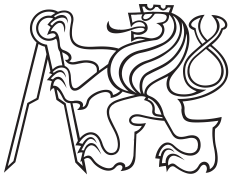
Diplomová práce



České
vysoké
učení technické
v Praze

F3

Fakulta elektrotechnická
Katedra radioelektroniky

Analýza nezávislých komponent pro řečové signály

Bc. Ondřej Brunner

Vedoucí: doc. Ing. Radoslav Bortel, Ph.D.
Obor: Audiovizuální technika a zpracování signálů
Studijní program: Elektronika a komunikace
Květen 2022

ZADÁNÍ DIPLOMOVÉ PRÁCE

I. OSOBNÍ A STUDIJNÍ ÚDAJE

Příjmení: **Brunner** Jméno: **Ondřej** Osobní číslo: **474390**

Fakulta/ústav: **Fakulta elektrotechnická**

Zadávající katedra/ústav: **Katedra radioelektroniky**

Studijní program: **Elektronika a komunikace**

Specializace: **Audiovizuální technika a zpracování signálů**

II. ÚDAJE K DIPLOMOVÉ PRÁCI

Název diplomové práce:

Analýza nezávislých komponent pro řečové signály

Název diplomové práce anglicky:

Independent Component Analysis for Speech Signals

Pokyny pro vypracování:

Obeznamte se s analýzou nezávislých komponent a s možnostmi jejího použití pro řečové signály.
Navrhněte algoritmy pro:
- spektrální rozklad řečových signálů,
- slepou separaci řečových signálů v jednotlivých pásmech (pomocí analýzy nezávislých komponent),
- zkombinování a rekonstrukci rozseparovaných signálů.

Algoritmy naimplementujte v prostředí Matlab a ověřte jejich funkci na reálných řečových signálech.
Vyhodnoťte kvalitu separace.

Seznam doporučené literatury:

[1] Hyvärinen, A., Karhunen, J., Oja, E.: Independent Component Analysis, John Wiley & Sons, 2004.
[2] Benesty, J., Chen, J., Huang, Y.: Microphone Signal Array Processing, Springer, 2008.
[3] Bingham, E., Hyvärinnen, A.: A fast fixed-point algorithm for independent component analysis of complex valued signals. International Journal of Neural Systems, Vol. 10, No. 1, pp 1-8, 2000.

Jméno a pracoviště vedoucí(ho) diplomové práce:

doc. Ing. Radoslav Bortel, Ph.D. katedra teorie obvodů FEL

Jméno a pracoviště druhé(ho) vedoucí(ho) nebo konzultanta(ky) diplomové práce:

Datum zadání diplomové práce: **26.01.2022** Termín odevzdání diplomové práce: _____

Platnost zadání diplomové práce: **30.09.2023**

doc. Ing. Radoslav Bortel, Ph.D. podpis vedoucí(ho) práce	doc. Ing. Stanislav Vítek, Ph.D. podpis vedoucí(ho) ústavu/katedry	prof. Mgr. Petr Páta, Ph.D. podpis děkana(ky)

III. PŘEVZETÍ ZADÁNÍ

Diplomant bere na vědomí, že je povinen vypracovat diplomovou práci samostatně, bez cizí pomoci, s výjimkou poskytnutých konzultací. Seznam použité literatury, jiných pramenů a jmen konzultantů je třeba uvést v diplomové práci.

. _____ Datum převzetí zadání

_____ Podpis studenta

Poděkování

Děkuji svému vedoucímu, doc. Ing. Radoslavu Bortelovi, Ph.D., za trpělivé vedení mé práce a předané zkušenosti.

Prohlášení

Prohlašuji, že jsem předloženou práci vypracoval samostatně a že jsem uvedl veškeré použité informační zdroje v souladu s Metodickým pokynem o dodržování etických principů při přípravě vysokoškolských závěrečných prací.

V Praze, 16. května 2022

Abstrakt

Práce se zabývá slepou separací řečových signálů pomocí analýzy nezávislých komponent. Nejprve jsou letmo představeny její základní metody, které ale nejsou pro separování řeči přímo použitelné. Práce proto pokračuje zavedením takzvaného konvolučního modelu směsí řečových signálů, který již využití metod analýzy nezávislých komponent po určitých změnách umožní. Práce následně popisuje tyto nutné změny. Jedná se o spektrální rozklad promluv do úzkých frekvenčních pásem, vyřešení potřeby separovat komplexní signály a odstranění známých nejednoznačností metod analýzy nezávislých komponent. Dvě metody, konkrétně Fast fixed-point ICA a Fast fixed-point IVA budou popsány podrobně a následně implementovány v prostředí Matlab. Na závěr budou vzniklé algoritmy prověřeny na simulovaných i reálných datech a bude vyhodnocena kvalita separace.

Klíčová slova: slepá separace řeči, analýza nezávislých komponent, analýza nezávislých vektorů

Vedoucí: doc. Ing. Radoslav Bortel, Ph.D.
ČVUT v Praze,
Fakulta elektrotechnická,
Katedra teorie obvodů,
Technická 2,
166 27 Praha 6

Abstract

This thesis deals with the blind source separation of speech signals using the Independent Component Analysis. Firstly, it briefly describes basic methods of Independent Component Analysis which are not directly applicable on speech separation. Therefore, the work continues with the introduction of so-called convolutive model of speech signal mixtures, which allows utilization of Independent Component Analysis methods after certain changes. Subsequently, the thesis describes those necessary changes. It is a spectral decomposition of speech into narrow frequency bands, solving the need to separate complex signals and elimination of known ambiguities of Independent Component Analysis methods. Two methods, namely Fast Fixed-point ICA and Fast Fixed-point IVA will be described in detail and afterwards implemented using Matlab. Finally, the resulting algorithms will be tested on simulated and real data, and the quality of separation will be evaluated.

Keywords: Speech blind source separation, Independent Component Analysis, Independent Vector Analysis

Title translation: Independent Component Analysis for Speech Signals

Obsah

Obrázky

Tabulky

Kapitola 1

Úvod

Cílem této práce je popsat, implementovat a otestovat vybrané metody separace řečových signálů pomocí analýzy nezávislých komponent.

1.1 Motivace

Separace obecně řeší úlohu, kdy existuje známá směs několika zdrojových signálu, které je od sebe žádoucí zpětně oddělit. Slepá separace (dále jako BSS) pak takovou úlohu řeší pouze za znalosti vzniklé směsi signálů.

Algoritmy BSS se v praxi využívají při řešení nejrůznějších problémů z biomedicíny, ekonomie, telekomunikací nebo zpracování obrazu. V těchto případech je často uvažován poměrně jednoduchý (takzvaně lineární) model směsi signálů umožňující rychlou a efektivní separaci. [1]

Směsi řečových signálů, které je cílem separovat, ale pomocí lineárního modelu popsat nelze [2]. Nutnost použití složitějších modelů při popisu směsi řečových signálů zapříčiní, že i samotný algoritmus BSS bude výpočetně náročnější, než je tomu u směsí signálů, které popsat lineárním modelem lze. Právě výpočetní složitost algoritmů slepě separujících řečové signály činí z celého tématu téma výzkumné, které je v průmyslových aplikacích využívané prozatím méně.

Problém, který může BSS řečových signálů řešit (a kterým se bude tato práce výhradně zabývat), je takzvaný „Cocktail Party Effect“ [1, s. 2]. Jedná se o modelovou situaci, kdy se nacházíme na večírku a zaznamenáváme zvuk v místnosti. V místnosti mluví více lidí naráz a na záznamu se objeví směs těchto promluv. Cílem BSS je tyto promluvy od sebe zpětně odseparovat, a to bez znalosti pozic jednotlivých řečníků.

Algoritmy BSS řečových signálů ale naleznou po určitých modifikacích uplatnění i při řešení dalších problémů. Příkladem je takzvaná extrakce, během které dochází k zvýraznění určité promluvy a potlačení okolního šumového pozadí. Dále tyto algoritmy naleznou využití jako předzpracování například pro rozpoznávání řeči od více mluvčích zároveň.

1.2 Typická řešení

Metod BSS (ať už řečových nebo jiných signálů) je značné množství. Nabízí se například využití technik pracujících na základě statistik druhého řádu (SOBI), mezi které patří třeba „analýza kanonické korelace“ (CCA) nebo takzvaný JADE pracující na principu aproximace diagonalizace vlastních matic. Se statistikami vyšších řádů pracují například metody známé jako „analýza nezávislých komponent“ (ICA), mezi něž lze řadit i jejich rozšíření známé jako „analýza nezávislých vektorů“ (IVA).

Pro úlohu BSS je důležité vědět, kolik různých směsí zdrojových signálů je k dispozici. V případě směsi řečových signálů se jedná o otázku, kolik mikrofonů bylo při nahrávání mluvčích použito. Existují jak algoritmy provádějící BSS řeči na základě dat z jediného mikrofonu [3], tak algoritmy využívající celých mikrofonových polí, které umožňují skloubit metody BSS s technikou „beamforming“, kdy se tvaruje příjmová charakteristika senzorového pole [4]. Metody ICA (i její rozšíření IVA) sice ve svém základu uvažují stejný počet senzorů, jako zdrojových signálů, po jejich jednoduché úpravě lze ale s výhodou využít senzorů více.

Metody BSS jsou bez dalších úprav určeny pro separování směsí, pro které platí lineární model. Řeč je ale v běžné místnosti mísena takzvaně konvolučně v závislosti na patřičných impulzních odezvách mezi zdroji a mikrofony. Jedním z možných řešení BSS řečových signálů je tedy nejprve zaznamenané signály rozložit do úzkých frekvenčních pásem, kde lze s určitou nepřesností uvažovat lineární model směsi, a v teprve v těchto pásmech známé metody BSS aplikovat.

1.3 Obsah práce

Tato práce se zabývá pouze vybranými metodami analýzy nezávislých komponent. Metody ICA jsou nejprve letmo představeny ve svých základních verzích včetně jejich klíčových vlastností, jako jsou nejednoznačnosti v pořadí a amplitudě. Dále je popsána aplikace těchto metod při separování řečových signálů. Konkrétně se jedná o rozklad zaznamenaných směsí do úzkých frekvenčních pásem, uzpůsobení metod ICA pro účely separace signálů v těchto pásmech, vyřešení jejich nejednoznačností a zpětnou syntézu odseparovaných signálů. Na závěr teoretické části jsou detailně popsány dvě vybrané metody, konkrétně „Fast fixed-point ICA“ a „Fast fixed-point IVA“.

Praktická část obsahuje implementaci obou metod v prostředí Matlab a vyhodnocení kvality separace. Za tím účelem budou popsány a uskutečněny směšovací experimenty v simulované místnosti i skutečná měření v místnosti reálné. Vzniklé směsi budou pomocí implementovaných metod odseparovány a následně bude vyhodnocena kvalita separace pomocí metrik SNR, kepstrální vzdálenost a na samotný závěr také jednoduchým poslechovým testem.

Kapitola 2

Separace lineární směsi signálů pomocí analýzy nezávislých komponent

2.1 Popis problému

Lineární směšování signálů je situace, kdy jsou neznámé reálné signály s_j váhovány neznámými reálnými konstantami, sčítány mezi sebou a vzniklé nové signály x_i jsou zaznamenávány. Signály s_j jsou označovány jako zdrojové a x_i jako pozorované. Tyto signály mohou být závislé na libovolném počtu souřadnic. V této práci jsou ale uvažovány výhradně akustické signály závislé na diskrétních vzorcích $n \in \mathbb{Z}$. Pro jednotlivé signály x_i pak platí

$$x_i(n) = a_{i1}s_1(n) + a_{i2}s_2(n) + \cdots + a_{im}s_m(n), \quad i = 1,\ldots,m, \tag{2.1}$$

kde m je počet zdrojových i pozorovaných signálů. Tuto rovnici je možné zapsat vektorově jako

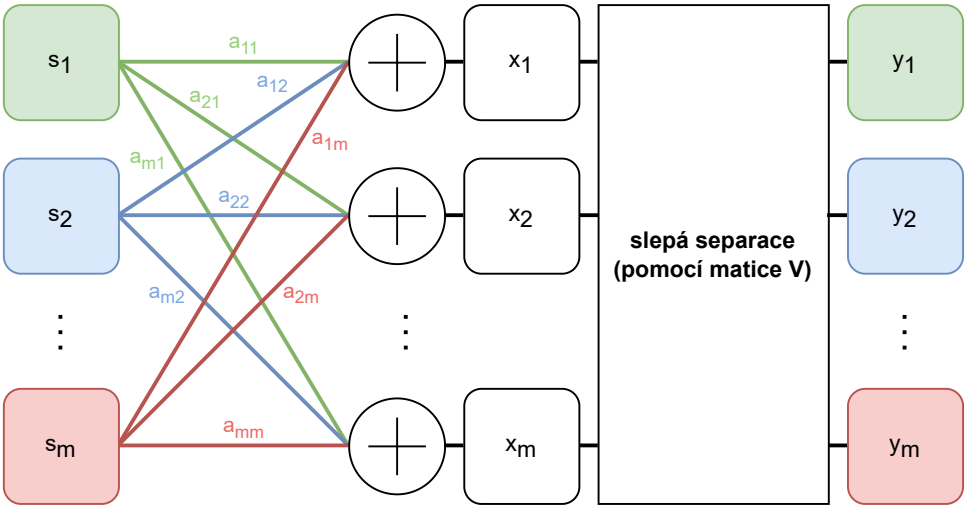
$$\mathbf{x} = \mathbf{A}\mathbf{s}, \tag{2.2}$$

kde $\mathbf{A}$ je čtvercová mixážní matice koeficientů a_{ij}, $\mathbf{x} = (x_1(n),\ldots,x_m(n))^{\mathrm{T}}$ je vektor pozorovaných signálů (jednotlivé signály jsou pak v řádcích) a $\mathbf{s} = (s_1(n),\ldots,s_m(n))^{\mathrm{T}}$ je vektor zdrojových signálů.

