ADAPTATION OF SPEAKER RECOGNITION SYSTEMS

Ondřej Novotný
Master Degree Programme (2), FIT BUT
E-mail: xnovot96@stud.fit.vutbr.cz

Supervised by: Oldřich Plchot
E-mail: iplchot@fit.vutbr.cz

Abstract: In this paper we propose techniques for adaptation of speaker recognition systems. The aim of this work is to create adaptation for Probabilistic Linear Discriminant Analysis (PLDA). Special attention is given to unsupervised adaptation. Our test shows appropriate clustering techniques for estimation of speaker identity and estimation of the number of speakers in adaptation dataset.

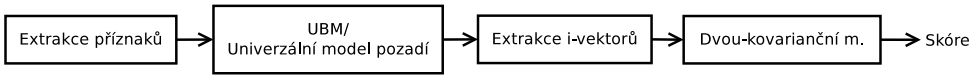
Keywords: Recognition, speaker recognition, clustering, adaptation.

1 ÚVOD

Systémy na rozpoznávání mluvčích jsou dnes široce využívány v mnoha oblastech lidského života. Své uplatnění našly v civilní, policejní i vojenské sféře. Využívají se v aplikacích pro autorizaci mluvčích, či forenzní dokazování. Pro vytvoření rozpoznávacího systému je potřeba nemalého množství anotovaných nahrávek. Přesto pak nemusí systém dosahovat požadované přesnosti. Je to způsobené rozdílnou variabilitou (jazyk, akustické podmínky atd.) trénovacích a rozpoznávaných nahrávek. Proto je v současnosti věnováno velké úsilí vývoji metod pro přizpůsobení systému novým podmínkám.

2 SYSTÉM PRO ROZPOZNÁVÁNÍ MLUVČÍCH

Pro tuto práci byl použit rozpoznávací systém vyvinut výzkumnou skupinou **Speech@Fit** [1]. Zaměříme



Obrázek 1: Základní struktura systému.

se na adaptaci Dvou-kovariančního modelu. Vstupem tohoto kroku jsou pro jednotlivé nahrávky příznakové i-vektory (příznakové vektory konstantní velikosti, bez ohledu na délku nahrávky, viz [3]). Výstupem je skóre odpovídající podobnosti dvojici i-vektorů (dále jen podobnost). Dvou-kovarianční model se snaží postihnout a rozlišit dvě variability dat, variabilitu mluvčích a variabilitu kanálu. V tomto modelu se předpokládá variabilita v gaussovském rozložení. Variability mluvčích a kanálu jsou poté reprezentovány kovariančními maticemi Σ_{ac} a Σ_{wc}.

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}; \mu, \Sigma_{ac}) \tag{1}$$

$$p(\phi|\hat{\mathbf{y}}) = \mathcal{N}(\phi; \hat{\mathbf{y}}, \Sigma_{wc}) \tag{2}$$

Identita mluvčího je reprezentována skrytou proměnnou $\mathbf{y}$, jejíž rozložení odpovídá (1). Následně rozložení i-vektorů ϕ známého mluvčího, který je reprezentován vektorem $\hat{\mathbf{y}}$ odpovídá (2). Proměnná μ reprezentuje střední hodnotu identit mluvčích.

3 ADAPTACE BEZ UČITELE

V tomto případě nám stačí dostatek cílových nahrávek bez nutnosti k nim příslušných anotací. Popíšeme si zde dva způsoby jak adaptaci provést. První možností je **Shlukování a přetrénování**.V tomto způsobu je neadaptovaný model využit k výpočtu skóre vzájemné podobnosti adaptačních i-vektorů. Poté se pomocí shlukovacího algoritmu provede odhad identit mluvčích. Následně je natrénován nový model pomocí nové trénovací sady tvořené původními trénovacími vektory a adaptačními vektory s nově odhadnutou identitou mluvčích. Využití pouze nových adaptačních dat pro trénování není vhodné. Nemáme jich zpravidla dostatek a odhad identit není dokonalý. Pro zpřesnění této techniky je vhodné tento adaptační postup opakovat. Dochází tak k postupné úpravě variability kanálu i mluvčích vlivem postupného zpřesňování odhadu identifikace mluvčích.

