Rozšířený Abstrakt

Zlepšování řečových nahrávek pořízených v reálném prostředí

Bc. Tomáš Kounovský, Ing. Jiří Málek, Ph.D.

Abstrakt

Tato práce se zabývá problematikou zlepšování řečových nahrávek reálného charakteru odstraněním šumu v logaritmické výkonové spektrální oblasti pomocí hlubokých neuronových sítí. Dvě sítě jsou natrénovány na odlišně sestavených trénovacích sadách tak, aby se naučily vztah mezi zašuměnou a čistou řečí. Jejich schopnosti odstranění šumu z nahrávek jsou otestovány ve známých i neznámých podmínkách a porovnány s některými konvenčními metodami. Výsledky ukazují, že síť natrénovaná na databázi obsahující více druhů šumu je robustnější při zlepšování řeči degradované neznámým druhem šumu. Zároveň je tato síť schopna konkurovat konvenčním metodám při odstraňování šumu stacionárního charakteru a překonat je při odstraňování šumu silně nestacionárního.

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, dále jen jako autoenkodéry) bez složitého procesu předtrénování společně s vlivem rozmanitosti šumových dat při trénování na výsledky.

Experiment a metody

Pro trénování autoenkodérů bylo 6000 čistých řečových nahrávek z databáze TIMIT [4] degradováno šumy z databáze CHiME challenge [5]. Pro úzce zaměřený autoenkodér byl vybrán šum Kavárna, pro obecný autoenkodér šumy Kavárna, Lidé a Autobus. Všechny nahrávky byly degradovány na úrovně SNR = 0, 5 a 10 dB. Výsledné trénovací sady dosahují délek 13 hodin pro úzce zaměřený autoenkodér a 26 hodin pro obecný autoenkodér. Všechny nahrávky byly decimovány na vzorkovací frekvenci 8 kHz.

Pro účely testování bylo zbývajících 256 nahrávek zašuměno všemi čtyřmi druhy šumu (Kavárna, Lidé, Autobus a Ulice) na úrovních SNR = -3, 0, 2, 5, 7, a 10 dB. Pro účely otestování robustnosti při zlepšování nahrávek s neznámým jazykem řečníka byla sestavena a nahrána databáze 100 českých vět, které byly zašuměny stejným způsobem jako ostatní testovací nahrávky.

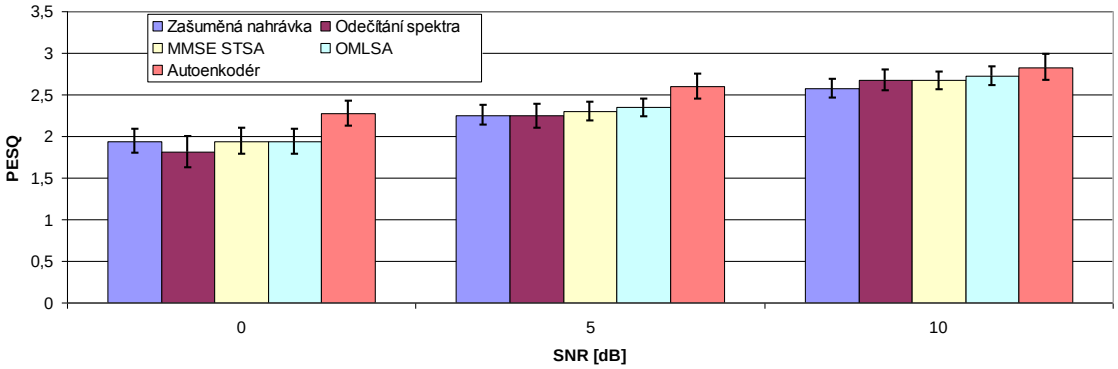
Samotné autoenkodéry využívají dopředné topologie a obsahují 3 skryté vrstvy s 1024 neurony v každé vrstvě. Aktivační funkcí je hyperbolická tangenta ve skrytých vrstvách a lineární funkce ve vrstvě výstupní. Jako vstup autoenkodéry přijímají zašuměné řečové signály transformované pomocí STFT (hammingovo okénko délky 256 vzorků, překryv 128 vzorků) do normalizované logaritmické výkonové časově-spektrální oblasti. Ke každému rámci je přidáno 5 rámců z obou stran pro obsažení kontextové informace. Výstup sítě tvoří rámce odhadnutého zlepšeného spektra. Pro rekonstrukci signálu ve fázi testování je využito fázové spektrum původní zašuměné nahrávky.

Výsledky a diskuze

Nahrávky zlepšené všemi uvedenými metodami byly porovnány s čistými nahrávkami pomocí metrik PESQ (označuje srozumitelnost a poslechovou kvalitu řeči na stupnici od 1 (nejhorší) do 4,5 (nejlepší)) a LSD (označuje vzdálenost mezi signály v logaritmickém spektru, čím menší tím lepší).

Obecný autoenkodér dosahoval konzistentně lepších výsledků než autoenkodér úzce zaměřený. Obecný autoenkodér také překonával konvenční metody ve všech testovaných případech ve smyslu přiblížení signálu ideální (čisté) nahrávce. Ve smyslu poslechové kvality překonával obecný autoenkodér konvenční metody pro silně dynamické druhy šumu (např. šum Kavárna) na všech úrovních SNR, viz obr. 1, a stacionární druhy šumu na nízkých úrovních SNR (< 5 dB). Nahrávky zlepšené pomocí autoenkodéru obsahovaly nejméně šumových residuí. Doba potřebná ke zpracování nahrávky je o řád nižší než doba potřebná pro metodu OMLSA a srovnatelná s ostatními metodami.

Autoenkodéry natrénované v této práci dosahují konzistentních výsledků pro známé i neznámé podmínky, což poukazuje i přes závislost na trénovacích datech na vysokou robustnost použité metody. Mezi nevýhody autoenkodérů lze zařadit vysoké výpočetní nároky procesu trénování neuronové sítě a lehké zkreslení čistého řečového signálu při zpracování.



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

Závěr

Bylo ukázáno, že využití hlubokých neuronových sítí bez předtrénování pro účely robustního zlepšování řečových nahrávek může být validním přístupem. Schopnosti zlepšování nahrávek jsou srovnatelné s konvenčně používanými metodami při vysokém SNR a stacionárním šumu, přičemž pro nízká SNR a nestacionární šumy dosahuje navrhovaná metoda lepších výsledků. Autoenkodér trénovaný na větší množině druhů šumu dosahuje lepších výsledků než autoenkodér úzce zaměřený.

Výsledky autoenkodérů by mohly být zlepšeny hlavně přidáním čistých nahrávek do trénovací sady k odstranění efektu zkreslení čistého řečového signálu. Zvětšení trénovací sady o další druhy šumů s více úrovněmi SNR by také mohlo zlepšit výsledky navrhované metody.

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. *IEEE Automatic Speech Recognition and Understanding Workshop*. 2015.