Matice $\mathbf{A}$ by v rámci zavedeného modelu mohla být i obdélníková, což by vyjadřovalo situaci, kdy by počty zdrojových a pozorovaných signálů nebyly stejné. Při separaci ale většina algoritmů pracuje s předpokladem čtvercové mixážní matice a stejná situace bude uvažována i v této práci. V případě, kdy by existovalo více pozorovaných signálů než zdrojových, se provede předzpracování pomocí redukce dimenzionality pozorovaných signálů na stejný počet jako signálů $\mathbf{s}$, aby byla mixážní matice opět čtvercová. Opačný případ (s nižším počtem pozorovaných signálů než zdrojových) pomocí analýzy nezávislých komponent separovat nelze a bylo by nutné použít pokročilejší metody, které jsou mimo zaměření této práce [3].

Cílem slepé separace je pouze se znalostí signálů $\mathbf{x}$ určit zdrojové signály $\mathbf{s}$. Toho je dosaženo pomocí separační matice $\mathbf{V}$, pro kterou platí

$$\mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s} = \mathbf{s}. \tag{2.3}$$



Obrázek 2.1: Nákres lineární směsi a separace při stejném počtu signálů s_j a x_i

V praxi je tato úloha řešena separačními algoritmy, které při splnění určitých předpokladů (uvedeny dále) separační matici pouze odhadnou. Výsledkem je odhad zdrojových signálů $\mathbf{y}$, pro který platí

$$\mathbf{y} = \mathbf{V}\mathbf{x} = \mathbf{V}\mathbf{A}\mathbf{s} \approx \mathbf{s}. \tag{2.4}$$

Podstata lineární směsi a následné separace je patrná z obrázku 2.1.

2.2 Principiální řešení

Tato práce se soustředí na slepou separaci pomocí analýzy nezávislých komponent (dále jako ICA). Tu je možné provést pomocí řady metod, které ale všechny fungují na podobném principu. Jedná se o postupné iterování separační matice $\mathbf{V}$ za účelem maximalizace míry separace, přičemž metody ICA jako měřítko separace používají míru nezávislosti odseparovaných signálů. Různé metody ICA se mezi sebou liší způsobem, jakým onu nezávislost měří.

Jedním z možných měřítek nezávisosti je „negaussianita“ [1, s. 6]. Intuitivně jde o důsledek centrální limitní věty. Pokud bychom signály chápali jako náhodné proměnné se stejným rozdělením, pak by se rozdělení lineárních směsí těchto proměnných blížilo Gaussovskému rozdělení. A naopak zpětně separované proměnné by měly mít takové rozdělení, které se Gaussovskému blíží co nejméně, tedy má co největší negaussianitu. V praxi se pak negaussianita přibližně určuje pomocí různých kontrastů, jako jsou například špičatost (kurtotis) nebo negentropie. Na principu maximalizace neagaussianity pracuje často používaný a účinný algoritmus „FastICA“ [1, s. 14].

Lineární směs lze separovat i bez odhadování neagaussianity signálů. Nabízí se separace pomocí metod maximální věrohodnosti (MLE ICA), mezi něž se řadí například Infomax [1, s. 10]. Další metoda pak pracuje na principu minimalizace vzájemné informace separovaných signálů [1, s. 9].

Všechny tyto algoritmy fungují pouze při splnění určitých předpokladů. Protože cílem je co největší nezávislost separovaných signálů, je základním předpokladem, aby zdrojové signály byly vzájemně nezávislé. Z podstaty úlohy (ukázáno [1, s. 5]) je také nemožné separovat stacionární signály s normálním rozdělením. Splnění těchto dvou předpokladů je dáno charakterem zdrojových signálů, který v praxi nelze měnit. Je například běžné, že budou zdrojové signály obohaceny o určité množství šumu s normálním rozdělením. V takovém případě budou algoritmy slepé separace funkční, ale úspěšnost separace bude klesat s množstvím tohoto šumu v signálech. Dalším předpokladem je také stejné množství pozorovaných a zdrojových signálů, čehož je v praxi možné dosáhnout adekvátním množstvím senzorů snímajících mixované signály. Posledním předpokladem je nulová střední hodnota zdrojových signálů. Toho obecně dosáhnout nelze, protože zdrojové signály jsou neznámé. V praxi je ale postačující řešit tento problém centrováním signálů pozorovaných [1, s. 2].

Před samotnou separací signálů je výhodné (a u některých metod i nutné) provést bělení (dekorelaci) pozorovaných signálů [1, s. 12]. Toho je možné dosáhnout například pomocí kovarianční matice zdrojových signálů $\mathbf{R}_{\mathbf{x}}$. Lepší způsob je společně s bělením provést i redukci dimenzionality pozorovaných signálů (v případě, že je jejich počet větší než separovaných zdrojů $\mathbf{s}$). Obojí je možné provést pomocí analýzy hlavních komponent (PCA) [5].

Vektor $\mathbf{x}_s = (x_{s1}(n), \ldots, x_{sk}(n))^{\mathrm{T}}$ obsahuje nevybělené zdrojové signály o počtu $k \geq m$. Cílem PCA je získat vektor $\mathbf{x} = (x_1(n), \ldots, x_m(n))^{\mathrm{T}}$ vybělených signálů, na základě kterých bude odhadnuta čtvercová separační matice $\mathbf{V}$, a které budou separovány v odhady zdrojových signálů $\mathbf{y}$, jak vyjadřuje rovnice (2.4).

Toho je dosaženo následovně [5]. Nejprve je proveden SVD rozklad kovarianční matice $\mathbf{R}_{\mathbf{x}_s}$ signálů $\mathbf{x}_s$

$$\mathbf{R}_{\mathbf{x}_s} = \mathrm{E}\left[(\mathbf{x}_s - \bar{\mu}_{\mathbf{x}_s})(\mathbf{x}_s - \bar{\mu}_{\mathbf{x}_s})^{\mathrm{T}}\right] = \mathbf{U}\mathbf{\Sigma}\mathbf{W}^{\mathrm{T}}, \tag{2.5}$$

kde matice $\mathbf{U}$, $\mathbf{\Sigma}$ a $\mathbf{W}$ představují výsledek SVD rozkladu, $\bar{\mu}_{\mathbf{x}_s}$ jsou odhady středních hodnot signálů $\mathbf{x}_s$ a E představuje střední hodnotu, přičemž pro její výpočet bude použit konzistentní vychýlený odhad. Matice $\mathbf{U}$ je následně upravena na matici $\mathbf{U}_m$, a sice takovým způsobem, že se zachová pouze jejích prvních m sloupců, které odpovídají nejsilnějším hlavním komponentám. Tím dojde ke zredukování dimenzionality, přičemž rozměry matice $\mathbf{U}_m$ budou $k \times m$. Úpravou projde i matice $\mathbf{\Sigma}$. Jedná se o diagonální matici s rozměry $k \times k$, na jejíž diagonále se nacházejí singulární hodnoty $\sigma_1, \ldots, \sigma_k$. Upravená matice $\mathbf{\Sigma}_m$ bude diagonální o rozměrech $m \times m$, na jejíž diagonále se bude nacházet pouze prvních m singulárních hodnot σ_i.

Celý právě popsaný proces, kdy se na základě signálů $\mathbf{x}_s$ a hodnoty m určí pozměněné matice $\mathbf{U}_m$ a $\mathbf{\Sigma}_m$, bude v této práci souhrně označen jako PCA s notací $[\mathbf{U}_m, \mathbf{\Sigma}_m] = \mathrm{PCA}[\mathbf{x}_s, m]$.

Po provedení PCA dojde k určení vektoru $\mathbf{x}$ jako

$$\mathbf{x} = \mathbf{\Sigma}_m^{-\frac{1}{2}} \mathbf{U}_m^{\mathrm{T}} (\mathbf{x}_s - \bar{\mu}_{\mathbf{x}_s}), \tag{2.6}$$

kde vektor $\mathbf{x}$ obsahuje vycentrované a vybělené signály o stejném počtu, jako je signálů zdrojových. Tyto signály splňují předpoklady pro jejich separování pomocí metod ICA.

Určité předpoklady se týkají i separační matice $\mathbf{V}$. Z důvodů popsaných v [1] je nutné, aby byla matice $\mathbf{V}$ mezi jednotlivými iteracemi algoritmu stále ortogonální. Toho lze dosáhnout například Gram - Schmidtovou ortogonalizací a normalizací řádků matice. Tento postup ale závisí na pořadí ortogonalizovaných vektorů a často je nutné použít symetrickou ortogonalizaci, při které na tomto pořadí nezáleží. Zde ze dá využít pokročilejší iterativní postup ukázaný v [6]. V této práci bude využívána třetí možnost ortogonalizace matice $\mathbf{V}$ daná následujícím vzorcem [1, s. 15]

$$\mathbf{V} \coloneqq \left(\mathbf{V}\mathbf{V}^{\mathrm{T}}\right)^{-\frac{1}{2}} \mathbf{V}. \tag{2.7}$$

Jak již bylo řečeno, algoritmy ICA separující lineární směsi signálů v iteracích postupně upravují matici $\mathbf{V}$, sledují míru separace a je nutné určit, kdy tyto iterace ukončit a separaci prohlásit za dostatečnou. Nejjednodušší možností je provést předem stanovené množství iterací. Tento postup je v praxi funkční, ale vzhledem k poměrně vysoké výpočetní náročnosti samotných iterací není příliš efektivní. Lepší je mezi iteracemi sledovat buď změnu separovaných signálů, nebo ještě lépe změnu samotné separační matice. Jak plyne z rovnice (2.4), jednotlivé signály y_i jsou separovány řádky separační matice $\mathbf{v}_i^{\mathrm{T}}$ a platí pro ně $y_i = \mathbf{v}_i^{\mathrm{T}}\mathbf{x}$. Pokud se mezi iteracemi nebudou významně měnit řádky matice $\mathbf{V}$, nebudou se měnit ani separované signály a separaci lze prohlásit za dostatečnou. Změnu řádků lze vyjádřit jejich skalárním součinem, který se bude u dvou podobných vektorů $\mathbf{v}_i$ blížit ± 1. Záporné znaménko je dovoleno, protože vektory by mohly mít i přesně opačný směr, což by mělo vliv pouze na polaritu signálů a tu stejně, jak bude ukázáno dále, určit nelze. Hodnota 1 vyplývá z faktu, že je matice $\mathbf{V}$ stále ortogonalizována a její řádky mají jednotkovou normu. Pro ukončení iterací je pak samozřejmě nutné, aby se skalární součin řádků blížil ± 1 u všech řádků $\mathbf{v}_i^{\mathrm{T}}$. Ukončovací pravidlo iterací pak lze formulovat jako [1, s. 14]

$$\max|(1 - |\mathbf{V} \cdot \mathbf{V}_o|)| < \epsilon, \tag{2.8}$$

kde $\mathbf{V}_o$ představuje separační matici minulé iterace, operace $\cdot$ vyjadřuje skalární součin řádků matic a ϵ je zvolená toleranční konstanta blízká nule.

