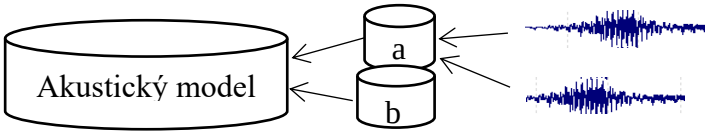
Vliv generovaných dat na úspěšnost akustického modelu

Jakub Nedvěd[1]

1 Úvod

Akustický model se skládá z řetězců příznaků jednotlivých fonémů. Základní podoba je znázorněna na obrázku 1. Čím více vzorů od daného fonému je k dispozici, tím přesněji bude rozpoznávač fungovat.



Obrázek 1: Příklad akustického modelu

Problém nastává při pořizování dat – získat velké množství dat je finančně i časově velmi náročné. Odpovědí na tento problém by mohla být určitá modifikace získaných dat. Cílem této práce je zjistit vliv generovaných dat na výslednou úspěšnost akustického modelu.

2 Generovaná data

K testu jsem se rozhodl využít dva typy dat – se změnou rychlosti a úpravou hlasu. Změnu rychlosti v dnešní době umí prakticky kterákoliv aplikace, zabývající se úpravou zvukových souborů a byl k ní využit softwarový nástroj SOX. K úpravě hlasu byl využit software pocházející z katedry kybernetiky, vytvořený Zdeňkem Hanzlíčkem (2009). Program na základě vstupního parametru za využití dynamického borcení časové osy pozmění hodnotu f0 v signálu. Hodnota f0 udává frekvenci základního hlasivkového tónu a výšku hlasu.

3 Trénování akustického modelu

Nejpoužívanějším způsobem trénování modelu je využití statistických metod, při kterých jsou slova či menší jednotky např. fonémy modelovány pomocí HMM. Akustický procesor převede řečové kmity na posloupnosti vektorů příznaků (O), zatímco lingvistický dekodér překládá tyto řetězce na řetězce slov (W). Cílem je nalézt posloupnost slov $\widehat{W}$, která maximalizuje podmíněnou pravděpodobnost $P(W|O)$. Hledáme tedy nejpravděpodobnější posloupnost slov pro danou akustickou informaci (vektor příznaků) – viz rovnice 1.

$$\widehat{W} = arg \max_{W} P(W|O) \quad (1)$$

Jako další byla využita neuronovou síť, konkrétně TDNN (time-delay), která simuluje síť s pamětí pomocí zpoždění mezi neurony, což je důležitě pro řečová data, která jsou závislá na konkrétním čase a kontextu promluvy.

Trénování proběhlo na datech o celkové době 5 603 minut, tedy více jak 93 hodin. Celkem bylo natrénováno a otestováno 6 modelů s následujícími parametry:

- Model 1 – triphonový model s 3 500 HMM stavy a 30 000 mixturami

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: nedvedj@students.zcu.cz

- Model 2 – triphonový model s 4 000 HMM stavy a 70 000 mixturami
- Model 3 – triphonový model s 6 000 HMM stavy a 140 000 mixturami
- Model 4 – triphonový model s 8 500 HMM stavy a 200 000 mixturami
- Model 5 – TDNN s 6 skrytými vrstvami s dimenzí 1024

4 Test

Test proběhl na datovém korpusu o době 107 minut, velikosti 14 753 slov. Datový korpus byl obohacen o generovaná data, vždy s odlišnými parametry

- Korpus 1 – základní data, neupravená
- Korpus 2 – základní data + data 1 s modifikací F0
- Korpus 3 – základní data + data 2 s modifikací F0
- Korpus 4 – základní data + data s modifikací F0 – data 1 a data 2
- Korpus 5 – základní data + data s modifikací rychlosti – zrychlení a zpomalení

5 Vyhodnocení

Pro vyhodnocení úspěšnosti akustického modelu byla využita Word Error Rate (WER) metrika. Jedná se o porovnání textového přepisu promluvy s rozpoznaným textem – výstupem z akustického modelu. Ve výsledku je zohledněn počet vkládání, záměn či mazání daných slov. Definujeme ji následovně

$$WER = \frac{S+D+I}{N}, \tag{2}$$

Přičemž

- S – počet slov, která jsou rozdílná v obou textech (Substitution)
- D – počet slov, která chybí v rozpoznaném textu (Deletion)
- I – počet slov, která jsou navíc v rozpoznaném textu (Insertion)
- N – počet slov v referenčním přepisu

V tabulce 1 jsou tučně vyznačeny nejlepší dosažené hodnoty pro daný model. Jak je patrné, nejlepších hodnot dosahujeme při trénování modelu pomocí TDNN.

Model\Dataset	**Korpus 1**	**Korpus 2**	**Korpus 3**	**Korpus 4**	**Korpus 5**
Model 1	**51,61**	52,03	52,31	52,25	51,97
Model 2	**50,08**	51,27	50,34	50,89	50,28
Model 3	44,07	44,07	44,01	**43,94**	44,49
Model 4	46,58	46,38	45,86	45,68	**45,5**
Model 5	37,48	39,96	40,4	40,05	**36,62**

Tabulka 1: WER [%]

6 Závěr

Z provedeného testu lze usuzovat, že rozšíření základního datového korpusu o generovaná data může vypomoci k dosažení lepších výsledků. K ověření této hypotézy byly porovnány výsledky z Model 5 pro Korpus 1 a Korpus 5. Korpus 5 dosahuje z **97,28%** lepších výsledků jak Korpus 1 s intervalem důvěryhodnosti 95%.

Literatura

Hanzlíček Zdeněk. (2009) *Automatická konverze hlasu v systému syntézy řeči*. Plzeň