Robustní odhad odstupu řeči od šumu pomocí hlubokých neuronových sítí

Bc. Michal Mužíček, Ing. Jiří Málek, PhD.

michal.muzicek@tul.cz, jiri.malek@tul.cz

TECHNICKÁ UNIVERZITA V LIBERCI
Fakulta mechatroniky, informatiky a mezioborových studií

Studentská Konference Fakulty Mechatroniky, informatiky a mezioborových studií 2. červen 2016, Liberec, Česká republika

Abstract

This documentation describes a creation of a neural network that is capable of locating the location of speech in audio sample. Database containing additive mixture of noise and speech signals is used as an input for training of the neural network.
Output from this network is then processed by an algorithm, which computes an estimation of signal to noise ratio. Performance of this algorithm is then compared against performance of WADA, a conventionally used software.

Cíl

Cílem práce je vytvořit algoritmus pro robustní odhad SNR v nahrávce za pomocí hlubokých neuronových sítí a srovnat efektivitu algoritmu s již existující technologií WADA [1]. Přínos této práce spočívá v předvedení efektivity neuronové sítě v problematice detekce přítomnosti řeči (VAD) a následného odhadu SNR úrovně.

Úvod

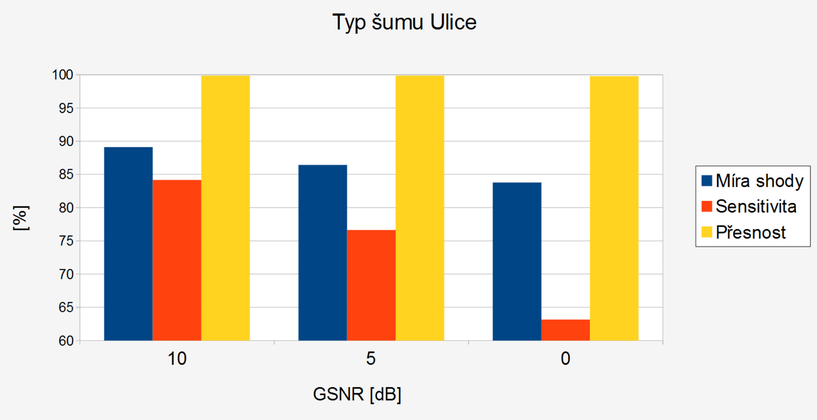
- každý reálný řečový signál je součet užitečné komponenty (pro mojí aplikaci to je řeč) a neužitečné komponenty (šum)
  - je třeba zjistit, jak moc zašuměná je zpracovávaná nahrávka
  - hovoříme o odstupu řeči od šumu (Signal to Noise Ratio, dále jen SNR)
  - určuje poměr energií užitečné komponenty vůči neužitečné komponentě v digitální nahrávce
  - zjistit přesně lze pouze v laboratorních podmínkách -> v reálném světě je třeba odhadnout

Metodika

- natrénování VAD sítě
  - vytvoření trénovací a testovací množiny z aditivní směsi
    - databáze reálných šumů ChiME [2]
    - databáze čistých řečových nahrávek TiMIT [3]
  - rozdělení na rámce o 512 vzorcích s překryvem 256 vzorků
  - získání reprezentačního vektoru 39 logaritmických frekvenčních příznaků
  - vytvoření cílových VAD dat (pro trénink sítě) pomocí hranice LSNR -5 dB
- výstup VAD sítě předán algoritmu pro odhad úrovně SNR
  - adaptivní okénko s faktorem zapomínání pro odhad energie šumu a následně řeči

Obrázek 1: Schéma neuronové sítě vygenerované prostředím Matlab

Výsledky



Obrázek 2: Efektivita nejlepší VAD sítě na testovací sadě - Ulice

V Grafu je vidět efektivita VAD sítě pro testovací množinu s neviděnými daty (jak šumové tak řečové signály nebyly použity při trénování).

- jedná se o šum, kde jsou v pozadí slyšet projíždějící auta (stacionární šum)
  - má velmi vysokou Přesnost klasifikace
  - Sensitivita sítě je podstatně nižší ()síť nebyla schopná správně rozeznat všechny řečové segmenty)
- síť je schopná rozpoznávat zašuměné nahrávky jak se stacionárním šumem, tak i nestacionárním
- síť byla schopná rozpoznat neviděný typ šumu s dobrými výsledky, proto bych tuto síť označil za robustní

Ulice

VAD síť + SNR odhad

SNR	Bias	Variance	MSE
0	-4,5	2,3	101,1
5	-2,5	0,8	4,4
10	-1,2	0,8	0,8

Ideální VAD + SNR odhad

SNR	Bias	Variance	MSE
0	-3,9	0,7	8,3
5	-2,2	0,6	1,6
10	-1,2	0,7	0,7

WADA

SNR	Bias	Variance	MSE
0	-0,7	3,1	4,1
5	-0,5	1,0	0,3
10	-0,4	1,2	0,2

Tabulka 1: Srovnání odhadů pro šum typu Ulice

- u cílové úrovně 0 dB pro VAD síť se Variance zdá být podstatně vyšší (než u 5 a 10 dB) -> příčinou je tzv. outlier
  - nahrávka, kterou algoritmus natolik špatně odhadl, že dostala podstatně menší hodnotu -> větší vzdálenost od cílové hodnoty
  - stejný problém měl i WADA s několika nahrávkami, jelikož jinak se i jeho Variance pohybuje okolo 1 dB (byl ovšem přesnější s celkovým odhadem)
- WADA má lepší odhady ale zároveň větší množství hodnot, které se vzdálily od cílového SNR

Závěr

Výsledné odhady algoritmu navrženého v práci jsou srovnatelné s metodou WADA. WADA je lepší na stacionárních datech a navržená metoda je lepší na nestacionárních datech (až na Varianci s outlierem). WADA má výhodu, že je to statistický přístup, tedy nemá Neviděná data. Ovšem navržená metoda je robustní i na těchto neviděných datech.
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] ELLIS, Dan. Objective measures of speech quality/SNR. Labrosa. [online]. 8.4.2011 [cit. 2016-02-15]. Dostupné z: http://labrosa.ee.columbia.edu/projects/snreval/

[2] BARKER, John, MARXER, Ricard, VINCENT, Emmanuel a WATANABE, Shinji. The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. *IEEE 2015 Automatic Speech Recognition and Understanding Workshop*. 2015.

[3] GAROFOLO, J.S., LAMEL, L.F., FISHER, W.M., FISCUS, J.G. a PALLETT, D.S. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. NASA STI/Recon Technical Report N, 93. 1993.

Poděkování

Tato práce byla podpořena z projektu Studentské grantové soutěže (SGS) na Technické univerzitě v Liberci v roce 2016.