2.3 Stručný popis algoritmů

Algoritmy ICA pro separaci lineární směsi signálů vycházející z řešení nastíněných v této kapitole jsou shrnuty v pseudokódu 1. Použité funkce f a h záleží na příslušném algoritmu a jejich konkrétní definice jsou v [1].

Pseudokód 1 Algoritmy ICA separující lineární směs signálů

1: $[\mathbf{U}_m, \mathbf{\Sigma}_m] \coloneqq \mathrm{PCA}\,[\mathbf{x}_s, m]$ ▷ určení $\mathbf{U}_m$ a $\mathbf{\Sigma}_m$ pomocí PCA
2: $\mathbf{x} \coloneqq \mathbf{\Sigma}_m^{-\frac{1}{2}} \mathbf{U}_m^{\mathrm{T}} (\mathbf{x}_s - \bar{\mu}_{\mathbf{x}_s})$ ▷ centrování, bělení a redukce dimenze
3: $\mathbf{V} \coloneqq \mathbf{I}_m$ ▷ inicializace jednotkovou maticí
4: **repeat**
5: $\quad \mathbf{y} \coloneqq \mathbf{V}\mathbf{x}$ ▷ separace
6: $\quad \mathbf{V}_o \coloneqq \mathbf{V}$
7: $\quad \mathbf{V} \coloneqq \mathbf{V} + f(\mathbf{y}) + h(\mathbf{y})\mathbf{V}$ ▷ aktualizace matice $\mathbf{V}$
8: $\quad \mathbf{V} \coloneqq \left(\mathbf{V}\mathbf{V}^{\mathrm{T}}\right)^{-\frac{1}{2}} \mathbf{V}$ ▷ ortogonalizace
9: **until** $\max|(1 - |\mathbf{V} \cdot \mathbf{V}_o|)| < \epsilon$ ▷ podmínka konvergence
10: $\mathbf{y} \coloneqq \mathbf{V}\mathbf{x}$ ▷ výsledná separace

2.4 Nejednoznačnosti výsledku

Separační matice $\mathbf{V}$ (nalezená některým z algoritmů ICA) by se v ideálním případě blížila inverzi mixážní matice $\mathbf{A}$ a platilo by

$$\mathbf{y} = \mathbf{V}\mathbf{x} \approx \mathbf{A}^{-1}\mathbf{x} = \mathbf{A}^{-1}\mathbf{A}\mathbf{s} = \mathbf{s} \tag{2.9}$$

a signály $\mathbf{y}$ by byly přímými odhady signálů $\mathbf{s}$. Jak je ale ukázáno v [1, s. 3], pomocí uvedených postupů není možné signály takto jednoznačně separovat. U separovaných signálů bude vždy existovat nejednoznačnost (nebo také nejistota) v amplitudě a v jejich pořadí, jak lze vyjádřit následující rovnicí.

$$\mathbf{y} = \mathbf{V}\mathbf{x} \approx \mathbf{P}\mathbf{D}\mathbf{A}^{-1}\mathbf{x} = \mathbf{P}\mathbf{D}\mathbf{A}^{-1}\mathbf{A}\mathbf{s} = \mathbf{P}\mathbf{D}\mathbf{s}. \tag{2.10}$$

$\mathbf{P}$ je permutační matice vyjadřující nejistotu v pořadí signálů a $\mathbf{D}$ je diagonální matice s různými a neznámými škálami separovaných signálů [2]. Ačkoli při separování lineárních směsí tyto nejednoznačnosti mohou být v praxi často ignorovány, v případě separování řečových signálů je bude nutné řešit.

Kapitola 3

Separace konvoluční směsi řečových signálů

3.1 Popis problému

Cílem této práce je dosáhnout separace řečových signálů smíchaných v běžné a uzavřené místnosti. Do modelu této situace je tedy nutné zahrnout co nejvíce aspektů šíření zvuku uzavřeným prostorem jako jsou odrazy, zpoždění, ozvěny, difuzní pole a mnoho dalších [7]. Protože se ale jedná o separaci slepou, kdy jsou známy pouze pozorované signály $\mathbf{x}$, není možné vliv místnosti na šíření řečových signálů nijak kompenzovat (například známou impulsní odezvou místnosti). Z tohoto důvodu se tato práce nezabývá detailním popisem šíření zvuku v místnosti, ale snahou bude vytvořit obecný model směsi signálů umožňující následnou slepou separaci. Z uvedených poznatků je ale zřejmé, že lineární model směsi, představený kapitole 2, nebude dostatečný. [2]

Situaci lze lépe modelovat pomocí takzvané konvoluční směsi signálů, pro kterou platí

$$x_i(n) = \sum_{j=1}^{m} \sum_{\tau} a_{ij}(\tau) s_j(n-\tau) = \sum_{j=1}^{m} a_{ij}(n) * s_j(n), \tag{3.1}$$

kde $a_{ij}(n)$ odpovídá impulzní odezvě cesty ze zdroje j na senzor i. Prozatím bude pro jednoduchost uvažován stejný počet zdrojových a pozorovaných signálů, ale později bude jednoduše přidána možnost separace signálů z vícero senzorů pomocí redukce dimenzionality. Vztah (3.1) je možné zapsat analogicky k rovnici (2.2) jako

$$\mathbf{x} = \mathbf{A} * \mathbf{s}. \tag{3.2}$$

Separace takové směsi by byla možná nalezením separační matice $\mathbf{V}$, pro kterou by platilo

$$\mathbf{y} = \mathbf{V} * \mathbf{x} = \mathbf{V} * (\mathbf{A} * \mathbf{s}) \approx \mathbf{PDs}, \tag{3.3}$$

kde matice $\mathbf{P}$ a $\mathbf{D}$ vyjadřují opět permutační, respektive amplitudovou nejistotu v separovaných signálech.

3.2 Principiální řešení

Řešit takto definovanou konvoluční směs v časové oblasti by byl ale náročný problém a je výhodné převést celý problém do časově - frekvenční oblasti, konkrétně pomocí krátkodobé Fourierovy transformace (dále jako STFT) [2].

STFT $X(\omega_f, t_s)$ signálu $x(n)$ je definována jako

$$X(\omega_f, t_s) = \sum_n \mathrm{e}^{-\mathrm{j}\omega_f n} x(n) w(n - t_s), \tag{3.4}$$

$$\omega_f = 2\pi \frac{f-1}{F}, \quad f = 1, \ldots, F, \tag{3.5}$$

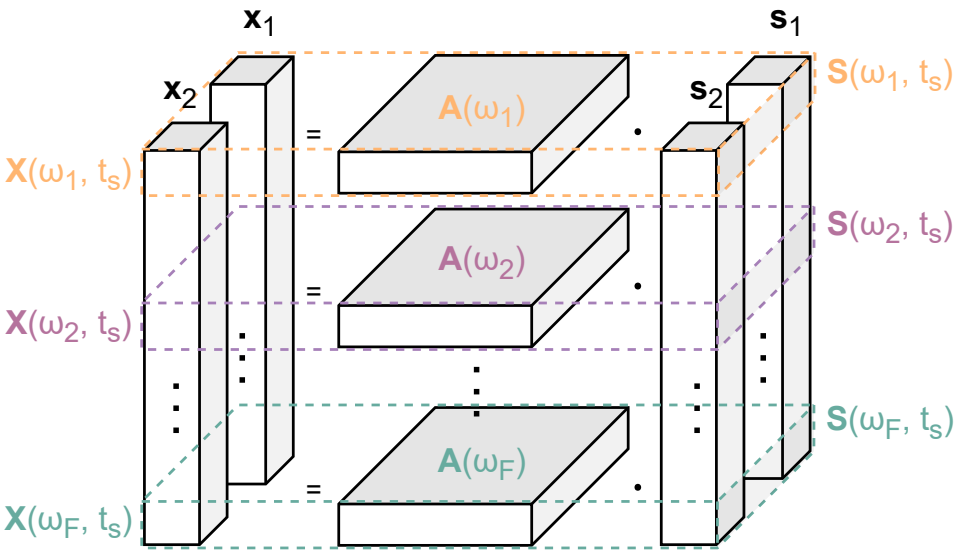
$$t_s = 0, \Delta T, 2\Delta T, \ldots, \tag{3.6}$$

kde ω_f je normalizovaná frekvence, F je počet bodů diskrétní Fourierovy transformace, t_s představuje pořadí segmentu, w je použité okno a ΔT je časový posun jednotlivých segmentů. Výsledek $X(\omega_f, t_s)$ je také často nazýván jako spektrogram signálu $x(n)$. V této práci je uvažováno více signálů $x_i(n)$ sdružených ve vektoru $\mathbf{x}$. Spektrogramy všech signálů v takovém vektoru by pak byly vyjádřeny jako $\mathbf{X}(\omega_f, t_s)$.

Aplikováním STFT na rovnici (3.2) vnikne vztah

$$\mathbf{X}(\omega_f, t_s) = \mathbf{A}(\omega_f)\mathbf{S}(\omega_f, t_s), \tag{3.7}$$

kde $\mathbf{S}(\omega_f, t_s)$ je STFT zdrojových signálů a $\mathbf{A}(\omega_f)$ představuje Fourierovu transformaci mixážní matice $\mathbf{A}$. Popsaný problém je pro úplnost nastíněn i graficky na obrázku 3.1.



Obrázek 3.1: Nákres rozkladu konvoluční směsi 2 signálů do frekvenčních pásem [8]

Díky reprezentaci směsi ve frekvenční oblasti přešla konvoluce v prosté násobení a vztah (3.7) představuje opět lineární směs, přičemž mixážní matice $\mathbf{A}(\omega_f)$ je pouze frekvenčně závislá. Časová nezávislost matice $\mathbf{A}(\omega_f)$ implikuje, aby i prostředí, kde k směšování dochází, bylo v čase neměnné. Proto je nutným a velmi důležitým předpokladem statická pozice zdrojů a senzorů řečových signálů. Pohybující se zdroje nebo senzory by měnily geometrickou podstatu směsi v čase a slepá separace by nebyla dostatečná. Druhým předpokladem vyplývajícím z časové nezávislosti matice $\mathbf{A}(\omega_f)$ je, aby vlastnosti prostředí ovlivňující šíření zvuku (jako teplota, rozměry místnosti apod.) byly po celou dobu trvání signálů konstantní, což je ale v praxi většinou zajištěno automaticky. [2]

Samotná separace pak spočívá v nalezení separačních matic $\mathbf{V}(\omega_f)$, pro které platí

$$\mathbf{Y}(\omega_f, t_s) = \mathbf{V}(\omega_f)\mathbf{X}(\omega_f, t_s) \approx \mathbf{PDS}(\omega_f, t_s). \tag{3.8}$$

Spektrogramy $\mathbf{Y}(\omega_f, t_s)$ je po provedení separací v jednotlivých pásmech nutné transformovat zpět do časové oblasti a získat tak akustické signály $\mathbf{y}$. K této transformaci slouží inverzní krátkodobá Fourierova transformace (dále jako ISTFT) daná vztahy [2]

$$y(n) = \frac{1}{2\pi}\frac{1}{W(n)}\sum_{t_s}\sum_{\omega_f} \mathrm{e}^{\mathrm{j}\omega_f(n-t_s)}Y(\omega_f, t_s), \tag{3.9}$$

$$W(n) = \sum_{t_s} w(n - t_s). \tag{3.10}$$

Pro lepší přehlednost proměnných z uvedených rovnic bude zavedena odlišná notace, ve které bude vynechán časový index t_s a frekvence ω_f bude vyjádřena pouze pomocí indexu f. Vzniklá notace má podobu

$$\mathbf{X}(\omega_f, t_s) \to \mathbf{x}^f = \left(x_1^f, \ldots, x_m^f\right)^{\mathrm{T}}, \tag{3.11}$$

$$\mathbf{S}(\omega_f, t_s) \to \mathbf{s}^f = \left(s_1^f, \ldots, s_m^f\right)^{\mathrm{T}}, \tag{3.12}$$

$$\mathbf{Y}(\omega_f, t_s) \to \mathbf{y}^f = \left(y_1^f, \ldots, y_m^f\right)^{\mathrm{T}}, \tag{3.13}$$

$$\mathbf{A}(\omega_f) \to \mathbf{A}^f, \tag{3.14}$$

$$\mathbf{V}(\omega_f) \to \mathbf{V}^f, \tag{3.15}$$

přičemž rovnice (3.7) a (3.8) přejdou ve tvar

$$\mathbf{x}^f = \mathbf{A}^f\mathbf{s}^f, \tag{3.16}$$

$$\mathbf{y}^f = \mathbf{V}^f\mathbf{x}^f. \tag{3.17}$$

Separaci signálů $\mathbf{x}$, tedy nalezení matic $\mathbf{V}^f$ pro všechna frekvenční pásma, lze řešit vícero způsoby. V dalších sekcích bude nastíněno použití obecných metod ICA a její modifikace pro separaci vektorů známou jako IVA.

