Rozšířený Abstrakt

Robustní odhad odstupu řeči od šumu pomocí hlubokých neuronových sítí

Bc. Michal Mužíček, Ing. Jiří Málek, Ph.D.

Abstrakt

Práce se zabývá tvorbou neuronové sítě, která je schopná, i přes výskyt různorodého šumu, odhadnout, kde se v řečové nahrávce vyskytuje řeč. Jako vstupní data pro trénování neuronové sítě slouží databáze aditivní směsi šumu a čistých řečových nahrávek. Data zpracovaná neuronovou sítí jsou následně předána algoritmu, který vypočítá odhad odstupu řeči od šumu. Správnost výstupu navrženého algoritmu je hodnocena dle porovnání s konkurenční metodou WADA. Výsledné hodnoty naznačují, že využití neuronových sítí pro detekci přítomnosti řeči a následného odhadu SNR úrovně jsou reálnou alternativou existujícím metodám.

Úvod

Každý reálný signál je součet užitečné(pro mojí aplikaci) komponenty a neužitečné komponenty (označujeme jako šum, interference). Jedním ze základních problémů při zpracovávání signálů v oblasti rozpoznávání řeči je pak zjištění, jak moc zašuměná je zpracovávaná nahrávka.

Cílem této práce je vytvořit algoritmus pro odhad odstupu řeči od šumu (SNR) v nahrávce s použitím detektoru řečové aktivity (VAD) [1], který je realizovaný pomocí neuronové sítě. Stejnou problematikou se již zabýval Zhang Wu [2]. Jeho řešení je podobné, avšak liší se použitými prvky v algoritmu.

Takový algoritmus pak lze použít pro získání informace o míře zarušení signálu, což je stěžejní informace pro hlasové detekční úlohy typu rozeznání řečníka, rozeznání řeči apod. Tedy tam, kde je třeba odhadnout intenzitu šumu a jeho vliv na užitečnou informaci (řeč).

Experiment a metody

První částí algoritmu je VAD založený na hluboké neuronové síti. Tato sít má 4 vrstvy. Z toho 3 vrstvy jsou skryté a každá z nich používá Tansig aktivační funkci a má 128 neuronů, kvůli dostatečnému prostoru pro extrakci vlastností řečového signálu. Výstupní vrstva používá Softmax funkci a má 2 neurony, protože výstupem sítě je klasifikace do 2 kategorií (řeč/neřeč). Pro učení používá optimalizační kritérium Cross Entropy, které je navrženo pro hodnocení správnosti klasifikace.

Jakmile algoritmus získá binární VAD vektor o nahrávce, tak je na řadě odhad SNR bez referenčního signálu [3]. Toho dosáhne pomocí adaptivního okénka o velikosti 30 vzorků a faktoru zapomínání 0.98 spočítá pro každý vzorek signálu odhad okamžitého výkonu šumu. Tento odhad je pak použit pro výpočet odhadu výkonu řeči v řečových vzorcích tím, že se odečte od okamžitého výkonu daného vzorku. Čímž má k dispozici jednotlivé výkony šumu a řeči a je již schopen vypočítat odhadovanou hodnotu SNR úrovně.

Výsledné testování proběhlo na neviděné testovací množině, která bylo vytvořena z nahrávek čisté řeči a aditivního šumu z reálného prostředí Ulice (hlavní charakteristikou tohoto šumu jsou projíždějící auta v pozadí, tento šum je stacionárního charakteru).

Výsledky a diskuze

Tabulka 1. Hodnocení odhadu SNR úrovně pomocí VAD sítě a SNR odhadu

GSNR [dB]	Bias	Variance	MSE
0	-4.5	2.3	101.1
5	-2.5	0.8	4.4
10	-1.2	0.8	0.8

Tabulka 2. Hodnocení odhadu SNR úrovně pomocí WADA

GSNR [dB]	Bias	Variance	MSE
0	-0.7	3.1	4.1
5	-0.5	1.0	0.3
10	-0.4	1.2	0.2

Z Tabulky č. 1 je vidět, že čím intenzivnější je řeč v nahrávce, tím přesněji algoritmus cílovou úroveň odhadl. Je třeba poznamenat, že důvod proč se Variance liší od ostatních úrovní u úrovně 0 dB GSNR, je ten, že se v testovacích datech vyskytl tzv. outlier. Jedná se o signál, který se svými vlastnostmi silně odlišuje od ostatních v dané skupině. V tomto případě to znamená, že algoritmus nebyl schopen danou nahrávku správně odhadnout a nahrávka dostala velmi vzdálenou hodnotu od cíle.

V tabulce č. 2 je možné vidět efektivitu odhadu SNR pomocí WADA [4]. WADA má odhady blíže k cílové hodnotě (malý Bias), ovšem častěji jsou jednotlivé hodnoty odhadů od sebe vzdáleny ve větší míře (větší Variance).

Závěr

Výsledné odhady algoritmu navrženého v práci jsou srovnatelné s metodou WADA. WADA je lepší na stacionárních datech a naše metoda je lepší na nestacionárních datech (až na Varianci s outlierem). Wada má výhodu, že je to statistický přístup, tedy nemá Neviděná data. Ovšem naše metoda je robustní i na těchto neviděných datech.

K výsledným odhadům je třeba podotknout, že jelikož algoritmus odhadu globálního SNR je založen na vypočítání energie řečových segmentů, tak čím méně řečových segmentů se vyskytovalo v nahrávce, tím méně přesnější byl samotný odhad

Reference

[1] Zhang XL, Wu J. Deep *belief networks based voice activity detection*. Audio, Speech, and Language Processing, IEEE Transactions on. 2013 Apr;21(4):697-710.

[2] Vondrasek, Martin, and Petr Pollak. *Methods for speech SNR estimation: Evaluation tool and analysis of VAD dependency*. Radioengineering 14.1 (2005): 6-11.

[3] Adamec, M. Moderní *rozpoznávače řečové aktivity*. (2008):32-33 [online]. [cit. 2016-04-25] https://dspace.vutbr.cz/bitstream/handle/11012/16807/Diplomova_prace_Michal_Adamec.pdf

[4] Ellis, Dan. Objective measures of speech quality/SNR [online]. [cit. 2016-04-25]. http://labrosa.ee.columbia.edu/projects/snreval/