Druhou možností je **Shlukování a adapatace**. Opět využijeme neadaptovaného model k získání podobnosti adaptačních i-vektrorů následovaného shlukovacím algoritmem pro odhad identit mluvčích. Nyní ovšem nedochází k natrénování nového modelu na kombinované trénovací sadě, ale pouze na adaptačních nahrávkách. Ke kombinaci variabilit dochází na úrovni modelů (kovariančních matic, rovnice 3, kde *out* značí neadaptované model a *in* nově natrénované) pomocí váženého průměru.

$$(\Sigma_{ac}, \Sigma_{wc}) = arg \max \alpha(\Sigma_{ac}, \Sigma_{wc})_{in} + (1-\alpha)(\Sigma_{ac}, \Sigma_{wc})_{out} \tag{3}$$

Tímto způsobem pak lze adaptovat pouze určitou variabilitu, například v případě, že se nezmění mluvčí, ale pouze akustické podmínky.

4 ODHAD IDENTIT MLUVČÍHO A POČTU MLUVČÍCH V DATOVÉ SADĚ

Pro odhad identity mluvčích jsou využívány shlukovací algoritmy. Vhodnými algoritmy jsou algoritmy pro shlukováni grafových struktur. Uvědomme si, že při užití skóre máme k dispozici pouze vzájemnou podobnost i-vektorů. Některé shlukovací algoritmy pracují se vzdáleností (euklidova, kosínova) je proto nutné podobnost i-vektorů (skóre z rozpoznávacího systému) přepočítat na virtuální vzdálenost. Toho lze docílit vztahem $d_i = \max(S) - s_i$, kde S odpovídá množině jednotlivých skóre mezi všemi kombinacemi i-vektorů v adaptační sadě, s_i jednotlivým hodnotám skóre a d_i příslušné hodnotě virtuální vzdálenosti. Vhodnými algoritmy tedy jsou: Aglomerativní hierarchické shlukování nebo Spojování komponent ([2]).

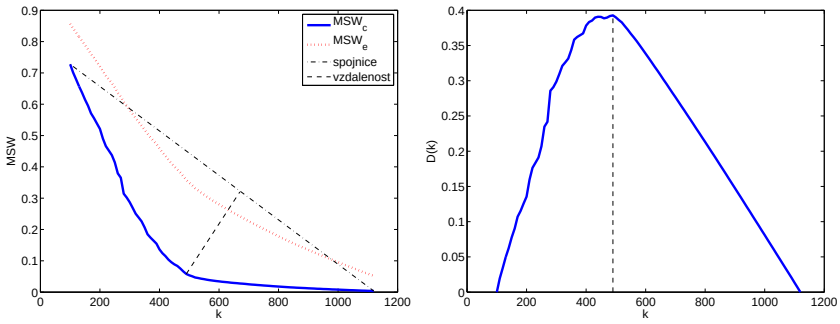
Tyto algoritmy vyžadují kromě matice vzájemné podobnosti (virtuální vzdálenosti) vektorů i informaci o počtu shluků, do kterých mají být vstupní data roztříděna. Existuje velké množství algoritmů pro odhad ideálního počtu shluků. Většina algoritmů pro odhad využívá vývoj střední čtvercové vzdálenosti dat a středu shluků (MSW-*mean square within cluster*). Jedná se o hledání příznačného ohybu této křivky u optimálního počtu shluků (obr. 2, vlevo). Zde je nutné dát si pozor při rozlišování virtuální vzdálenosti (vychází ze skóre rozpoznávacího systému) a skutečné vzdálenosti (euklidova, kosínova). Pro samotný odhad identit, je použíta virtuální vzdálenost. Pro hodnoty křivky středních vzdáleností se používá skutečná vzdálenost (vycházíme ze vzdálenosti vektoru a středu shluků a je tedy použita poloha bodu určeného vektorem v prostoru). Primárně používanou vzdáleností dvou bodů v prostoru je euklidovská vzdálenost, v našich experimentech se ukázala být vhodnější kosínova vzdálenost (pořád se, ale počítá průměr čtverců těchto vzdáleností).