3.3 Separace pomocí metod ICA

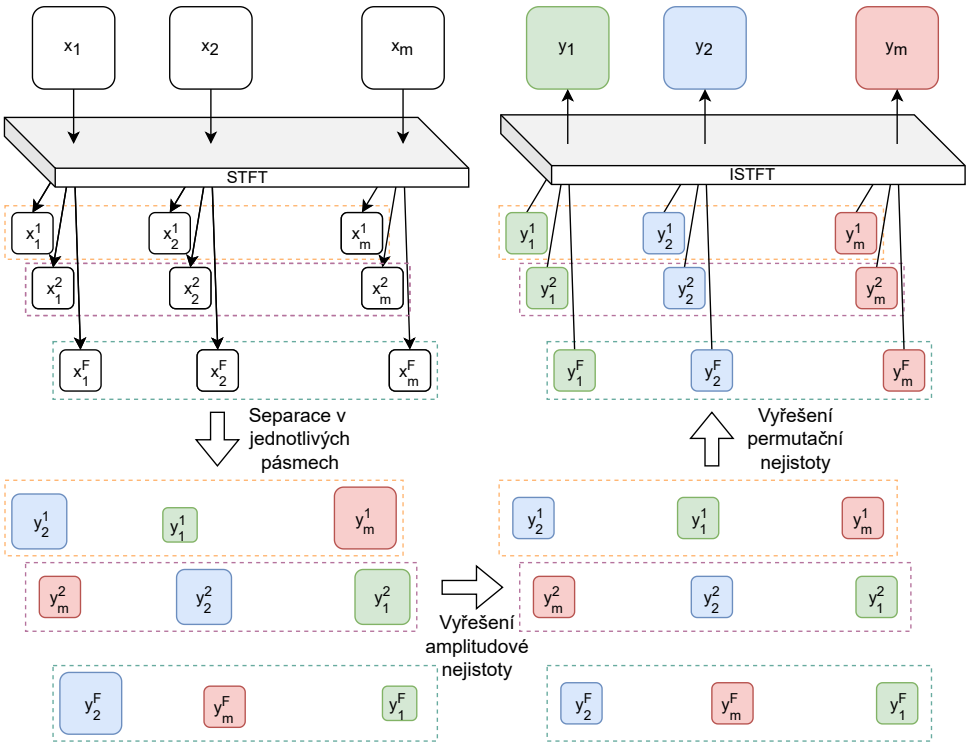
První možnost, jak popsanou konvoluční směs odseparovat, je poměrně intuitivní. Spektrogramy zdrojových signálů jsou analyzovány po jednotlivých frekvenčních pásmech f. V těchto pásmech je na základě rovnice (3.16) předpokládána lineární směs signálů a ty je možné separovat některou z metod ICA. Tento postup obnáší několik problémů.

Separované signály $\mathbf{x}^f$ nejsou reálné, ale komplexní. Jedná se totiž o výsledky STFT v jednotlivých frekvenčních pásmech f. Separační algoritmy je proto nutné přizpůsobit separaci komplexních signálů.

Kvůli použití některé z metod ICA budou separované signály zatížené permutační a amplitudovou nejistotou. Tento fakt lze vyjádřit jako

$$\mathbf{y}^f = \mathbf{V}^f \mathbf{x}^f \approx \mathbf{P}^f \mathbf{D}^f \mathbf{s}^f, \tag{3.18}$$

kde matice $\mathbf{P}^f$ a $\mathbf{D}^f$ jsou opět neznámé permutační, respektive diagonální matice, přičemž jejich tvar je ale tentokrát pro každé pásmo f jiný. Různě zpermutované a přeškálované tedy nebudou výsledné akustické signály $\mathbf{y}$, ale jednotlivá frekvenční pásma $\mathbf{y}^f$ a jejich naivním složením a převedením do časové oblasti by vznikly náhodné a nesrozumitelné signály. V každém frekvenčním pásmu je tedy nutné odseparovaná pásma jak správně třídit, tak přenásobit vhodnými škálami. [2]



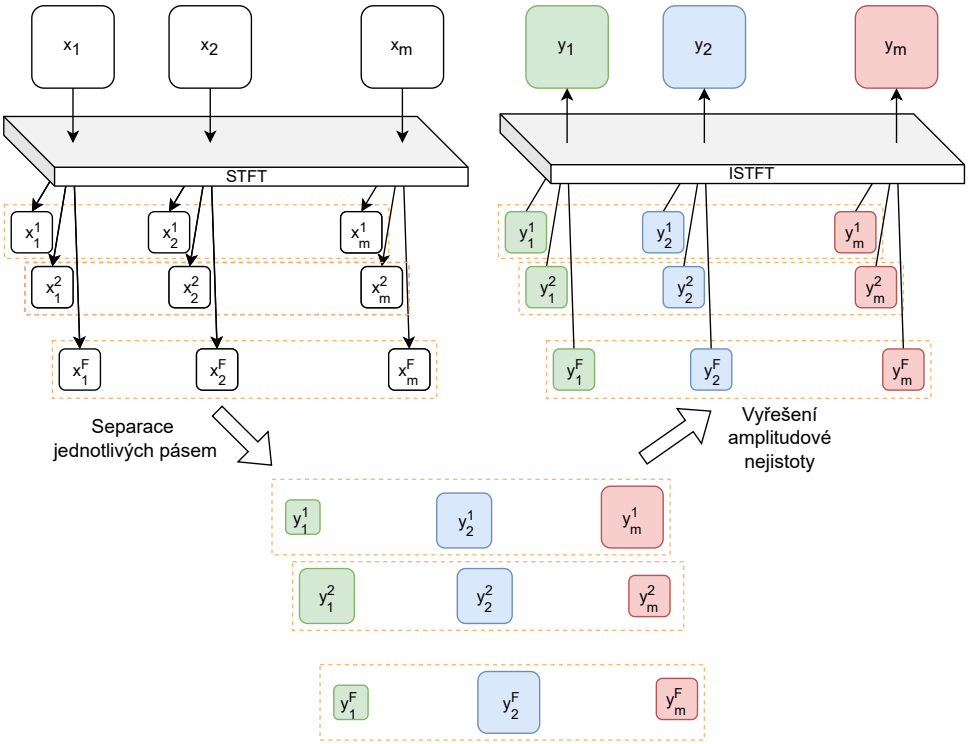
Obrázek 3.2: Základní schéma separace řečových signálů metodami ICA

Konkrétní řešení uvedených problémů bude podrobně rozebráno v následující kapitole. Separace metodami ICA s rozkladem do frekvenčních pásem je pak schématicky znázorněna na obrázku 3.2.

3.4 Separace pomocí metod IVA

Zajímavou modifikací metod ICA, kterou lze pro separaci řečových signálů s výhodou použít, je takzvaná „analýza nezávislých vektorů“ (dále jako IVA). Jedná se o soubor algoritmů schopných separovat lineární směs zdrojových signálů s tím rozdílem, že tyto signály jsou vícerozměrné nebo také takzvaně vektorové [9].

Vektorový model směsi, kterou metody IVA separují, lze na separaci řečových signálů přímo aplikovat a je popsán rovnicí (3.16), přičemž ony vektory se skládají z jednotlivých pásem spektrogramů s_i^f, x_i^f a y_i^f. Počet prvků vektorů odpovídá počtu frekvenčních pásem F.



Obrázek 3.3: Základní schéma separace řečových signálů metodami IVA

Princip metod IVA je v podstatě úplně stejný, jako u metod ICA, akorát je rozšířen na celý vektor. Tato metoda tedy neseparuje jednotlivé lineární mixáže popsané rovnicí (3.16) zvlášť v jednotlivých pásmech, ale pracuje se všemi pásmy f najednou. Předpoklady metod IVA proto zůstávají takřka stejné, jako u metod ICA. Požadavek nezávislosti zdrojových signálů se mění na požadavek nezávislosti zdrojových vektorů a dále se přidává požadavek

vzájemné závislosti prvků jednotlivých vektorů. V případě slepé separace řečových signálů to znamená vzájemnou závislost jednotlivých frekvenčních pásem, což lze v případě řeči do určité míry očekávat automaticky. [8] [9]

Pokud bude předpoklad vzájemné závislosti prvků vektorů splněn, nedojde při separaci k náhodnému promíchání signálů v rámci jednotlivých pásem, díky čemuž zmizí permutační nejistota. To je hlavní výhoda a motivace pro použití metod IVA pro separaci řečových signálů. Nicméně amplitudová nejednoznačnost bude přítomná nadále a kromě ní je potřeba vyřešit i potřebu separace komplexních signálů. Oba problémy se ale řeší velmi podobně jako u metod ICA a budou představeny v další kapitole. Základní schéma separace konvoluční směsi řečových signálů metodami IVA je uvedeno na obrázku 3.3.

Dlužno dodat, že při použití libovolné představené metody pro separaci řečových signálů dojde u výsledných akustických signálů $\mathbf{y}$ k nejistotě v pořadí a amplitudě, a to i přes vyřešení těchto nejistot v jednotlivých frekvenčních pásmech. Tuto vlastnost je ale možné vzhledem k povaze úlohy ignorovat. Pořadí signálů není možné určit (pozice mluvčích je neznámá) a změna amplitudy má vliv pouze na hlasitost výsledných signálů. Tyto nejistoty bude nutné plně kompenzovat pouze v případě vyhodnocování kvality separace, kdy ale budou zdrojové signály $\mathbf{s}$ známé.

3.5 Separace řečníků pomocí vícero mikrofonů

V této kapitole byl pro jednoduchost a názornost uvažován stejný počet zdrojových a pozorovaných signálů. To odpovídá situaci, kdy je směs řečových promluv zaznamenávána na stejný počet mikrofonů, jako je v místnosti mluvčích. V praxi je ale výhodné při separaci využít celá mikrofonová pole o desítkách mikrofonů. Důvody pro tento přístup tkví zejména v přítomnosti šumů a hluků okolí při nahrávání skutečné směsi promluv v reálné místnosti, jak bude podrobněji diskutováno v praktické části práce.

Nicméně je žádoucí nastíněné metody ICA i IVA upravit tak, aby bylo možné separovat určitý počet promluv na základě dat z více mikrofonů, přičemž je předpokládáno, že počet mluvčích m je předem známý. Tato potřeba se jednoduše vyřeší pomocí redukce dimenzionality, jako tomu bylo v kapitole 2. Vektor $\mathbf{x}_s$ obsahuje signály z k mikrofonů, přičemž platí, že $k \geq m$. Tento vektor bude pomocí STFT rozložen do frekvenčních pásem sdružených ve spektrogramech $\mathbf{x}_s^f$. Na základě $\mathbf{x}_s^f$ jsou následně pomocí PCA získány spektrogramy $\mathbf{x}^f$.

$$\left[\mathbf{U}_m^f, \mathbf{\Sigma}_m^f\right] = \mathrm{PCA}\left[\mathbf{x}_s^f, m\right], \tag{3.19}$$

$$\mathbf{x}^f = {\mathbf{\Sigma}_m^f}^{-\frac{1}{2}} {\mathbf{U}_m^f}^{\mathrm{T}} \left(\mathbf{x}_s^f - \bar{\mu}_{\mathbf{x}_s^f}\right). \tag{3.20}$$

Pásma spektrogramu $\mathbf{x}^f$ obsahují vycentrovaná a vzájemně vybělená data z m signálů. Takto předzpracované spektrogramy $\mathbf{x}^f$ budou pomocí metod ICA a IVA separovány způsobem nastíněným v této kapitole.

Kapitola 4

Separace řeči pomocí ICA

Základní princip využití metod ICA při slepé separaci konvoluční směsi řečových signálů byl nastíněn v sekci 3.3 a také na obrázku 3.2. Jak bylo zmíněno, je především potřeba vyřešit přizpůsobení separačních algoritmů ICA komplexním signálům a vyřešit permutační a amplitudovou nejistotu.

