BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF INFORMATION TECHNOLOGY

Excel@FIT 2023
http://excel.fit.vutbr.cz

Kombinace informací více-kanálových nahrávek

Jan Procházka*

Abstrakt
Tento projekt se zabývá analýzou a porovnáním kombinací informací více-kanálových řečových dat pro úlohu verifikace mluvčího. Byly zvoleny tři úrovně/reprezentace pro fúzi dat: kombinace na úrovni **signálu**, **embeddingu** a **skóre**. Na úrovni signálu jsou implementovány prostorové filtry (algoritmy formování svazku – beamforming), konkrétně Delay and Sum a MVDR (celkem tři varianty: výpočet šumové kovarianční matice kardinálním sinem, ze vstupu a ze separované nahrávky na řeč a šum pomocí neuronové sítě **DPRNN**). Řečové nahrávky slouží jako vstup do neuronové sítě (architektura **ECAPA-TDNN**), která extrahuje embeddingy, vektorovou reprezentaci mluvčího. Vektory jsou dále zpracovány v modulu cosinové podobnosti, jehož výsledkem jsou skóre, reálná čísla. Nejlepšího relativního zlepšení v odhadu odpovědi na úlohu verifikace proti jedno-kanálovým nahrávkám dosahuje fúze na úrovni skóre (**až 70 %**), nejkonzistentnější výsledky pro různé podmínky pořizování nahrávek poskytuje fúze na úrovni embeddingu.

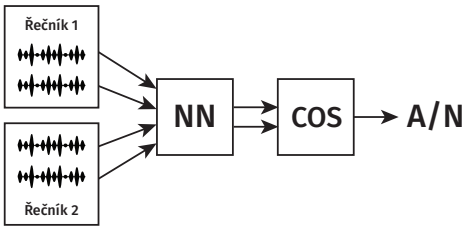
*xproch0g@vutbr.cz, *Faculty of Information Technology, Brno University of Technology*

1. Úvod

Tato práce má za cíl porovnat **metody pro zpracování a kombinaci dat více-kanálových nahrávek** (nahrávky z mikrofonního pole) řeči pro úlohu vzdálené (mikrofon se nachází daleko od zdroje řeči) **verifikace mluvčího**. Vzdálenému rozpoznání řečníka se věnujeme z důvodu nemožnosti použít blízký mikrofon pro danou aplikaci, nebo pouze z důvodu pohodlnosti pro uživatele. Očekáváme výrazné zlepšení přesnosti verifikace pro promluvu nahranou více mikrofony, relativně k nahrávce pořízené jedním mikrofonem.

Při použití **vzdáleného mikrofonu** do nahrávky se do nahrávky dostává velké množství **reverberace**, nežádoucího **šumu** z místnosti (např. lednice, klimatizace, projíždějící auto..) a snižuje se **dynamický rozsah**. Snížení poměru SNR mezi původním řečovým signálem a ostatními zvukovými signály v místnosti vede ke zhoršení vlastností nahrávky, je méně srozumitelná a náročná na poslech. Špatně se odhaduje, kdo na nahrávce mluví, systém snižuje **jistotu odhadu** na otázku verifikace mluvčího.

Na Obr. 1 je vidět základní schéma zpracování řečových nahrávek. Srdcem celého systému je **neuronová síť** s architekturou **ECAPA-TDNN** [1], která ze vstupního signálu extrahuje embeddingy, vektorovou formu charakterizace mluvčího. Pro obě vstupní



Obrázek 1. Schéma vyhodnocování

nahrávky extrahujeme tyto embeddingy, které slouží jako vstup funkce cosinové podobnosti. Výstup cosinové podobnosti je skóre, které určuje podobnost vstupních vektorů. Prostou metodou prahování skóre získáme odpověď na otázku verifikace mluvčího.

Potvrdily se počáteční předpoklady výrazného zlepšení pro úlohu verifikace mluvčího při použití více informací, které poskytují více-kanálové nahrávky oproti jedno-kanálovým.

- Až 70,2 % pro fúzi na úrovni skóre.
- Až 63,2 % a konzistentnější výsledky pro fúzi na úrovni embeddingu.
- Až 11 % pro fúzi na úrovni signálu.

Pro účely vyhodnocení verifikace je použit datový korpus MultiSV [2], který obsahuje nahrávky se čtyřmi kanály a disponuje rozdělením podle podmínek pro pořízení nahrávky (rozměry nahrávací místnosti, zdroje šumu (hudba, televize, či řeč v pozadí) a vzdálenosti mikrofonů od zdroje řečové nahrávky).