Většina algoritmů je založena na hledání největší změny směrnice přímky mezi jednotlivými body křivky se čtvercovou vzdáleností. Vzhledem k poměru dat (velké množství tříd, s malým množstvím vektorů) se nejlépe uplatní algoritmus zkoumající vývoj křivky v globálním měřítku. Optimální počet je pak určen podle bodu jenž má největší vzdálenost od spojnice koncových bodů křivky (obr. 2, vpravo).

5 VYHODNOCENÍ

V tabulce 1, lze vidět srovnání zmíněných adaptačních technik s neadaptovaným systémem.

Zmíněných výsledků u obou systémů bylo dosaženo nastavením hodnot parametrů, na hodnoty jenž se při vývoji ukázali jako nejvhodnější ($\alpha = 0.3$). V případě metody Shlukování a přetrénování je získané



Obrázek 2: (Vlevo) Vývoj střední čtvercové euklidové a kosínové vzdálenosti při odhadu počtu mluvčích (osa x). Čerchovaná čára naznačuje spojnici mezi koncovými body, čárkovaná poté měřenou vzdálenost. (Vpravo) Vzdálenost bodů MSW_c křivky od spojnice koncových bodů.

Metoda	EER (Equal Error Rate)
Shlukování a přetrénování	4.6%
Shlukování a adaptace	6.4%
Neadaptovaný systém	6.6%

Tabulka 1: Srovnání adaptačních algoritmů.

zlepšení dosaženo zanesením nové variability do trénovacích dat. Vzhledem ke spojení trénovacích sad, ještě před trénováním nového systému je trénování ve výsledku méně náchylný na chybný odhad identit v adaptační sadě. Celková variabilita všech dat je vhodně rozšířena a dochází tedy ke zlepšení. Nemáme, ovšem sloučení obou systému dostatečně pod kontrolou. Podstatně horší výsledek druhého způsobu adaptace je zřejmě zapříčiněn větší náchylnosti této metody na kvalitu odhadu počtu mluvčích a bude nutné navrhnout doplňující techniky pro zkvalitnění těchto výsledků (např. filtrování výsledků shlukování, adaptování variability mluvčích a kanálu odděleně).

Pro trénování a evaluace rozpoznávacího systému je použito několik datových korpusů. Jako zdroj dat pro neadaptovaný systém slouží korpus **Switchboard** (Fáze 1, Fáze 2 a Cellular). Jako adaptační data jsou použité nahrávky z korpusů **NIST** z let 2004, 2005, 2006 a 2008. Jako evaluační sada poté slouží korpus **NIST** 2010.

6 ZÁVĚR

V této práci jsme popsali způsoby adaptace systémů na rozpoznávání mluvčího a možnosti při nasazení takového systému v praxi. Navrhli jsme odhad počtu mluvčích v adaptační datové sadě a možnosti odhadu identit mluvčích. Uvedené postupy lze dále aplikovat i při prvotním trénování systému v případech, kdy nemáme anotovanou dostatečnou zásobu nahrávek.

REFERENCE

[1] Niko, B.; Lukáš, B.; Patrick, K.; aj.: ABC System description for NIST SRE 2010. In Proc. NIST 2010 Speaker Recognition Evaluation, Brno, CZ, 2010, s. 1-20.

[2] Everitt, B. S.; Landau, S.; Leese, M.; aj.: Cluster Analysis. Kings College London,UK: John Wiley & Sons, Ltd, páté vydání, 2011, ISBN 978-0-470-97780-4.

[3] Dehak, N.; Kenny, P. J.; Dehak, R.; aj.: Front-End Factor Analysis for Speaker Verification. In IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, ročník 19, 2011, s. 19-41.