Zlepšování řečových nahrávek pořízených v reálném prostředí

Bc. Tomáš Kounovský, Ing. Jiří Málek, Ph.D.
tomas.kounovsky@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 thesis is focused on enhancement of real speech recordings by noise removal in the log-power spectrum using deep neural networks. Two networks are trained with differently oriented training sets to learn the mapping from noisy to clean speech. The resulting network performances are evaluated in matched and mismatched conditions and compared to some conventionally used methods. Results show that a network trained with a database containing more noise types has more robust speech enhancement capabilities in mismatched conditions. Performance of this network is comparable with performance of conventionally used methods in stationary noise conditions and surpasses them when the noise is highly non-stationary.

Úvod

Zlepšování jednokanálových řečových nahrávek ve spektrální oblasti je dlouho řešeným tématem. Konvenční přístupy (odečítání spektra, MMSE STSA) často nejsou schopny vystihnout povahu dynamických šumů s nízkými úrovněmi SNR [1]. Modernější metody, např. OMLSA [2], dosahují lepších výsledků pro nestacionární šumy, nicméně jsou výpočetně velmi náročné.
V poslední době se vracejí do obliby umělé neuronové sítě. Bylo ukázáno, že je možné efektivně odstraňovat šum v řečových nahrávkách pomocí hlubokých neuronových sítí s využitím předtrénování pomocí skládání RBM a algoritmu kontrastivní divergence [3].
Tato práce se snaží prozkoumat možnost využití hlubokých neuronových sítí pro stejný účel (tzv. autoenkodéry pro odstranění šumu) bez složitého procesu předtrénování společně s vlivem rozmanitosti trénovací sady na schopnosti robustně odstraňovat šum.

Cíle práce

1. Natrénovat 2 různě zaměřené autoenkodéry pro odstranění šumu
2. Porovnat výsledky obecně (3 druhy šumu) a úzce (1 druh šumu) zaměřeného autoenkodéru
3. Porovnat výsledky autoenkodérů s konvenčně používanými metodami

Metodika

Trénování

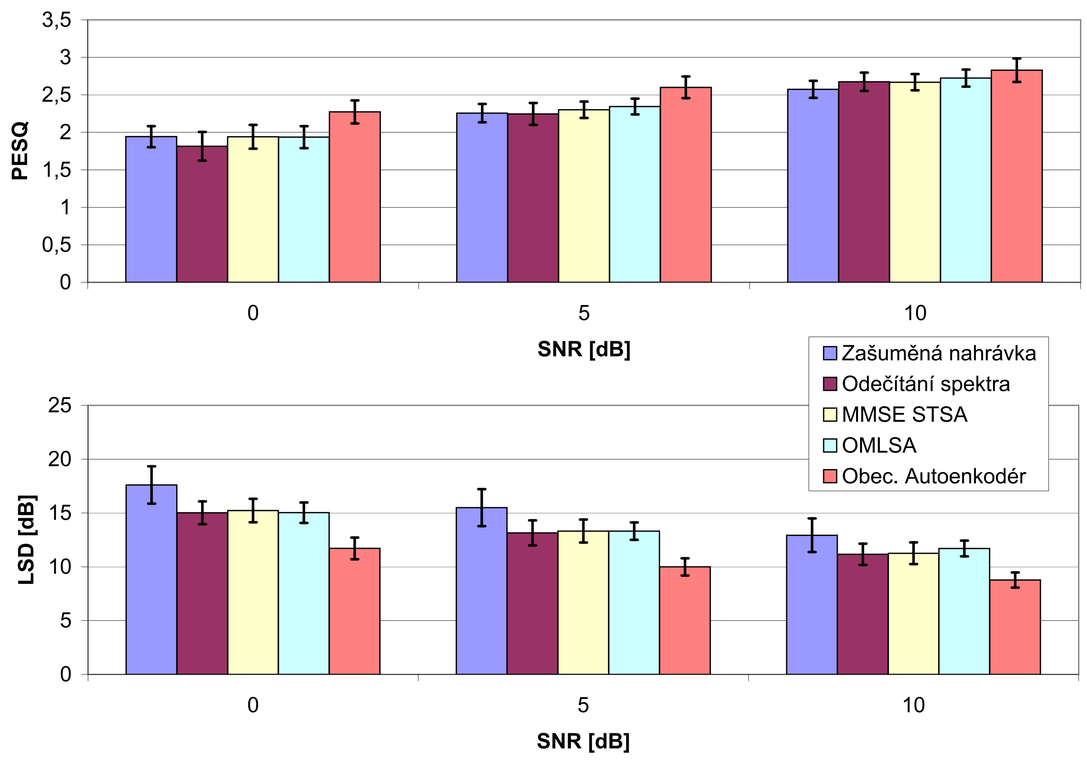
- Trénovací sada
  - 6000 nahrávek čisté řeči TIMIT [4], šumy CHiME Challenge [5] (Kavárna pro úzce zaměřený autoenkodér, Kavárna, Lidé a Autobus pro obecně zaměřený autoenkodér), 8 kHz vzorkovací frekvence, SNR = 0, 5 a 10 dB
  - 13 hodin zašuměné řeči pro úzce zaměřený autoenkodér, 26 hodin pro obecně zaměřený autoenkodér
  - STFT (Hammingovo okénko, 256 vzorků, překryv 128 vzorků), logaritmické výkonové spektrum, kontext +- 5 rámců
- Neuronová síť
  - 3 skryté vrstvy, každá 1024 neuronů, aktivační funkce tanh
  - vstup - zašuměná řeč, výstup - čistá řeč
  - SCG trénovací algoritmus, 500 rámců mini-dávka, 20 epoch