4.1 Separace komplexních signálů

Cílem této podkapitoly je pozměnit některou ze základních separačních metod ICA představených v kapitole 2 tak, aby byla schopná separovat lineární směsi komplexních signálů. Tu je možné vyjádřit rovnicí (2.2), kde prvky $\mathbf{s}$, $\mathbf{x}$ i $\mathbf{A}$ mohou nabývat komplexních hodnot a vektor $\mathbf{x}$ je uvažován už jako vycentrovaný, vybělený a s případně sníženou dimenzí, jak je popsáno v předchozí kapitole. Separace probíhá pomocí komplexní matice $\mathbf{V}$ na základě vztahu

$$\mathbf{y} = \mathbf{V}^{\mathrm{H}}\mathbf{x} = \mathbf{V}^{\mathrm{H}}\mathbf{A}\mathbf{s} \approx \mathbf{P}\mathbf{D}\mathbf{s}. \tag{4.1}$$

Rozdíl oproti rovnici (2.4) je tedy v Hermitovské transpozici matice $\mathbf{V}$. Tato transpozice se pochopitelně projeví i u dalších vztahů typických pro metody ICA. Děje se tak čistě z formálních důvodů popsaných v [10].

Algoritmů řešících tuto úlohu je opět více. V této práci je využíván pouze jeden z nich, konkrétně „Fast fixed-point ICA“ algoritmus, který byl vybrán pro svoji jednoduchost, rychlost a robustnost [10]. Zde bude stručně uveden a okomentován jeho pseudokód 2.

Použitý výraz $|\mathbf{y}|^2$ je vektor druhých mocnin modulů jednotlivých prvků komplexního vektoru $\mathbf{y}$ a oba vektory budou mít tedy stejný rozměr. Operace $\circ$ představuje Hadamardův součin a $\mathrm{diag}\,[\mathbf{a}]$ převede vektor $\mathbf{a}$ na diagonální matici, kde prvky na hlavní diagonále odpovídají prvkům vektoru $\mathbf{a}$.

Funkce g je použitá nelinearita (také někdy označována jako kontrast), kdy g′ vyjadřuje její derivaci. Výraz $\mathrm{g}\left(|\mathbf{y}|^2\right)$ představuje vektor výsledných hodnot nelinearit po aplikování na jednotlivé vzorky vektoru $|\mathbf{y}|^2$, takže vektory $|\mathbf{y}|^2$ a $\mathrm{g}\left(|\mathbf{y}|^2\right)$ budou mít stejný rozměr.

Pseudokód 2 Algoritmus „Fast fixed-point ICA“ separující lineární směs komplexních signálů

$\mathbf{V} \coloneqq \mathbf{I}_m$ ▷ inicializace jednotkovou maticí
repeat
  $\mathbf{y} \coloneqq \mathbf{V}^{\mathrm{H}}\mathbf{x}$ ▷ separace
  $\mathbf{V}_o \coloneqq \mathbf{V}$
  $gy \coloneqq \mathrm{g}\left(|\mathbf{y}|^2\right)$ ▷ nelinearity
  $dgy \coloneqq \mathrm{g}'\left(|\mathbf{y}|^2\right)$
  $\mathbf{V} \coloneqq \mathrm{E}\left[\mathbf{x}\,(\mathbf{y} \circ gy)^{\mathrm{H}}\right] - \mathrm{diag}\left[\mathrm{E}\left[gy + |\mathbf{y}|^2 \circ dgy\right]\right]\mathbf{V}$ ▷ aktualizace $\mathbf{V}$
  $\mathbf{V} \coloneqq \mathbf{V}\left(\mathbf{V}^{\mathrm{H}}\mathbf{V}\right)^{-\frac{1}{2}}$ ▷ ortogonalizace
until $\max\left|\left(1 - \left|\mathbf{V}^{\mathrm{H}} \cdot \mathbf{V}_o^{\mathrm{H}}\right|\right)\right| < \epsilon$ ▷ podmínka konvergence
$\mathbf{y} \coloneqq \mathbf{V}^{\mathrm{H}}\mathbf{x}$ ▷ výsledná separace

Tvar funkce g nemůže být libovolný. V rámci představeného algoritmu jsou využívány 3 možné definice, a sice [10]

$$\mathrm{g}_1(y) = \frac{1}{2\sqrt{a+y}}, \quad \mathrm{g'}_1(y) = \frac{-1}{4\sqrt{(a+y)^3}}, \tag{4.2}$$

$$\mathrm{g}_2(y) = \frac{1}{a+y}, \quad \mathrm{g'}_2(y) = \frac{-1}{(a+y)^2}, \tag{4.3}$$

$$\mathrm{g}_3(y) = y, \quad \mathrm{g'}_3(y) = 1, \tag{4.4}$$

kde a je konstanta, již je možné zvolit tak, aby $a \approx 0{,}1$. Kontrast g_1 je založen na odmocnině, g_2 na logaritmu a g_3 je inspirován špičatostí. Volba nelinearity sice kvalitu separace do určité míry ovlivní, nelze ale prohlásit, že by jedna z nich vedla vždy na lepší výsledky. V této práci bude konkrétní nelinearita vybrána experimentálně na základě dosažených výsledků separace. [1] [10]

Představený algoritmus pro separaci komplexních signálů bude při separování řečových signálů aplikován na každé pásmo f zvlášť. Výsledkem budou separační matice $\mathbf{V}^f$. Spolu s provedením samotné separace je z důvodů uvedených v kapitole 3.3 nutné odstranit nejistotu v pořadí a amplitudě.

4.2 Vyřešení amplitudové nejistoty

Je uvažována situace, kdy byly v rámci slepé separace řečových signálů nalezeny odseparované spektrogramy $\mathbf{y}^f$ a je nutné vyřešit jejich amplitudovou nejistotu vyjádřenou vztahem (3.18). Toho lze dosáhnout pomocí jednoduché úpravy separačních matic $\mathbf{V}^f$, ale pouze ve chvíli, kdy je známá informace o amplitudových poměrech příslušných signálů nebo spektrogramů [11].

Je ale nutné připomenout, že separované spektrogramy $\mathbf{x}^f$ prošly bělením (výkony pásem jsou jednotkové a tím došlo ke ztrátě informace o amplitudě), separační matice $\mathbf{V}^f$ jsou udržovány ortogonálními a tím i odseparovaná

pásma spektrogramů $\mathbf{y}^f$ mají jednotkový výkon. Mixážní matice $\mathbf{A}^f$ a zdrojové spektrogramy $\mathbf{s}^f$ jsou neznámé. Jediná dostupná informace o amplitudových poměrech je součástí signálů $\mathbf{x}_s$, což jsou původní smíšené signály zaznamenané jednotlivými mikrofony. Ty jsou před samotnou slepou separací pomocí STFT podrobeny rozkladu do spektrogramů $\mathbf{x}_s^f$, na jednotlivá pásma je aplikována PCA, díky čemuž vzniknou vybělené spektrogramy $\mathbf{x}^f$, jak popisuje vztah (3.20).

Výsledkem separačního algoritmu jsou matice $\mathbf{V}^f$ separující pásma spektrogramů $\mathbf{x}^f$ dle vztahu (4.1). Aby součástí separovaných signálů nebo separační matice byla informace o amplitudových poměrech spektrogramů (a mohla být odstraněna amplitudová nejistota), bude separační matice $\mathbf{V}^f$ jednoduše pozměněna v matici $\mathbf{V}_s^f$ separující nevybělené signály $\mathbf{x}_s^f$ dle vztahů

$$\mathbf{V}_s^f = {\mathbf{V}^f}^{\mathrm{H}} {\mathbf{\Sigma}_m^f}^{-\frac{1}{2}} {\mathbf{U}_m^f}^{\mathrm{T}}, \tag{4.5}$$

$$\mathbf{y}^f = {\mathbf{V}^f}^{\mathrm{H}} \mathbf{x}^f = {\mathbf{V}^f}^{\mathrm{H}} {\mathbf{\Sigma}_m^f}^{-\frac{1}{2}} {\mathbf{U}_m^f}^{\mathrm{T}} \mathbf{x}_s^f = \mathbf{V}_s^f \mathbf{x}_s^f. \tag{4.6}$$

Vzniklé separační matice $\mathbf{V}_s^f$ se dále upraví dle vztahu [11]

$$\mathbf{V}_s^f := \mathrm{diag}\left[\mathbf{V}_s^{f\dagger}\right] \mathbf{V}_s^f, \tag{4.7}$$

kde $\mathbf{V}_s^{f\dagger}$ vyjadřuje pseudoinverzi matice $\mathbf{V}_s^f$ a výraz $\mathrm{diag}\left[\mathbf{V}_s^{f\dagger}\right]$ vytvoří diagonální matici z prvků na hlavní diagonále matice $\mathbf{V}_s^{f\dagger}$.

Díky této úpravě bude po provedení separace pro všechna pásma platit

$$\mathbf{y}^f = \mathbf{V}_s^f \mathbf{x}_s^f \approx \mathbf{P}^f \mathbf{D} \mathbf{s}^f, \tag{4.8}$$

kde $\mathbf{P}^f$ je permutační matice vyjadřující zatím nevyřešenou nejistotu v pořadí a $\mathbf{D}$ je diagonální matice vyjadřující stále přítomnou nejistotu v amplitudě, která je ale na úrovni výsledných akustických signálů a jak bylo zmíněno, bez znalosti zdrojových signálů $\mathbf{s}$ ji nelze odstranit a může být ignorována. [11]

4.3 Vyřešení permutační nejistoty

Na závěr je nutné odseparovaná pásma $\mathbf{y}^f$ správně setřídit. Formálně je tedy nutné pro každé frekvenční pásmo f určit permutaci π^f permutující pásma y_i^f dle vztahu

$$\mathbf{y}_p^f = \left(y_{p1}^f, \ldots, y_{pm}^f\right)^{\mathrm{T}} = \left(y_{\pi^f(1)}^f, \ldots, y_{\pi^f(m)}^f\right)^{\mathrm{T}}, \tag{4.9}$$

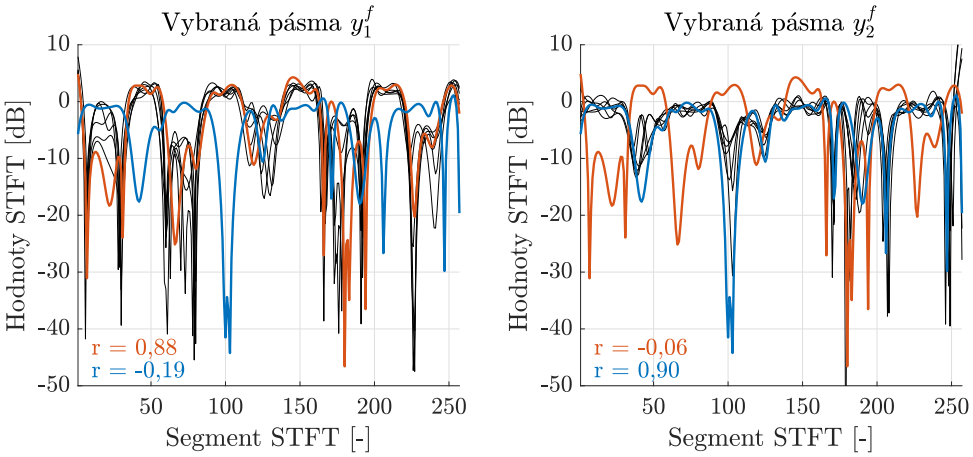
kde $\mathbf{y}_p^f$ je vektor správně setříděných pásem, ve kterém pro každé f platí

$$\mathbf{y}_p^f \approx \mathbf{P}\mathbf{D}\mathbf{s}^f, \tag{4.10}$$

kde matice $\mathbf{P}$ a $\mathbf{D}$ opět vyjadřují neřešitelné nejednoznačnosti na úrovni výsledných signálů.

V této práci se pro nalezení správné permutace využívá postup (podobný postupu používanému v [2]), kdy se frekvenční pásma třídí postupně na základě podobnosti tříděných pásem s pásmy již zatříděnými.

Intuitivně se jedná o důsledek silné vzájemné závislosti frekvenčních pásem v promluvách. Díky tomu lze pásma do patřičných spektrogramů správně zatřídit na základě podobnosti pásem s ostatními pásmy ve spektrogramu. Tuto podobnost pak lze vyjádřit pomocí korelace modulů frekvenčních pásem.



Obrázek 4.1: Příklad několika správně zatříděných pásem (černě) a aktuálně zatřiďovaných pásem dvou signálů

Malý příklad je pro představu tohoto principu uveden na obrázku 4.1, kde je nutné vyřešit zatřídění dvou pásem (červeně a modře) k již správně zatříděných pásmům y_1^f nebo y_2^f. Korelace modulů zatřiďovaných pásem se součtem modulů pásem již zatříděných jsou vypsány v obrázcích. Je vidět, že korelace dobře postihla grafickou podobnost modulů pásem a na základě jejích hodnot se zatřídí „červené“ pásmo k pásmům y_1^f a „modré“ k y_2^f.