2. Fúze na úrovni skóre

Skóre je výsledek *cosinové podobnosti* vektorů reprezentujících řečníka, které nám poskytuje neuronová síť. Pro zkoumané dvě nahrávky dostaneme skóre, pomocí kterého pak **prahováním získáme odpověď** na úlohu verifikace mluvčího. Pro více-kanálová data získáváme unikátní skóre pro každé porovnání kanálů. Skóre jsou reprezentovány pomocí reálných čísel, lze nad nimi snadno provádět jakékoliv matematické operace. Výsledky verifikace pro metodu průměrování (`avg`), nebo výběru maxima (`max`) ze získaných skóre dosahují výborných hodnot (viz Tabulka 1).

Následující tabulky obsahují hodnoty **EER [%]** a **relativní zlepšení** oproti jedno-kanálovým datům [%].

Tabulka 1. Vyhodnocení MultiSV s „retransmitovanými" nahrávkami s fúzí na úrovni skóre

Retransmit	eval1	%
CE (max)	1.44	70.27
SRE (max)	1.72	65.47
MRE (max)	2.22	51.64
MRE_H (max)	4.51	37.85

3. Fúze na úrovni embeddingu

Embeddingy, vektorová **charakterizace řečníka**, konkrétně jeho celé **hlasové ústrojí**. Pro každou vstupní nahrávku extrahuje neuronová síť jednu reprezentaci mluvčího. V případě porovnání všech kanálů jedné nahrávky dostáváme více takových reprezentací mluvčího, které jsou dále zpracovány. Varianta `avg` sčítá tyto vektory se stejnou váhou a normalizuje na jednotkovou kružnici. Varianta (`vah`) přiděluje jednotlivým vektorům vlastní váhy. Děje se tak na základě odhadu odstupu signálu od šumu (SNR) pro jednotlivé nahrávky. Z více implementovaných metod se jako nejpřesnější jeví algoritmus Wada[1] viz Tabulka 2.

Tabulka 2. Vyhodnocení MultiSV s „retransmitovanými" nahrávkami s fúzí na úrovni embeddingů

Retransmit	eval1	%
CE (vah)	1.78	63.20
SRE (vah)	2.06	58.59
MRE (vah)	2.11	54.07
MRE_H (vah)	4.31	40.64

[1] https://gist.github.com/johnmeade/d8d2c67b87cda95cd253f55c21387e75

4. Fúze na úrovni signálu

4.1 Delay and Sum

Algoritmus Delay and Sum **zohledňuje časové zpoždění**, které vychází z propagace rovinné vlny prostorem. Zpožďuje jednotlivé kanály nahrávky, které představují výstupy z mikrofonů rozmístěných po místnosti, tedy v **různých vzdálenostech** od zdroje zkoumaného řečového signálu. K odhadu posunu je použita křížová korelace přes frekvenční spektrum GCC-PHAT[2]. Nejlepších výsledků (viz Tabulka 3) dosahuje varianta (`norm`), která normalizuje nahrávky před aplikací DaS algoritmu.

Tabulka 3. Vyhodnocení MultiSV s „retransmitovanými" nahrávkami s fúzí na úrovni signálu (Delay and Sum s omezením)

Retransmit	eval1	%
CE (norm)	4.39	9.36
SRE (norm)	4.82	3.24
MRE (norm)	4.42	3.77
MRE_H (norm)	6.70	7.63

4.2 MVDR

MVDR adaptivně **potlačuje prostorově korelovaný šum**, čehož lze dosáhnout změnou váhového vektoru beamformeru tak, aby se **minimalizoval rozptyl šumu a interference** na výstupu algoritmu s ohledem na podmínku zkreslení (distortionless response), tj. výstup beamformeru bez šumu není ani zesílen, ani zeslaben. Z více variant implementace nejlepších hodnot dosahuje metoda se separací řeči a šumu pomocí neuronové sítě DPRNN [3] viz Tabulka 4.

Tabulka 4. Vyhodnocení MultiSV s „retransmitovanými" nahrávkami s fúzí na úrovni signálu (MVDR)

Retransmit	eval1	%
CE (dprnn)	4.31	11.02
SRE (dprnn)	4.66	6.49
MRE (dprnn)	4.18	9.19
MRE_H (dprnn)	6.81	6.12

Poděkování

Rád bych poděkoval Ing. Ladislavu Mošnerovi za jeho neúnavnou podporu, vedení a poskytování četných konzultací v průběhu vytváření této práce.

[2] http://www.xavieranguera.com/phdthesis/node92.html

Literatura

[1] Kris Demuynck Brecht Desplanques, Jenthe Thienpondt. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification, Aug 2020. Dostupné z https://arxiv.org/pdf/2005.07143.pdf.

[2] Ladislav Mošner, Oldřich Plchot, Lukáš Burget, and Jan Černocký. Multisv: Dataset for far-field multi-channel speaker verification, 2021.

[3] Takuya Yoshiokak Yi Luo, Zhuo Chen. Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation, Mar 2020. Dostupné z https://arxiv.org/pdf/1910.06379.pdf.