Testování

- 256 nahrávek připravených stejně jako trénovací sada + šum Ulice, SNR = -3, 0, 2, 5, 7 a 10 dB
- Metriky
  - PESQ (poslechová kvalita, srozumitelnost; od 1 (nejhorší) do 4,5 (nejlepší))
  - LSD (vzdálenost v logaritmickém spektru, čím menší tím lepší)
- Robustnost v neznámých jazykových podmínkách - 100 nahrávek české řeči

Výsledky

- Obecně vs. úzce zaměřený autoenkodér
  - Obecně zaměřený autoenkodér dosahuje lepších výsledků
- Obecně zaměřený autoenkodér vs. konvenční metody
  - PESQ - Autoenkodér lepší pro nestacionární šumy přes všechna SNR (viz obr.), pro stacionární šumy při nízkých úrovních SNR (< 5 dB)
  - LSD - Autoenkodér lepší ve všech případech
  - Doba zpracování nahrávky - o řád nižší než OMLSA, srovnatelná s ostatními metodami
  - Nejméně šumových residuí a hudebních artefaktů
- Vysoká konzistence výsledků autoenkodéru pro známé i neznámé podmínky - velmi robustní metoda
- Nevýhody
  - Zkreslení řeči při vysokém SNR
  - Vysoká výpočetní náročnost při trénování



Obrázek 1: Střední hodnota a standardní odchylka PESQ a LSD testovaných metod, šum Kavárna.

Závěr

Bylo ukázáno, že využití hlubokých neuronových sítí bez předtrénování pro účely robustního zlepšování řečových nahrávek může být validním přístupem, obvzláště při využití rozmanité trénovací sady.
Schopnosti zlepšování nahrávek překonávají schopnosti konvenčně používaných metod v obtížných podmínkách (nestacionární šum, nízké úrovně SNR).
Zlepšení by mohlo být dosaženo přidáním čistých nahrávek do trénovací sady (pro zmírnění zkreslení čisté řeči) a zvětšením trénovací sady o další druhy šumů a úrovně SNR.

Reference

[1] KAWAMURA, Arata, et al. Single Channel Speech Enhancement Techniques in Spectral Domain. *ISRN Mechanical Engineering*. 2012 [cit. 18.12.2015]. doi:10.5402/2012/919234

[2] COHEN, Israel a BERDUGO, Baruch. Speech enhancement for non-stationary noise environments. *Signal processing*. 2001, 81(11), s. 2403-2418. ISSN 0165-1684.

[3] XU, Yong, et al. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal Processing Letters*. 2014, 21(1), s. 65-68. ISSN 1070-9908.

[4] GAROFOLO, John S., et al. DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon Technical Report N 93*. 1993.

[5] BARKER, John, et al. The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines. *IEEE Automatic Speech Recognition and Understanding Workshop*. 2015.

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