Obecně lze celý postup zatřiďování vyjádřit jako

$$\begin{aligned}
&\mathbf{y}_p^1 = \mathbf{y}^1, \\
&\text{pro každé} \quad f = 2, \ldots, F \\
&\qquad \pi^f = \underset{\pi^f}{\operatorname{argmax}} \sum_{i=1}^{m} \mathrm{r}\left[\left|y^f_{\pi^f(i)}\right|, \sum_{\varphi=1}^{f-1}\left|y^{\varphi}_{pi}\right|\right], \\
&\qquad \mathbf{y}_p^f = \left(y^f_{\pi^f(1)}, \ldots, y^f_{\pi^f(m)}\right)^{\mathrm{T}},
\end{aligned} \tag{4.11}$$

kde r představuje korelaci. První frekvenční pásmo se zatřídí libovolně a dál se postupuje přes všechna frekvenční pásma a ty se zatřiďují takovým způsobem, aby se maximalizovala korelace zatříděných pásem se sumou pásem již zatříděných.

Tento způsob zatřiďování je sice jednoduchý, nemusí ale obecně vést k perfektním výsledkům a některá pásma mohou být zatříděna špatně. Jedná se zejména o pásma obsahující malé množství energie. Tato pásma na okolních

pásmech dostatečně závislá nejsou, korelace s ostatními pásmy nemusí být dostatečně signifikantní a kvůli tomu mohou být zatříděna špatně. Protože se ale jedná o pásma s malou energií, na výsledném akustickém signálu se tolik neprojeví a jejich kvalitu příliš neovlivní.

Ke špatnému setřídění frekvenčních pásem ale může dojít také u promluv s podobnými amplitudovými obálkami, protože v takovém případě od sebe jednotlivá pásma nelze (na základě jejich modulů) snadno odlišit. V tu chvíli tato metoda selhává a dává neuspokojivé výsledky. Separaci je poté lepší řešit metodami IVA, které amplitudovou nejistotu odstraňují lepším způsobem.

Druhým neduhem této metody je její vysoká výpočetní náročnost. Pro nalezení optimálního zatřídění je nutné vyzkoušet všechny možné permutace a s rostoucím počtem separovaných signálů jejich počet strmě roste. Pro rozumnou náročnost i při vysokých počtech separovaných signálů je možné využít Maďarský algoritmus. Úloha správného setřídění pásem je ve své podstatě „přiřazovací problém“, který je možné pomocí Maďarského algoritmu efektivně řešit [12].

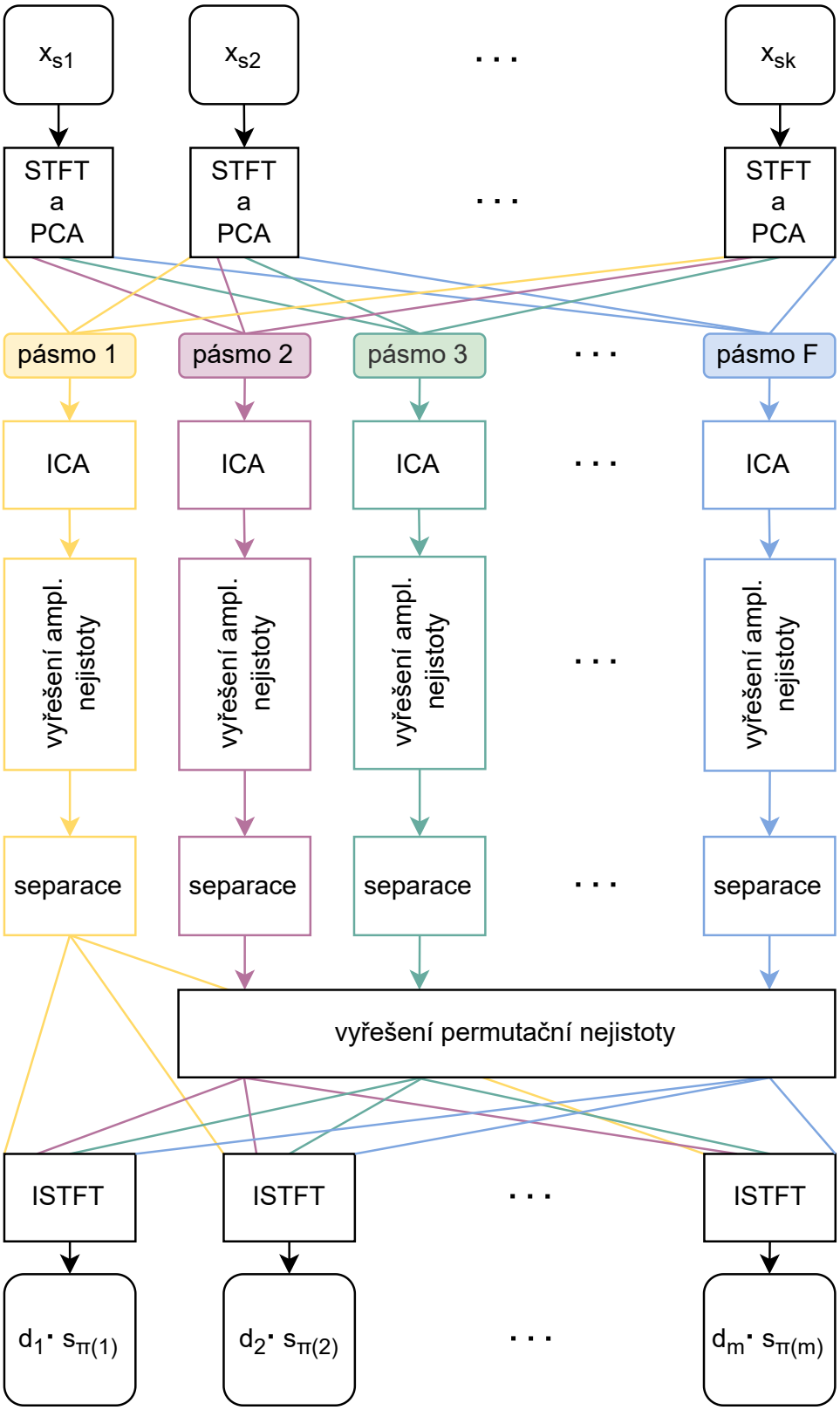
4.4 Finální schéma a pseudokód separace řečových signálů metodou ICA

Pro úplnost a přehlednost je zde uveden kompletní pseudokód 3 algoritmu, který slepě separuje směsi řečových signálů za využití metody Fast fixed-point ICA. Na obrázku 4.2 je také uvedeno jeho schéma. Jak je ve schématu naznačeno, výsledné signály budou v ideálním případě pouze náhodně promíchané a přeškálované zdrojové signály $\mathbf{s}_i$.

Nakolik je tento způsob slepé separace řečových signálů skutečně efektivní a přesný, bude vyhodnoceno v dalších kapitolách této práce.

Pseudokód 3 Algoritmus „Fast fixed-point ICA“ pro slepou separaci řečových signálů

1: $\mathbf{x}_s^f \coloneqq \mathrm{STFT}\left[\mathbf{x}_s\right]$ ▷ rozklad akustických signálů do frekv. pásem
2: **for** $f = 1, \ldots, F$ **do**
3: $\quad\left[\mathbf{U}_m^f, \mathbf{\Sigma}_m^f\right] \coloneqq \mathrm{PCA}\left[\mathbf{x}_s^f, m\right]$
4: $\quad\mathbf{x}^f \coloneqq {\mathbf{\Sigma}_m^f}^{-\frac{1}{2}} {\mathbf{U}_m^f}^{\mathrm{T}} \left(\mathbf{x}_s^f - \bar{\mu}_{\mathbf{x}_s^f}\right)$ ▷ centrování, bělení a redukce dimenze
5: $\quad\mathbf{V}^f \coloneqq \mathbf{I}_m$
6: $\quad$**repeat**
7: $\quad\quad\mathbf{y}^f \coloneqq {\mathbf{V}^f}^{\mathrm{H}} \mathbf{x}$
8: $\quad\quad\mathbf{V}_o \coloneqq \mathbf{V}^f$
9: $\quad\quad gy \coloneqq \mathrm{g}\left(\left|\mathbf{y}^f\right|^2\right)$
10: $\quad\quad dgy \coloneqq \mathrm{g}'\left(\left|\mathbf{y}^f\right|^2\right)$
11: $\quad\quad\mathbf{V}^f \coloneqq \mathrm{E}\left[\mathbf{x}^f \left(\mathbf{y}^f \circ gy\right)^{\mathrm{H}}\right] - \mathrm{diag}\left[\mathrm{E}\left[gy + \left|\mathbf{y}^f\right|^2 \circ dgy\right]\right] \mathbf{V}^f$
12: $\quad\quad\mathbf{V}^f \coloneqq \mathbf{V}^f \left({\mathbf{V}^f}^{\mathrm{H}} \mathbf{V}^f\right)^{-\frac{1}{2}}$
13: $\quad$**until** $\max\left|\left(1 - \left|{\mathbf{V}^f}^{\mathrm{H}} \cdot \mathbf{V}_o^{\mathrm{H}}\right|\right)\right| < \epsilon$
14: $\quad\mathbf{V}_s^f \coloneqq {\mathbf{V}^f}^{\mathrm{H}} {\mathbf{\Sigma}_m^f}^{-\frac{1}{2}} {\mathbf{U}_m^f}^{\mathrm{T}}$
15: $\quad\mathbf{V}_s^f \coloneqq \mathrm{diag}\left[{\mathbf{V}_s^f}^{\dagger}\right] \mathbf{V}_s^f$ ▷ odstranění amplitudové nejistoty
16: $\quad\mathbf{y}^f \coloneqq \mathbf{V}_s^f \mathbf{x}_s^f$
17: **End for**
18: $\mathbf{y}_p^1 \coloneqq \mathbf{y}^1$
19: $\mathbf{y}_p^{\mathrm{sum}} \coloneqq 0$
20: **for** $f = 2, \ldots, F$ **do** ▷ třídění pásem
21: $\quad\mathbf{y}_p^{\mathrm{sum}} \coloneqq \mathbf{y}_p^{\mathrm{sum}} + \left|\mathbf{y}_p^{f-1}\right|$
22: $\quad r^f \coloneqq \mathrm{r}\left[\left|\mathbf{y}^f\right|, \mathbf{y}_p^{\mathrm{sum}}\right]$
23: $\quad\pi^f \coloneqq \mathrm{Hung.Alg.}\left[\max\left[r^f\right] - r^f\right]$ ▷ Maďarský algoritmus
24: $\quad\mathbf{y}_p^f \coloneqq \left(y^f_{\pi^f(1)}, \ldots, y^f_{\pi^f(m)}\right)^{\mathrm{T}}$
25: **End for**
26: $\mathbf{y} \coloneqq \mathrm{ISTFT}\left[\mathbf{y}_p^f\right]$ ▷ zpětná syntéza odseparovaných spektrogramů



Obrázek 4.2: Schéma kompletní separace řečových signálů pomocí metod ICA

Kapitola 5

Separace řeči pomocí IVA

Aplikace metod IVA na separaci konvoluční směsi řečových signálů byla popsána v kapitole 3.4. Zde bude představen konkrétní algoritmus pro separaci spektrogramů řečových signálů. Na rozdíl od metody ICA bude nutné upravit podmínku konvergence algoritmu, ale nebude nutné řešit permutační nejistotu. Amplitudová nejistota bude řešena stejným způsobem, jako v minulé kapitole.

5.1 Popis algoritmu

Z metod IVA byl opět vybrán algoritmus „Fast fixed-point IVA“ uzpůsobený separaci spektrogramů řečových signálů. Zde je uveden pseudokód 4 kompletní separace řečových signálů využívající tohoto algoritmu. [8]

Schéma algoritmu zde v tomto případě uvedeno není, protože je v podstatě úplně stejné, jako schéma na obrázku 4.2 popisující metodu ICA (pouze by chyběly bloky pro řešení permutační nejistoty a blok IVA by se neomezoval na jednotlivá frekvenční pásma, ale pracoval by se všemi pásmy současně).

