Recognizing Slips of the Tongue With Kaldi ASR

Gaynora van Dommelen

Institute for ICT, HU University of Applied Sciences, Utrecht



Background

More insight into the speech errors made by people with neurodegenerative diseases could possibly help with early diagnosis of these conditions.

Objectives

This study aims to prove whether the Kaldi ASR toolkit could be used to recognise 2 speech errors: 1) $\$ (five \rightarrow fipe, seven \rightarrow sepen) and 2) intended word \rightarrow combination of intended word and closely related word (one \rightarrow tw'one, five \rightarrow fou'five, nine \rightarrow eigh'nine).

Materials

OS: Ubuntu ASR toolkit: Kaldi

Language: Python & Bash

Train audio data: FSDD (Free Spoken Digit Dataset)
Test audio data: 2 male speakers (Dutch accent)

Hypothesis

The acoustic and language model probability scores for respectively the phone and word sequences can be used to determine speech errors in audio.

eight:eight/2.8064

five:five/2.7576

four:four/2.8577 nine:nine/2.6241

one:one/2.1386

seven:seven/2.7576

six:six/2.8064

three:three/2.5832

two:two/2.5063

zero:zero/2.2758

0/1.4077

two:two/2,5063

three three \$2,5832

seven seven 27576

1/1.4077

inex 6241

Prepare acoustic and Compute probabilty Record audio language data for scores of the LM and with speech ASR model AM for transcriptions errors Concatenate audio Train MONO and Analyse probabilties files to form TRI1 ASR model for recognizing utterances of 3 digits speech errors

The language model created from the lexicon and corpus.

 CER
 WER
 SER

 MONO
 0.85
 0.83
 2.50

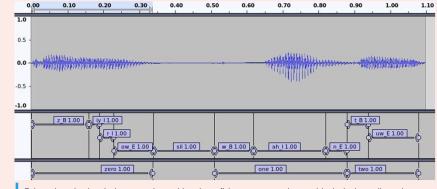
 TRI1
 0.85
 0.83
 2.50

Comparison of MONO and TRI1 Character Error Rate (CER), Word Error Rate (WER) and Sentence Error Rate (SER)



Lattice for the sentence 'nine zero one' which has only one path, The average number of paths found for each lattice is one.

Kaldi toolkit doesn't consider phone and word level probabilty scores of the AM and LM of importance and therefore retrieving these scores was deemed too big a task for the scope of this project.



Otherwise obtained phone and word level confidence scores along with their time aligned transcription. The sound wave is plotted along the time with on the y-axis the amplitude. Right underneath is the transcritpion of the phones found in the audio follwed by their confidence scores. Beneath the phone transcription is the word transcription follwed by their confidence scores.

Conclusion



For this dataset and model implementation, the confidence scores are not usable for determining speech errors, mainly due to the limited available data.