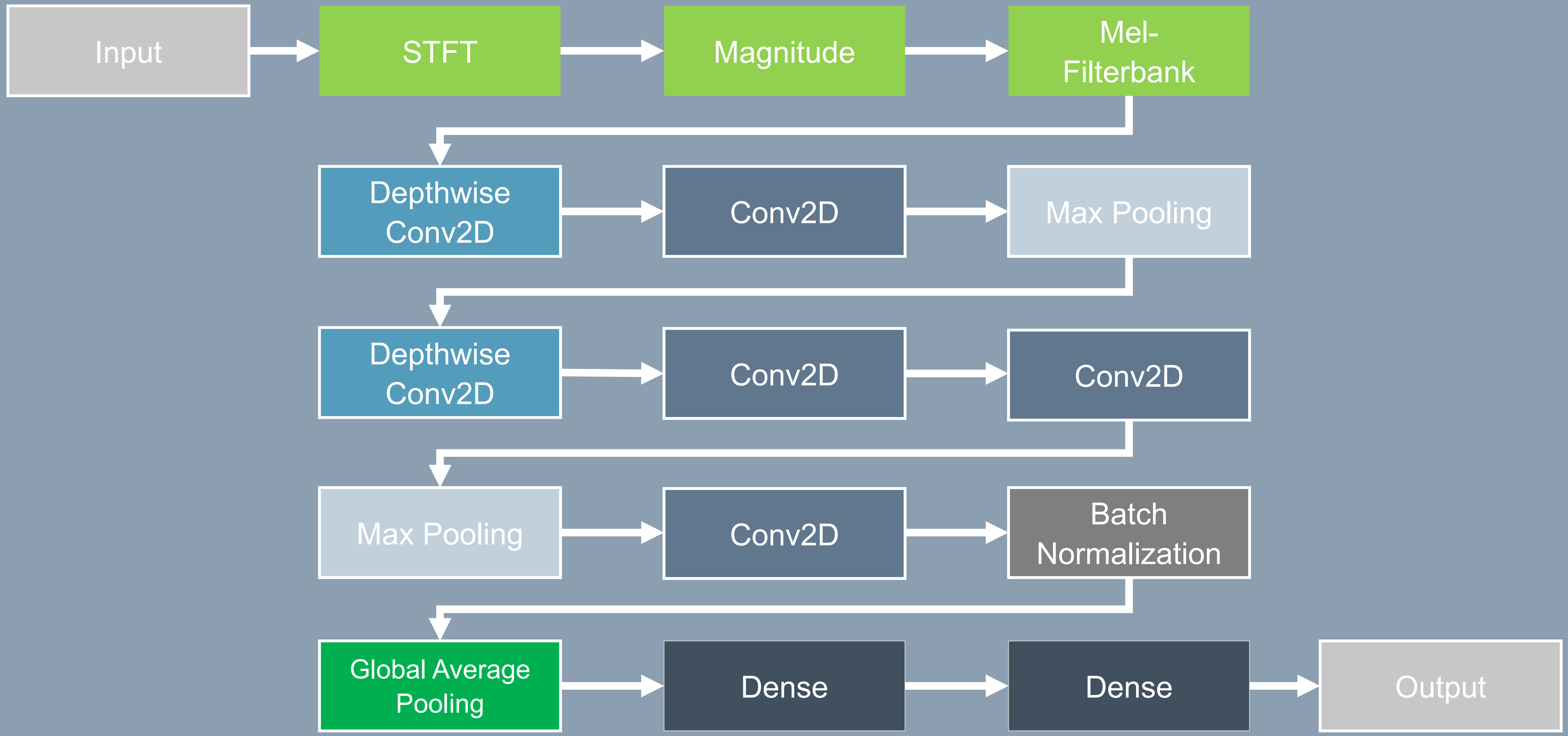


Architecture
EfficientNet-B0 CNN
Most accurate
Fastest
Fittest

→ **Proof-of-concept: huge optimization potential**

Results

Fittest model architecture



Conclusion

- SpokeN-100: Artificially generated speech recognition dataset with spoken numbers from 0 to 99 in four different languages
- It can serve as a benchmark for tinyDL datasets
- Data is not noisy, as with most other data sets → perfect for deep learning training, as you can control the amount of noise yourself
- High practical relevance for tinyDL applications: robot navigation, interaction with wearable devices, …
- Possibility to expand the database as required (voice cloning)

SpokeN

https://zenodo.org/records/10810044

Acknowledgements

Martin Fernholz
Prof. Dr. Tobias Bonhoeffer

Prof. Dr. Dave Berry
Dr. Dinesh Chhetri

Prof. Dr. Rebecca Leonard

Prof. Dr. Nicole Li-Jessen

Universitätsklinikum Erlangen

Prof. Dr. Michael Döllinger
PD Dr. Anne Schützenberger

Prof. Dr. Youri Maryn
Dr. Monique Verguts

Prof. Dr. Melda Kunduk

Prof. Dr. Matthias Echternach

Prof. Dr. Cara Stepp

Prof. Dr. Aaron Johnson

Prof. Dr. Cate Madill

Bayerisches Staatsministerium für Wissenschaft und Kunst

Bayerische Forschungsallianz

Fonds de recherche Santé Québec

Bundesministerium für Bildung und Forschung

IMPRS From Biology to Medicine

Joachim Herz Stiftung

Boehringer Ingelheim Fonds Stiftung für medizinische Grundlagenforschung

Emerging Talents Initiative

@rene__gr
@anki_xyz

rgroh1996
ankilab

Copyright Notice