V pseudokódu použitá nelinearita g je derivací účelové funkce G odrážející úroveň separace. V této práci jsou uvažovány následující tvary funkcí [8]

$$\mathrm{G}_1(y) = \sqrt{a+y}, \quad \mathrm{g}_1(y) = \frac{1}{2\sqrt{a+y}}, \quad \mathrm{g'}_1(y) = \frac{-1}{4\sqrt{(a+y)^3}}, \tag{5.1}$$

$$\mathrm{G}_2(y) = \sqrt{\frac{2}{F}}\sqrt{a+y} + \left(F - \frac{1}{2}\right)\ln(a+y), \tag{5.2}$$

$$\mathrm{g}_2(y) = \frac{1}{\sqrt{2F(a+y)}} + \frac{(F-\frac{1}{2})}{a+y}, \quad \mathrm{g'}_2(y) = -\frac{1}{\sqrt{8F(a+y)^3}} - \frac{(F-\frac{1}{2})}{(a+y)^2},$$

$$\mathrm{G}_3(y) = \ln(a+y), \quad \mathrm{g}_3(y) = \frac{1}{a+y}, \quad \mathrm{g'}_3(y) = -\frac{1}{(a+y)^2}, \tag{5.3}$$

kde a je volitelná konstanta, která byl opět zavedena tak, aby $a \approx 0{,}1$ [10].

Jak z pseudokódu vyplývá, je tato metoda velmi podobná již představené metodě Fast fixed-point ICA. Stejné je schéma celé separace, některé použité nelinearity i způsob odstranění amplitudové nejistoty v jednotlivých pásmech. Některé vztahy týkající se separačních matic $\mathbf{V}^f$ doznaly jen malých změn plynoucí z faktu, že tato metoda při separaci nepracuje s Hermitovskou transpozicí separačních matic.

Zásadní rozdíl je ve způsobu, jakým obě metody pracují s frekvenčními pásmy $\mathbf{x}^f$. Zatímco ICA tato pásma prochází postupně a vždy separuje jedno konkrétní pásmo bez ohledu na ostatní, IVA pracuje vždy se všemi frekvenčními pásmy zároveň. Z tohoto důvodu bylo nutné změnit podmínku konvergence algoritmu.

5.2 Změna podmínky dostatečné konvergence

U metod ICA se jako podmínka dostatečné konvergence používalo vyjádření změny separační matice mezi iteracemi. Pokud byla tato změna dostatečně malá, došlo k ukončení iterování algoritmu. Metody ICA byly totiž aplikovány na jednotlivá pásma f bez znalosti ostatních pásem a samotné iterování tak bylo možné ukončit pouze na základě změn konkrétní matice $\mathbf{V}^f$, jak je ukázáno v kapitole 2. Díky tomu se při separování určitého pásma provede pouze tolik iterací, kolik je v daném pásmu skutečně nutné, a tím dochází k snížení časové náročnosti celé separace, než pokud by se každé pásmo separovalo stejným a předem daným počtem iterací.

Metody IVA na frekvenční pásma plně rozdělit nelze a daná iterace se stává z aktualizace separačních matic $\mathbf{V}^f$ pro všechna pásma, takže v každém pásmu bude provedeno vždy stejné množství iterací. Výhodnější, než sledovat jednotlivé separační matice, je pracovat s účelovou funkcí G vystihující úroveň separace. V praxi se využívá relativní změna této funkce mezi iteracemi a pokud bude nižší, než zadaná hodnota ϵ, dojde k ukončení iterací.

Následkem této změny může dojít k tomu, že pro některá pásma bude výsledný počet iterací nedostatečný a separace v nich nebude důsledná. Z principu půjde o pásma, která se nedostatečně projeví ve změně účelové funkce G (například pásma s nízkou energií). Neprojeví se tak ani ve výsledných audiosignálech a jejich nedostatečnou separaci lze ignorovat.

Pseudokód 4 Algoritmus „Fast fixed-point IVA“ pro slepou separaci řečových signálů

$\mathbf{x}_s^f \coloneqq \mathrm{STFT}\left[\mathbf{x}_s\right]$
for $f = 1, \ldots, F$ **do**
  $\left[\mathbf{U}_m^f, \mathbf{\Sigma}_m^f\right] \coloneqq \mathrm{PCA}\left[\mathbf{x}_s^f, m\right]$
  $\mathbf{x}^f \coloneqq {\mathbf{\Sigma}_m^f}^{-\frac{1}{2}} {\mathbf{U}_m^f}^{\mathrm{T}} \left(\mathbf{x}_s^f - \bar{\mu}_{\mathbf{x}_s^f}\right)$
  $\mathbf{V}^f \coloneqq \mathbf{I}_m$
End for
$obj \coloneqq \infty$ ▷ účelová funkce
repeat
  for $f = 1, \ldots, F$ **do**
    $\mathbf{y}^f \coloneqq \mathbf{V}^f \mathbf{x}^f$ ▷ separace
  End for
  $gy \coloneqq \mathrm{g}\left(\sum_{f=1}^{F} \left|\mathbf{y}^f\right|^2\right)$ ▷ nelinearity
  $dgy \coloneqq \mathrm{g}'\left(\sum_{f=1}^{F} \left|\mathbf{y}^f\right|^2\right)$
  for $f = 1, \ldots, F$ **do**
    $\mathbf{V}^f \coloneqq \mathrm{diag}\left[\mathrm{E}\left[gy + \left|\mathbf{y}^f\right|^2 \circ dgy\right]\right] \mathbf{V}^f - \mathrm{E}\left[\left(gy \circ \mathbf{y}^f\right) {\mathbf{x}^f}^{\mathrm{H}}\right]$
    $\mathbf{V}^f \coloneqq \left(\mathbf{V}^f {\mathbf{V}^f}^{\mathrm{H}}\right)^{-\frac{1}{2}} \mathbf{V}^f$ ▷ ortogonalizace
  End for
  $obj_o \coloneqq obj$
  $obj \coloneqq \sum_{i=1}^{m} \sum_{t_s} \mathrm{G}\left(\sum_{f=1}^{F} \left|\mathbf{y}^f\right|^2\right)$
until $\frac{obj_o - obj}{|obj|} < \epsilon$ ▷ podmínka konvergence
for $f = 1, \ldots, F$ **do**
  $\mathbf{V}_s^f \coloneqq \mathbf{V}^f {\mathbf{\Sigma}_m^f}^{-\frac{1}{2}} {\mathbf{U}_m^f}^{\mathrm{T}}$
  $\mathbf{V}_s^f \coloneqq \mathrm{diag}\left[{\mathbf{V}_s^f}^{\dagger}\right] \mathbf{V}_s^f$ ▷ odstranění amplitudové nejistoty
  $\mathbf{y}^f \coloneqq \mathbf{V}_s^f \mathbf{x}_s^f$ ▷ finální separace
End for
$\mathbf{y} \coloneqq \mathrm{ISTFT}\left[\mathbf{y}^f\right]$

Kapitola 6

Implementace algoritmů

Metody Fast fixed-point ICA a Fast fixed-point IVA byly implementovány v prostředí Matlab. Samotné implementace (až na drobnou změnu) přímo odpovídají pseudokódům 3 a 4 popsaným v předchozích dvou kapitolách.

Změna v implementaci se týká ukončení iterování separačních algoritmů. Metody jsou navrženy tak, že k ukončení iterací dojde po splnění konvergenčního kritéria. Tedy, že změna separační matice v případě metody ICA nebo relativní změna objektivní funkce v případě IVA jsou menší, než zvolená konstanta $\epsilon \approx 0$. V praxi by ale separační metoda v určitých speciálních případech (například při hrubém porušení jejích předpokladů) nemusela dostatečně zkonvergovat a algoritmus by nikdy nebyl ukončen. Proto jsou algoritmy navrženy tak, aby bylo možné zadat maximální počet iterací metody a pokud je tento počet překročen, dojde k ukončení algoritmu bez ohledu na splnění podmínky konvergence.

Výsledkem implementace jsou dvě funkce `BSS_Fast_ICA` a `BSS_Fast_IVA` realizující metody Fast fixed-point ICA, respektive Fast fixed-point IVA. Výstupem funkcí jsou vždy akustické signály $\mathbf{y}$, pro které by mělo platit

$$\mathbf{y} \approx \mathbf{PDs}, \tag{6.1}$$

kde matice $\mathbf{P}$ a $\mathbf{D}$ vyjadřují neřešitelnou nejistotu v pořadí a v amplitudě na úrovni výsledných signálů v časové oblasti. Tyto nejistoty budou řešeny pouze v případě vyhodnocování kvality separace, kdy jsou známé zdrojové signály $\mathbf{s}$, způsobem popsaným v následující kapitole.

Vstupem do obou funkcí jsou pozorované akustické signály $\mathbf{x}_s$, vzorkovací frekvence použitá při jejich nahrávání a zvolená šířka pásma STFT. Dalšími nepovinnými vstupy jsou zamýšlený počet signálů m po redukci dimenzionality pomocí PCA, maximální počet iterací dané metody, toleranční konstanta ϵ pro určení dostatečné konvergence a výběr nelinearity g používané separačním algoritmem. Konkrétní názvy vstupních argumentů, jejich rozměry a přesné hodnoty, kterých smějí nabývat, jsou popsány v nápovědách funkcí, jejichž kompletní kód je dostupný v přílohách této práce.

Zatímco zmíněné parametry separačních metod je možné zvolit pomocí vstupních parametrů implementovaných funkcí, existují další parametry, které

pro separační metody nejsou příliš kritické a byly zvoleny fixně. Konstanta a, používaná v rámci nelinearit g, byla zvolena jako $a = 0{,}1$, a sice na základě zmíněné podmínky $a \approx 0{,}1$. Pro realizaci STFT a ISTFT bylo zvoleno Hammingovo okno $w(n)$ a překryv byl nastaven na 75 %, což odpovídá hodnotám, se kterými se lze při zpracování řeči běžně setkat.

Kromě samotných separačních metod bylo nutné implementovat i další podpůrné algoritmy. Kde to bylo možné, byly využité volně dostupné implementace nebo zabudované funkce Matlabu. Jedná se například o Maďarský algoritmus (používaný při správném třídění frekvenčních pásem po separování metodou ICA), rozklad signálů do frekvenčních pásem pomocí STFT, nebo provedení SVD rozkladu matice (využíván v rámci PCA a redukce dimenzionality). Implementovat proto bylo nutné pouze zpětnou syntézu signálů pomocí ISTFT, k čemuž byla využita inverzní Fourierova transformace v krátkých váhovaných segmentech s překryvem [13]. Některé z těchto podpůrných algoritmů se nacházejí v separátních funkcích, které hlavní funkce `BSS_Fast_ICA` a `BSS_Fast_IVA` využívají. I tyto podpůrné funkce jsou (včetně svého kódu a nápovědy s popisem vstupů a výstupů) součástí příloh této práce.

Na závěr byly vytvořeny obslužné skripty sloužící k ověření implementovaných funkcí. Tyto skripty vytvářejí konvoluční směsi řečových signálů simulováním místnosti, zpracovávají naměřená data ze směsí v místnosti reálné, pomocí implementovaných funkcí provádějí slepou separaci a následně kvalitu separace vyhodnocují. Způsob, jakým to dělají, je popsán v následující kapitole.

Kapitola 7

Ověření algoritmů

Za účelem ověření a vyhodnocení úspěšnosti implementovaných separačních algoritmů bude provedena série experimentů v místnosti simulované (s 2 až 4 mluvčími) a v reálné (s pouze 2 mluvčími). Samotný experiment se skládá z rozmístění zdrojů (mluvčích) a senzorů (mikrofonů) na předem daná místa v místnosti. Mluvčí následně pronesou zdrojové signály $\mathbf{s}$, a ty se v místnosti smísí v pozorované signály $\mathbf{x}_s$, které zaznamenají mikrofony. Pozorované signály jsou implementovanými metodami slepě separovány v signály $\mathbf{y}$.

Na základě signálů $\mathbf{y}$ je vyhodnocena kvalita separace, tedy úroveň potlačení promluv okolních řečníků. V případě známých zdrojových signálů bude kvalita separace určena za pomoci objektivních metrik SNR (odstup signál - šum) a kepstrální vzdálenost. Při posledním experimentu bude v reálné místnosti vytvořen skutečný „Cocktail Party Effect“, zdrojové signály budou plně neznámé a kvalita separace bude vyhodnocena jednoduchým poslechovým testem.

