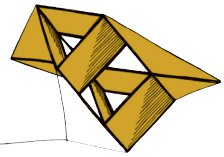

# Vyhodnocení rozpoznání řeči mluvčích s kognitivními vadami pomocí Wav2Vec

Filip Polák[1]

## 1 Úvod

Cílem této práce bylo vytvoření hlasového dialogového systému určenému pro sběr dat, které mohou být v budoucnu využité ve výzkumu předpovídání demence z řeči. Hlasový dialogový systém byl pojmenován Diadem (DIAgnóza DEMence). Jako data byly nahrány promluvy pacientů s mírnými kognitivními vadami, kdy pacienti 1 minutu popisovali živý obrázek pobřeží. Všechny promluvy byly poté ručně přepsány pro získání reference k vyhodnocení úspěšnosti prvních experimentů rozpoznání řeči na úrovni slov i znaků. Pro rozpoznání řeči byl použit state-of-the-art zero-shot Wav2Vec 2.0 rozpoznávač řeči [1].

## 2 Dialogový systém Diadem

Dialogový systém Diadem byl vytvořen pomocí platformy SpeechCloud [2], která umožňuje komunikaci v reálném čase mezi dialogovým manažerem (DM), který musí být implementován pro každý dialog, a tzv. SpeechCloud workery, které zprostředkovávají rozpoznávání řeči (ASR), porozumění řeči (SLU) a převod textu do řeči (TTS).

Jelikož byl dialogový systém Diadem vytvořen prozatím primárně pro sběr dat, jedná se o jednoduchý hlasový dialogový systém vykonávající sadu předdefinovaných akcí. Na začátku dialogu vyplnil pacient své osobní údaje a poté byl pomocí TTS pokynů veden sadou testů. Testy se skládaly z opakování čísel, která pacientovi řekl dialogový systém Diadem, poté bylo popisování živého obrázku pobřeží po dobu 1 minuty a na závěr vyjmenování co nejvíce zvířat za 30 sekund. Pro následné vyhodnocení rozpoznávání řeči byly vybrány pouze promluvy vztahující se k popisování živého obrázku pobřeží.

## 3 Popis dat a jejich vyhodnocení

Sběr dat probíhal ve dvou fázích. Před vytvořením dialogového systému Diadem se popisování živého obrázku pobřeží nahrávalo pomocí mikrofonu telefonu, který ležel na stole před pacientem. Nahrávání probíhalo v AD centru pro poruchy paměti nemocnice Královské Vinohrady. Bylo nahráno 17 pacientů, kteří měli kromě poruch kognitivních funkcí i lehké poruchy řeči, a byli označeni jako *Dataset1*, a bylo také nahráno 11 zdravých pacientů, označeni jako *Control Group*. V druhé fázi byl k nahrávání dat použit dialogový systém Diadem. Zde bylo nahráno 16 pacientů s kognitivními poruchami a ti byli označeni jako *Dataset2*.

K rozpoznání řeči byl použit grafémový Wav2Vec 2.0 rozpoznávač trénovaný nad českými daty. Jeho úspěšnost byla porovnána s ručním přepisem promluv. Pro vyhodnocení úspěšnosti byla vypočtena Levenshteinova vzdálenost reference a hypotézy. Ta má tvar $Acc = \frac{N-I-D-S}{N}$,

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Hlasové dialogové systémy, e-mail: polakf@students.zcu.cz

kde $N$ je celkové číslo tokenů (slov či znaků) v referenci a $I/D/S$ představují počet vložení/ smazání/ záměn tokenů.

Výsledná data jsou shrnuta v tabulce 1. Z dat je vidět korelace mezi přesností na úrovni slov a znaků, ačkoliv stejná přesnost na úrovni znaků pro *Dataset2* a *Control Group* není zobrazena ve stejnou přesnost na úrovni slov. To je pravděpodobně způsobeno principem, podle kterého Wav2Vec 2.0 rozpoznává řeč, a to je rozpoznávání grafémů (písmen abecedy). Výsledkem je pak vysoká úspěšnost na úrovni znaků a menší na úrovni slov, jelikož při přepisování promluv má člověk tendenci přepisovat je do srozumitelné řeči a může tedy dojít například k vložení znaků do přepisu, ačkoliv tyto znaky nejsou v promluvě zřetelně slyšitelné.

**Tabulka 1:** Přesnost na úrovni znaků a slov pro každý dataset vyjádřeno v procentech. *avg* znamená průměrnou úspěšnost a *std* znamená směrodatnou odchylku pro dataset.

| | Úroveň znaků | | | Úroveň slov | | |
|---|---|---|---|---|---|---|
| Groups | *avg* | ± | *std* | *avg* | ± | *std* |
| *Dataset1* | 83.5 | ± | 10.2 | 64.6 | ± | 11.9 |
| *Dataset2* | 90.0 | ± | 7.8 | 67.4 | ± | 13.3 |
| *Control Group* | 90.1 | ± | 5.7 | 74.1 | ± | 11.6 |

## 4 Závěr

Cílem práce bylo vytvořit hlasový dialogový systém, který by prováděl sběr dat pacientů s kognitivními poruchami. Data byla následně vyhodnocena pomocí state-of-the-art Wav2Vec 2.0 rozpoznávače. Experimenty ukázaly vysokou (cca 85%) přesnost na úrovni znaků a menší úspěnost (cca 75%) na úrovni slov, což je motivací pro další sběr dat, aby mohl být Wav2Vec 2.0 rozpoznávač přetrénován na doménová data.

Budoucnost této práce je pak data pouze nezaznamenávat, ale také jim porozumět. Cílem je zmapovat průběh promluvy v souvislosti s umístěním konceptů v obrázku (jestli pacient obrázek popisuje zleva doprava, shora dolů, přeskakuje atd.) a zahrnout do vyhodnocování lingvistické a fonetické vlastnosti promluvy (časování fonému, pomlky, přeřeky atd.).

Dalším úkolem je pak také doplnit do dialogového systému Diadem další testy, s jejichž pomocí se již dnes odhaluje demence při osobní návštěvě lékaře. Implementace těchto testů do dialogového systému Diadem by poté násobně zvýšila efektivitu těchto vyšetření, jelikož by byly přístupné z jakéhokoliv počítače či chytrého telefonu.

## Poděkování

Příspěvek byl podpořen grantovým projektem SGS-2022-017, Inteligentní metody strojového vnímání a porozumění 5.

## Literatura

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460, 2020.

[2] Jan Švec, Petr Neduchal, and Marek Hrúz. Multi-modal communication system for mobile robot. 4 2022.