FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

STUDENTSKÁ VĚDECKÁ KONFERENCE 2013

Automatická detekce a vizualizace chyb konkatenační syntézy řeči

Jakub Vít[1]

1 Úvod

Syntéza řeči se snaží co nejlépe napodobit lidskou řeč. To je ale obtížné, neboť akustický signál lidské řeči je velmi pestrý a komplikovaný. Občas se v syntetické promluvě vyskytne úsek, který působí velmi rušivě. Pokud se jedná o lokální problém, hovoří se o tzv. *„artefaktu“*.

Práce se zabývá návrhem automatického systému detekce řečových artefaktů. S použitím tohoto systému by bylo možné nejen označit artefakty v syntetické promluvě, ale bylo by rovněž možné těmto artefaktům předcházet. Systém detekce chyb by měl automaticky odhalit artefakt v syntetické řeči. K tomu by měl použít dostupné parametry ze systému syntézy řeči či jiné snadno dostupné parametry. V práci je rozebírána syntéza řeči pomocí konkatenační metody *unit selection*. Ta je dnes jedna z nejpoužívanějších.

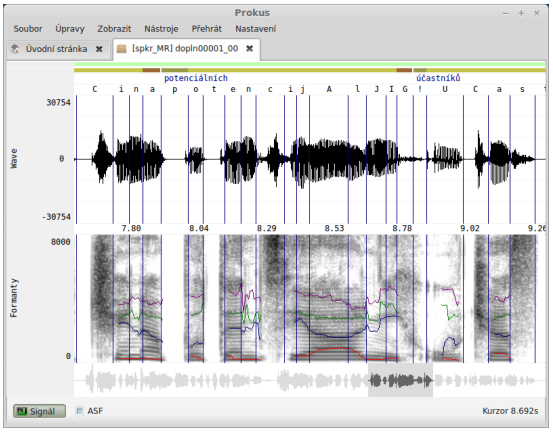
Před samotným návrhem systému je v práci nejdříve proveden rozbor problematiky řečových artefaktů. Je navržen způsob jak automatickou detekci sestavit. Ten je založen na sestavení a natrénování klasifikátoru z referenčních dat, která byla objektivně získaná pomocí poslechových testů, které byly prováděny na větším množství posluchačů.

2 Realizace

Systém detekce artefaktů je realizován klasifikátorem. Ten pro každou hlásku v syntetické řeči dokáže rozhodnout, zda dané místo je řečový artefakt. Jeho vstupem je vektor příznaků, který byl spočten z akustických a kontextových parametrů dané hlásky.

2.1 Analýza syntetické řeči

Pro pochopení příčin vzniku artefaktu je třeba procházet velké množství syntetických promluv. V nich je třeba studovat průběh audio vlny a také průběhy ostatních parametrů a spektra. Na všechny tyto funkce existují programy nebo jiné nástroje. Neexistuje však žádný program, který by všechny tyto funkce dokázal sjednotit a napojit na systém *ARTIC* (systém syntézy řeči na katedře kybernetiky FAV ZČU). V rámci práce byl proto takový program vytvořen a představen v jedné kapitole této práce. Tento program umožňuje vizualizovat a analyzovat proces syntézy řeči.



Obrázek 1: Analýza syntetické řeči

[1] student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika, e-mail: vit89@students.zcu.cz

2.2 Příprava dat

Pro trénování systému detekce artefaktů je nutné získat referenční data. Ta budou reprezentovat sadu objektivně označených artefaktů. Vnímání artefaktu je ale velmi subjektivní záležitost. K sestavení objektivních označení bylo nutné provést poslechový test na více posluchačích. Součástí práce je tedy i návrh a vytvoření poslechových testů. V poslechových testech odpovídalo *20* posluchačů. Z celkových *7200* odeslaných odpovědí bylo označeno *4700* podezřelých úseků, které sloužily jako referenční data pro trénování klasifikátoru.

Obrázek 2: Aplikace pro poslechové testy

3 Trénování klasifikátoru

Jako klasifikátor byl zvolen SVM (*support vector machines*). SVM je poměrně mladá metoda strojového učení. Jedná se o lineární binární klasifikátor. Klasifikátor byl učen pomocí *RBF kernelu*. Při hodnocení klasifikace se používala *10-fold cross validace*. Každý vzorek artefaktu měl přiřazenou svoji váhu. Ta byla zohledněna při klasifikaci. Hodnota váhy vyjadřovala jak moc věrohodný vzorek je. Pokud například v poslechových testech více posluchačů dané místo označilo, váha byla vyšší.

Samotné trénování klasifikátoru bylo provedeno ve čtyřech experimentech. Každý experiment měl jinak vybraná trénovací data. Experimenty EXP1 a EXP2 obsahovaly (narozdíl od EXP3 a EXP4) jen takové artefakty, které kolem sebe neměly další artefakty. Ověřovala se tak hypotéza, že příčinou artefaktu je vždy jen jedno místo a ne sekvence jednotek. Trénování bylo vždy provedeno jak s třetinou nejlepších vzorků (EXP1 a EXP3), tak se všemi vzorky (EXP2 a EXP4). Tím se ověřovalo správné nastavení vah vzorků. Ve všech experimentech byla použita vyvážená množina trénovacích dat.

Tabulka 1: Výsledky experimentů (R - *Recall*, P - *Precision*, A - *Accurracy*)

	Počet vzorků				Nevážený SVM				Vážený SVM			
	N_p	N_n	$N_p^{(vse)}$	$N_n^{(vse)}$	R	P	$F1$	A	R	P	$F1$	A
EXP1	500	500	1574	3605	0.72	0.80	0.76	0.74	0.79	0.88	0.83	0.78
EXP2	1574	1574	1574	3605	0.63	0.80	0.71	0.67	0.80	0.95	0.87	0.79
EXP3	1000	1000	2458	4025	0.68	0.81	0.74	0.71	0.79	0.91	0.85	0.78
EXP4	2458	2458	2458	4025	0.61	0.76	0.68	0.64	0.79	0.95	0.86	0.79

4 Závěr

V práci byl navržen a sestaven systém automatické detekce řečových artefaktů v syntetických promluvách. Pomocí programu speciálně vyvinutého pro tyto účely byly prozkoumány místa v okolí artefaktů. Na základě dat z poslechových testů byla sestavena referenční data pro testování klasifikátoru. V nejlepších konfiguracích dokázal klasifikátor při použití vah dosáhnout úspěšnosti téměř 80 %.

Takto natrénovaný klasifikátor by šlo použít přímo v systému syntézy řeči pro lepší výběr jednotek. S jeho pomocí by mělo být možné snížit četnost výskytu řečových artefaktů v syntetických promluvách. V budoucí práci je možné zaměřit se právě na takový experiment.