Common Phone

A Multilingual Dataset for Robust Acoustic Modelling
Philipp Klumpp, Tomás Arias-Vergara, Paula Andrea Pérez-Toro, Elmar Nöth, Juan Rafael Orozco-Arroyave

Introduction

Key features of Common Phone

- **Six Languages**
  - English
  - French
  - German
  - Italian
  - Russian
  - Spanish

- **Many Speakers**
  - More than 11,000 speakers recorded 116 hours of speech

**Phonetic Labels**
- 101 different phonetic symbols following IPA standards

**Made for Robustness**
- Countless number of microphones & environments

- **Training**
  - Fine-tuned Wav2Vec 2.0[3] base model (95 million parameters) with CP on 101 phonetic symbols
  - Model was pre-trained on English speech only
  - 3-step learning rate schedule:
    1. Warm-up (10 epochs)
    2. Plateau (30 epochs)
    3. Exponential decay (120 epochs)
  - Optimization with Adam & CTC loss

- **Testing**
  - Decoding beam width: 10
  - Phonetic symbol Error Rate (PER) as
    \[
    \text{PER} = \frac{N_{\text{dis}} + N_{\text{ins}} + N_{\text{err}}}{N_{\text{ref}}}
    \]

Material and Methods

**Motivation**
- Refined version of Common Voice[1]:
  - Eliminate imbalances
  - Enrich annotation
  - Preserve multilingual idea
- Provide reference dataset for:
  - Robust acoustic modelling
  - Testing in real-world environment

**Speaker selection**
- Only logged-in users were considered
- Gender-balanced distribution after 5-1-1 logic
- Age-balance: See Figure 1

**Phonetic Labelling**
- Automatic annotation with MAUS web-service[2]
- Pronunciation estimate as weighted output of G2P and ASR

**Distribution**
- Original MP3 files from CV
- Standard 16 kHz, 16 bits, single channel WAV
- Meta information for every speaker (age, gender, [dialect])
- Praat TextGrids with alignment information

Acoustic Modelling with CP

**Training**
- Fine-tuned Wav2Vec 2.0[3] base model (95 million parameters) with CP on 101 phonetic symbols
- Model was pre-trained on English speech only
- 3-step learning rate schedule:
  1. Warm-up (10 epochs)
  2. Plateau (30 epochs)
  3. Exponential decay (120 epochs)
- Optimization with Adam & CTC loss

**Testing**
- Decoding beam width: 10
- Phonetic symbol Error Rate (PER) as
  \[
  \text{PER} = \frac{N_{\text{dis}} + N_{\text{ins}} + N_{\text{err}}}{N_{\text{ref}}}
  \]

**Table 1:** PER (in %) observed for the six different languages of CP.

<table>
<thead>
<tr>
<th>Language</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>15.5</td>
<td>15.6</td>
</tr>
<tr>
<td>French</td>
<td>18.8</td>
<td>18.4</td>
</tr>
<tr>
<td>German</td>
<td>19.4</td>
<td>19.4</td>
</tr>
<tr>
<td>Italian</td>
<td>17.8</td>
<td>17.4</td>
</tr>
<tr>
<td>Russian</td>
<td>20.0</td>
<td>21.4</td>
</tr>
<tr>
<td>Spanish</td>
<td>14.5</td>
<td>15.0</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>17.8</td>
<td>18.1</td>
</tr>
</tbody>
</table>

**Conclusion**

**Long story short**
- Common Phone is a refined version of Mozilla’s Common Voice corpus collected from thousands of speakers
- The provided training, development and test splits resemble a more balanced distribution of speakers with respect to age, gender or language
- Reliable results for phonetic symbol recognition with SOTA acoustic model

**Who wants to use Common Phone**
- All speech researchers who want to
  - train models that are robust enough for deployment
  - test their models against a broad environment of signals
  - have phonetic labels for training/testing
  - want to work with multilingual data

**References**


Get Common Phone
Common Phone is available online via zenodo.com
A pre-print of the paper is available on arxiv.org

Acknowledgements