FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2017

Přerušení uživatele na základě výstupu fonémového rozpoznávače při inkrementálním dialogu

Adam Chýlek[1]

Úvod

Hlasové dialogové systémy umožňují komunikaci člověka a stroje. Tento dialog se však stále neblíží podobě typického dialogu mezi lidmi. Takzvané inkrementální nebo též kontinuální zpracování vstupních a výstupních signálů pro řízení dialogu je dalším krokem k tomuto ideálu.

Klasický přístup k řízení dialogu považoval promluvu za ukončenou, když po určitou dobu nebyla detekována řeč. Inkrementální systém umožňuje reagovat již na určitou minimální část vstupu, což může vést ke zrychlení zpracování vstupní promluvy, ale především k rychlejší reakci na chybné vstupy. Pokud například řízení dialogu dostane informaci o tom, že právě vyřčené slovo uživatele bylo systémem rozpoznávání řeči rozpoznáno s nízkou mírou důvěry, může po uživateli požadovat potvrzení daného slova už během jeho promluvy. Představme si, že uživatel diktuje posloupnost devíti číslic a pátá a osmá byly rozpoznány s nízkou mírou důvěry. U klasického systému by se po dokončení promluvy uživatele systém musel dotázat v lepším případě např. „Patrně jsem nerozuměl pátou a osmou číslici, zopakujte je.“ V horším případě pak „Některým číslicím jsem nerozuměl, zopakujte vstup.“ Inkrementální systém by se mohl zeptat ihned po páté číslici např. jejím zopakováním „Osm?“ Podobně se pak dá postupovat i při zjištění informací, které jsou v konfliktu s aktuálním stavem dialogu.

V takových případech musí být řízení dialogu schopno najít vhodný okamžik, kdy začít se syntézou systémové promluvy. Takovým okamžikem může být např. až konec uživatelovy promluvy, nicméně při závažných chybách může být vhodnější uživatele přerušit dříve, „skočit mu do řeči.“

Předpokládejme nyní, že má řízení dialogu k dispozici posloupnosti fonémů reprezentující části vstupu. Představený algoritmus má za úkol právě na základě tohoto vstupu odhadnout, zda je vhodné uživatele přerušit či nikoliv.

Navrhovaný algoritmus a použitá data

K dispozici byla data z projektu MALACH (Psutka et al. (2011)) obsahující výpovědi česky mluvících přeživších holokaustu. Jednalo se o zvukové nahrávky s 2 kanály; jeden z nich byl záznamem z mikrofonu reportéra a druhý se záznamem z mikrofonu dotazovaného.

V nahrávkách bylo lidskými anotátory označeno 2073 úseků řeči, kde dotazovaný a reportér mluvili současně. Pokud předpokládáme, že účastníci čekají na vhodný okamžik, kdy druhého přeruší, můžeme začátek těchto úseků použít pro trénování řízení dialogu, které je schopno takový vhodný okamžik odhadnout.

Ze zvukových nahrávek byla systémem rozpoznávání mluvené řeči dekódována časově zarovnaná nejlepší hypotéza o posloupnosti fonému pro každý kanál nahrávky zvlášť.

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

Problém odhadu vhodného okamžiku přerušení pak můžeme stanovit jako problém binární klasifikace posloupnosti fonémů (je/není vhodné přerušení).

Řešení takového problému vyžaduje především stanovení toho, jak posloupnost reprezentovat. V tomto případě bylo rozhodnuto o použití pouze historie o určité časové délce. Experimentálně byla zjištěna ideální hodnota (z hlediska přesnosti detekce) délky 1 vteřiny předcházející promluvy. Kvůli rozdílnému počtu fonémů v úseku 1 vteřiny nelze použít jako reprezentaci prostou posloupnost indexů fonémů, bylo tak nutné najít reprezentaci, která vede na příznaky stejné délky.

Reprezentace tohoto úseku byla dvojí:

- pytel fonémů - základem byl vektor o velikosti fonémové abecedy. Každému fonému příslušela jedna pozice ve vektoru. Tento vektor byl na začátku nulový. Při procházení vstupní sekvence fonémů byla k příslušné pozici vektoru přičtena vždy určitá váha. Pro experimenty byly postupně sestaveny vektory kde jako váhy byly použity: 1 za každý výskyt, délka fonému v milisekundách, míra důvěry z rozpoznávače. Nejlepších výsledků dosahovala konkatenace 2 vektorů: s vahami za výskyt a vahami z míry důvěry.

- izochronní posloupnost - vektor byl vytvořen vzorkováním původní posloupnosti s krokem $0,01$ s. Pro každý krok byl vektor rozšířen o tyto příznaky: ticho (-1 pokud vstup reprezentoval ticho, 1 pokud byl fonémem), délka fonému v sekundách, míra důvěry z rozpoznávače.

Pro trénování klasifikátorů byly vygenerovány také negativní příklady. K tomuto účelu byly příznaky vytvořeny vždy z úseku který předcházel $0,25$ s danému přerušení.

Vyhodnocení

Pro vyhodnocení byly použity míry accuracy, precision a recall. K stanovení těchto měr byla dostupná data rozdělena na testovací a trénovací sadu v poměru $33:67$.

V případě přístupu s pytlem fonémů dosáhla nejlepšího výsledku klasifikace pomocí support vector machines s accuracy $0,56$, precision $0,53$ a recall $0,70$. U přístupu s izochronní posloupností se jednalo také o tento typ klasifikátoru s accuracy $0,62$, precision $0,63$ a recall $0,52$. Ověřované klasifikátory byly SVC, AdaBoost a Random Forest Classifier.

Vyhodnocení objektivní mírou však nemusí vypovídat o vhodnosti přerušení. Je možné, že samotná trénovací sada obsahovala místa, kde byl uživatel nevhodně přerušen. V rámci dalšího výzkumu tedy budou provedeny poslechové testy, které zajistí subjektivní zhodnocení zvolených algoritmů.

Poděkování

Příspěvek byl podpořen grantovým projektem SGS-2016-039.

Literatura

Psutka, J., Švec, J., Psutka, J.V., Vaněk, J., Pražák, A., Šmídl, L. a Ircing, P. (2011) System for fast lexical and phonetic spoken term detection in a Czech cultural heritage archive. *EURASIP Journal on Audio, Speech, and Music Processing*. Springer.