LipsID

Ing. Miroslav Hlaváč[1]

1 Úvod

Úloha automatického odezírání ze rtů je v současnosti stále řešeným tématem, kde výsledky strojového rozpoznávání nedosahují zdaleka takové kvality jako odezírání řeči člověkem. Tato úloha je významná zejména z pohledu rozpoznávání mluveného slova v rušném prostředí, u postižených řečníků, při tichém diktování, atd.

State-of-the-art metody používaní k odezírání ze rtů primárně neuronové sítě. Na uzavřených datasetech s omezeným množstvím slov a přesně danou syntaxí dosahují současné metody přesnosti kolem 90%. Na otevřených datasetech vytvořených například z vysílání televizních zpráv dosahuje tato přesnost hodnoty kolem 60%. Tato oblast stále obsahuje velké množství prostoru ke zlepšení a jednou z možností je adaptace neuronové sítě na konkrétního řečníka. Tato práce se zabývá prvním krokem této adaptace a to získáním dodatečné informace a konkrétním řečníkovy a vytvořením příznakového vektoru reprezentující toho řečníka - LipsID.

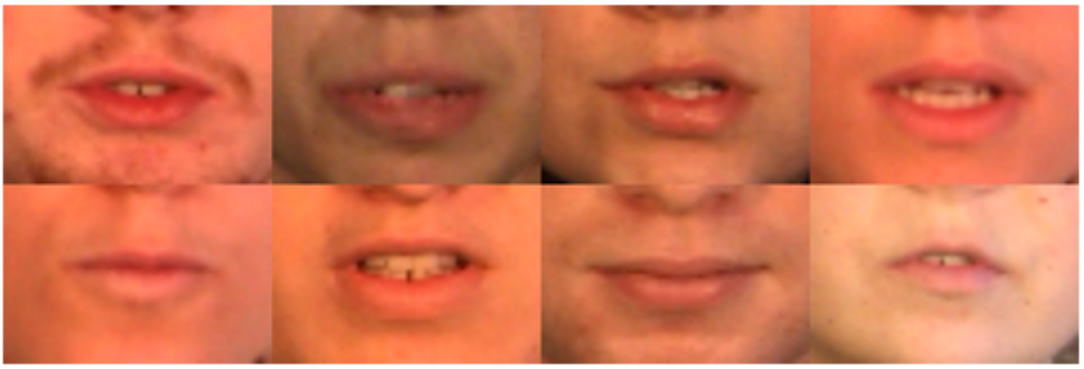
2 LipsID

Navržená reprezentace je inspirována metodou iVector popsanou v práci Saon at al. (2013), původně navrženou k identifikaci řečníka v úloze automatického rozpoznávání řeči. Jak se později ukázalo, tento příznakový vektor dokáže vylepšit i samotné rozpoznávání a díky své nízké dimensionalitě je též vhodný ke zpracování v reálném čase. LipsID reprezentace tedy staví na principu klasifikace řečníka pomocí jednotlivých snímků z videa obsahujícího řeč. Tato klasifikace je získána pomocí speciálně navržené neuronové sítě, které se učí identifikovat řečníka na základě snímků obsahujících pouze oblast rtů.

Počáteční experimenty byly navrženy tak, že se síť naučila identifikovat řečníky podle samostatných snímků s využitím 2D konvolučních vrstev. Toto zpracování dosáhlo úspěšnosti $99,1\%$. Jelikož se při rozpoznávání mluvené řeči z videa používají jako vstupní data sekvence obrázků, směřovala další snaha k vytvoření sítě, která bude provádět výše zmíněnou klasifikaci na základě těchto dat. Pomocí nově používaných 3D konvolucí publikovaných v Ji at al. (2013) tedy byla vytvořena síť klasifikující řečníky ze sekvencí čítajících 15 po sobě jdoucích snímků. Tyto snímky byly nejprve převedeny do odstínů šedi a poté spojeny je jedné matice o velikosti šířka $\times$ výška $\times$ délka sekvence. Výstupem sítí je v tomto případě softmax klasifikace do tříd, které odpovídají jednotlivým řečníkům. Parametrizační vrstva se nachází jako předposlední a LipsID tedy dostaneme odečtení výstupu předposlední vrstvy konkrétní sítě.

Trénovací data pro tuto síť byla získána z datasetu UWB-HSCAVC vytvořeného na Západočeské univerzitě v Plzni skupinou Císař at al. (2005). Z datasetu byla vybrána data 62 řečníků, která obsahují 200 nahraných vět. Z těchto vět je prvních 50 společných pro všechny řečníky a ostatní věty se liší. Pro trénování sítí byly využity snímky z vět, které jsou pro všechny

[1] student doktorského studijního programu Aplikované vědy a informatika, obor Kybernetika, specializace Počítačové vidění, e-mail: mhlavac@kky.zcu.cz



Obrázek 1: Trénovací data pro neuronové sítě.

řečníky společné a zbylé věty byly použity jako vývojová a testovací data.

3 Závěr

V této práci byla navržena metoda pro vytvoření parametrizace rtů konkrétních řečníků pro účely adaptace současných metod rozpoznávání vizuální složky řeči - odezírání ze rtů. Dalším úkolem bude implementovat navrženou metodu do stávajících sítí, například LipNet (Assael at al. (2016))), WLAS (Chung at al. (2017)) a LCANet (Xu at al. (2018)).

Poděkování

Přízpěvek byl podpořen grantovým projektem SVK1-2018-024. Tato práce vznikla za podpory projektů CERIT Scientific Cloud (LM2015085) a CESNET (LM2015042) financovaných z programu MŠMT Projekty velkých infrastruktur pro VaVaI.

Literatura

Saon, G., Soltau, H., Nahamoo, D., Picheny, M. (2013) Speaker adaptation of neural network acoustic models using i-vectors. *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 55–59

Císař, P., Železný, M., Krňoul, Z., Kanis, J., Zelinka, J., Müller, L. (2005) Design and recording of Czech speech corpus for audio-visual continuous speech recognition. *Proceedings of the Auditory-Visual Speech Processing International Conference 2005*, pp. 1–4

Ji, S., Xu, W., Yang, M., Yu, K. (2013) 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 221–231

Chung, J. S., Senior, A., Vinyals, O., Zisserman, A. (2017) Lip reading sentences in the wild. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, pp. 3444–3453

Assael, Yannis M., Shillingford, B., Whiteson, S., de Freitas, N. (2016) LipNet: End-to-End Sentence-level Lipreading. *eprint arXiv:1611.01599*, 2016

Xu, K., Li, D., Cassimatis, N., Wang, X. (2018) LCANet: End-to-End Lipreading with Cascaded Attention-CTC. *eprint arXiv:1803.04988*, 2018