SpokeN-100
A Cross-Lingual Benchmarking Dataset for The Classification of Spoken Numbers in Different Languages

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

Automatic recognition of a pronounced phone number

What systems are capable of:

+ 4 9 7 2 1 1 7 8 3 3 2 4 4 1

Reality:

+ 4 9 7 2 1 1 7 8 3 3 2 4 4 1

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 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
<table>
<thead>
<tr>
<th></th>
<th>English</th>
<th>German</th>
<th>French</th>
<th>Mandarin</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>&quot;Zero&quot;</td>
<td>&quot;Null&quot;</td>
<td>&quot;Zéro&quot;</td>
<td>&quot;零&quot;</td>
</tr>
<tr>
<td>1</td>
<td>&quot;One&quot;</td>
<td>&quot;Eins&quot;</td>
<td>&quot;Un&quot;</td>
<td>&quot;—&quot;</td>
</tr>
<tr>
<td>2</td>
<td>&quot;Two&quot;</td>
<td>&quot;Zwei&quot;</td>
<td>&quot;Deux&quot;</td>
<td>&quot;—&quot;</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
N=99

→ ChatGPT 3.5

Generative Audio AI
IIElevenLabs
32 speaker

Audio Segmentation
WebMAUS
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 speakers
- No clear clusters emerge when color-coded based on spoken number
We introduce two classification tasks:

Classification of

a) **Languages** (Four classes)
b) **Spoken numbers** for any language (100 classes)
Results

Baseline performance

Architecture

| 1D CNN | 16,368,340 | 0.815 ± 0.029 |
| 2D CNN | 1,891,972 | 0.927 ± 0.017 |
| RNN    | 5,709,700 | 0.920 ± 0.015 |
| EfficientNet-B0 CNN | 5,365,415 | 0.963 ± 0.011 |
| Transformer | 1,352,324 | 0.945 ± 0.012 |

→ 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

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

Benchmark results for MCU-optimized architectures

<table>
<thead>
<tr>
<th>Architecture</th>
<th>Parameters</th>
<th>Inference Time (ms)</th>
<th>Test Accuracy</th>
<th>Languages</th>
</tr>
</thead>
<tbody>
<tr>
<td>EfficientNet-B0 CNN</td>
<td>5,365,415</td>
<td>Not deployable</td>
<td>0.963 ± 0.011</td>
<td>0.848 ± 0.037</td>
</tr>
<tr>
<td>Most accurate</td>
<td></td>
<td>18,264</td>
<td>0.837 ± 0.022</td>
<td>0.175 ± 0.025</td>
</tr>
<tr>
<td>Fastest</td>
<td></td>
<td>6,496</td>
<td>0.364 ± 0.027</td>
<td>0.022 ± 0.003</td>
</tr>
<tr>
<td>Fittest</td>
<td></td>
<td>5,379</td>
<td>0.813 ± 0.021</td>
<td>0.126 ± 0.020</td>
</tr>
</tbody>
</table>

→ **Proof-of-concept:** huge *optimization potential*
Results
Fittest model architecture

Input -> STFT -> Magnitude -> Mel-Filterbank

Depthwise Conv2D -> Conv2D -> Max Pooling

Depthwise Conv2D -> Conv2D -> Conv2D

Max Pooling -> Conv2D -> Batch Normalization

Global Average Pooling -> Dense -> Dense -> Output
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)

https://zenodo.org/records/10810044
Acknowledgements

Prof. Dr. Nicole Li-Jessen
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

Martin Fernholz
Prof. Dr. Tobias Bonhoeffer
Prof. Dr. Dave Berry
Dr. Dinesh Chhetri
Prof. Dr. Rebecca Leonard

@anki_xyz
rgroh1996
@anki_xyz
Copyright Notice

This presentation in this publication was presented at the tinyML® Research Symposium 2024. The content reflects the opinion of the author(s) and their respective companies. The inclusion of presentations in this publication does not constitute an endorsement by tinyML Foundation or the sponsors.

There is no copyright protection claimed by this publication. However, each presentation is the work of the authors and their respective companies and may contain copyrighted material. As such, it is strongly encouraged that any use reflect proper acknowledgement to the appropriate source. Any questions regarding the use of any materials presented should be directed to the author(s) or their companies.

tinyML is a registered trademark of the tinyML Foundation.

www.tinyml.org