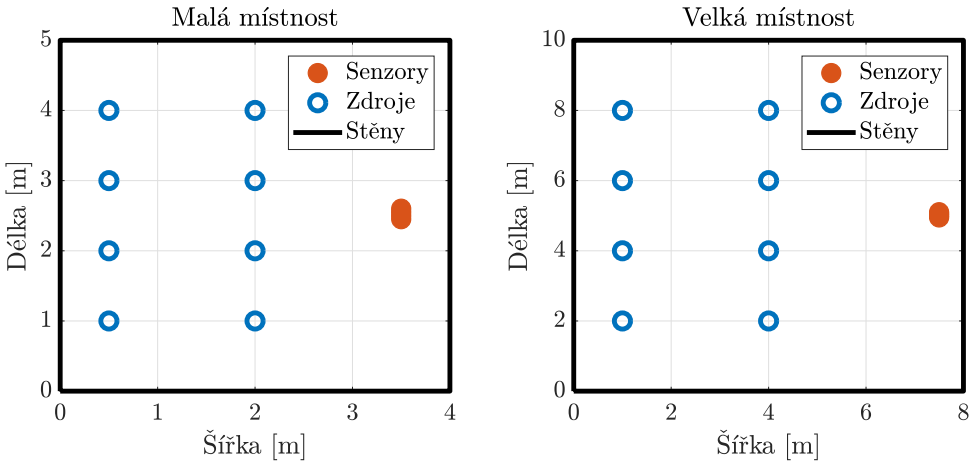
7.1 Experimenty v simulované místnosti

Základní sada experimentů bude probíhat v simulované místnosti, kde je možné snadno a rychle měnit parametry experimentu, jako jsou rozměry místnosti, odrazivost stěn nebo pozice zdrojů a senzorů promluv. Intuitivně je kvalita výsledné separace všemi těmito parametry ovlivněna. Proto je cílem provést takových experimentů větší množství (vždy pro rozdílné parametry místnosti), pro každý z nich vyjádřit kvalitu separace pomocí metrik SNR a kepstrální vzdálenost, ty následně zprůměrovat a tím vyjádřit odhad skutečné kvality implementovaných separačních metod.

Pro simulovanou místnost je uvažován model konvoluční směsi signálů popsaný rovnicí (3.1). Simulování místnosti proto spočívá v nalezení tvarů impulsních odezev $a_{ij}(n)$ cest ze zdroje j na mikrofon i. Za tím účelem je využita metoda „image - source“ [14], která je již pro Matlab implementována [15] a tato dostupná implementace bude přímo využita.

Před využitím této implementace je nutné zvolit rozměry simulované místnosti, pozice zdrojů a senzorů, odrazivosti stěn, dobu do určitého poklesu akustických odrazů, vzorkovací frekvenci, a trvání impulsní odezvy. Jak již bylo řečeno, cílem bude provést větší množství experimentů s různými parametry místnosti. Pro opravdu důkladné a komplexní ověření separačních metod by bylo nutné simulovat separaci v dostatečném množství místností s různými vzájemnými kombinacemi parametrů. Vzhledem k množství těchto parametrů a trvání separace je ale takový přístup mimo možnosti této práce. Bylo rozhodnuto, že většina parametrů bude pro všechny experimenty fixní a měnit se budou pouze parametry ovlivňující geometrické uspořádání místnosti, tedy její rozměry a vzájemné pozice zdrojů a senzorů.

Odrazivost svislých zdí je nastavena na 0,5 (hodnota 0 odpovídá pohltivému materiálu a 1 pak materiálu s dokonalou odrazivostí), v případě podlahy 0,3 (simulace koberce) a pro strop 0,7 (často bývá z jiného materiálu, než okolní zdivo). Trvání odrazů bylo zvoleno jako 100 ms, přičemž tato doba vyjadřuje čas, po kterém poklesne úroveň odrazů o 60 dB oproti původnímu signálu. Trvání impulsní odezvy bylo stanoveno taktéž na 100 ms a vzorkovací frekvence experimentů v simulované místnosti bude 16 kHz. Všechny tyto konkrétní hodnoty byly zvoleny buď na základě implicitních parametrů použité implementace [15], nebo dle již proběhlých experimentů popsaných v [8].

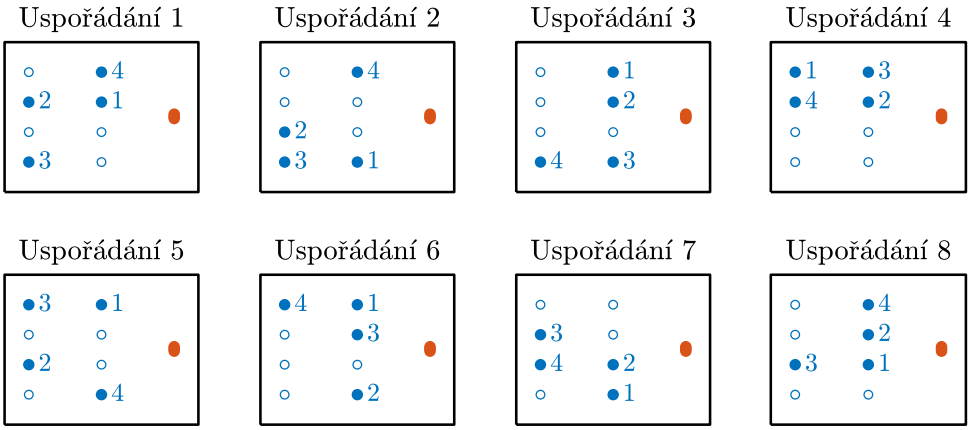


Obrázek 7.1: Nákres rozmístění zdrojů a senzorů v simulovaných místnostech

Rozměry místnosti jsou zvoleny buď jako „malé“ (šířka 4 m, délka 5 m, výška 3 m), nebo jako „velké“ (šířka 8 m, délka 10 m, výška 3 m). Jednotliví mluvčí jsou vždy náhodně rozmístěni na některou z předem vybraných pozic v místnosti. Senzory jsou sdruženy v krátkém mikrofonovém poli o stejném počtu mikrofonů jako řečníků, jak je to při slepé separaci v simulované místnosti (kde nejsou přítomny hluky a šumy okolí) běžné [8]. V rámci jednoho konkrétního rozmístění řečníků je vliv umístění tohoto pole na kvalitu separace sice velmi značný. Dostatečné množství pokusů s náhodným rozmístěním řečníků ale ukázalo, že pozice mikrofonového pole finální průměrnou kvalitu separace příliš neovlivní. Naopak na ni má značný vliv rozteč mikrofonů a jako

ideální byla empiricky zvolena vzdálenost 5 cm. Celé mikrofonové pole se pak fixně nachází v ose místnosti ve vzdálenosti 0,5 m od pravé stěny. Půdorys uspořádání malé i velké místnosti je patrný z obrázku 7.1. Výška zdrojů je vždy 1,7 m a v případě senzorů 1,5 m.

Pro obě místnosti bylo náhodně zvoleno 8 stejných vzorů pro výběr pozic řečníků při jednotlivých experimentech. Vzniklá uspořádání jsou patrná z nákresu na obrázku 7.2. Čísla u pozic identifikují řečníka stojícího na daném místě, přičemž daný řečník pronáší ve všech pokusných uspořádaních stále stejnou promluvu. Všichni 4 řečníci se ale mezi sebou pochopitelně liší, a to jak věkem a pohlavím, tak obsahem promluvy. Vždy se ale jedná o zhruba 40 s dlouhý souvislý proslov, který je vycentrován a amplitudově nanormován tak, aby všechny promluvy měly přibližně stejnou úroveň hlasitosti. Promluvy byly získány ze studiových nahrávek Českého rozhlasu a mají proto velmi dobrou kvalitu. Pokud je simulována místnost o pouze 3 řečnících, je řečník číslo 4 a jeden mikrofon ze všech pokusů odstraněn. Analogicky se postupuje při 2 řečnících.



Obrázek 7.2: Nákres uspořádání řečníků při jednotlivých experimentech

Pro všechna znázorněná uspořádání jsou simulovány místnosti obou velikostí a to vždy o 2, 3 i 4 řečnících. Celkem tak vznikne 48 modelových situací. V rámci jedné modelové situace jsou jasně dány všechny parametry pro výpočet impulsních odezev $a_{ij}(n)$ pomocí implementace metody „image - source“. Po výpočtu odezev je určena jejich konvoluce s patřičnými promluvami (zdrojovými signály s_j). Vzniklé konvoluce jsou následně na jednotlivých senzorech sečteny do pozorovaných signálů $\mathbf{x}_s$, čímž je simulováno nahrání vzniklé směsi mikrofony. Celkem vznikne 48 směsí k odseparování.

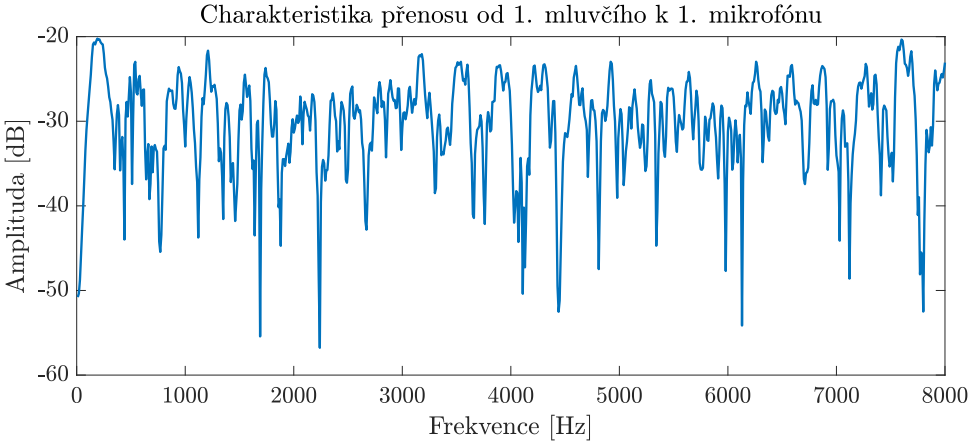
V reálném případě by se k pozorovaným signálům x_i na senzorech přičetly i nežádoucí akustické signály, jako je šum mikrofonu a různé hluky okolí (troubení aut, zvuky větru nebo hluk strojů v budově). V případě simulované místnosti nebudou tyto vlivy zahrnuty, neboť cílem je ověřit schopnosti implementovaných metod, které ve svých základních předpokladech tyto vlivy vylučují. Naopak při měření v reálné místnosti se tyto nežádoucí jevy projeví a bude zajímavé sledovat, nakolik to ovlivní kvalitu separace.

Vzniklé směsi jsou separovány implementovanými metodami ICA i IVA se stejnými parametry. Protože je cílem dosáhnout co nejlepší separace a její trvání není nutné optimalizovat, je konvergenční konstanta ϵ u obou metod nastavena na 0 a obě metody provedou předem stanovený počet iterací, konkrétně 200, což by u obou použitých metod mělo být více než dostatečné [8] [10]. Použité nelinearity (kontrasty) g sice kvalitu separace příliš neovlivňují, empiricky se ale v tomto případě ukázalo, že o něco lepších výsledků dosahují obě metody při použití shodné nelinearity ve tvaru

$$\mathrm{G}(y) = \ln(a+y), \quad \mathrm{g}(y) = \frac{1}{a+y}, \quad \mathrm{g}'(y) = -\frac{1}{(a+y)^2}, \tag{7.1}$$

kde a bylo v rámci implementace algoritmů zvoleno jako $a = 0{,}1$.

Šířka pásma (při aplikaci STFT na řečové signály) naopak kvalitu separace ovlivňuje značně. Pro představu vlivu šířky pásma je na obrázku 7.3 uveden příklad amplitudové frekvenční charakteristiky jednoho z přenosů v simulované místnosti. Protože je charakteristika poměrně proměnlivá, je nutné zvolit frekvenční pásma dostatečně úzká na to, aby v nich byla frekvenční charakteristika přibližně konstantní a bylo možné v těchto pásmech uvažovat (s přijatelnou nepřesností) lineární model směsi signálu. Na druhou stranu nemohou být pásma úzká příliš, protože v takovém případě by po aplikaci STFT došlo k hrubšímu rozlišení v časové oblasti a jednotlivá frekvenční pásma by obsahovala příliš málo vzorků na to, aby na jejich základě mohla být pásma spolehlivě separována. V praxi pak lze pro konkrétní mixáž vždy nalézt určitou kompromisní šířku pásma, pro kterou je kvalita separace nejlepší. Na základě dosažených výsledků s různými šířkami pásma se v tomto případě zcela jednoznačně jednalo o hodnotu 4 Hz, se kterou proto budou všechny simulované směsi separovány.



Obrázek 7.3: Amplitudová frekvenční charakteristika přenosu mezi první pozicí a prvním mikrofonem v malé místnosti

Po separování všech 48 simulovaných směsí metodami ICA a IVA s uvedenými parametry je pomocí metrik SNR a kepstrální vzdálenost vyhodnocena kvalita separace. Výsledné hodnoty jsou uvedeny v kapitole 8.

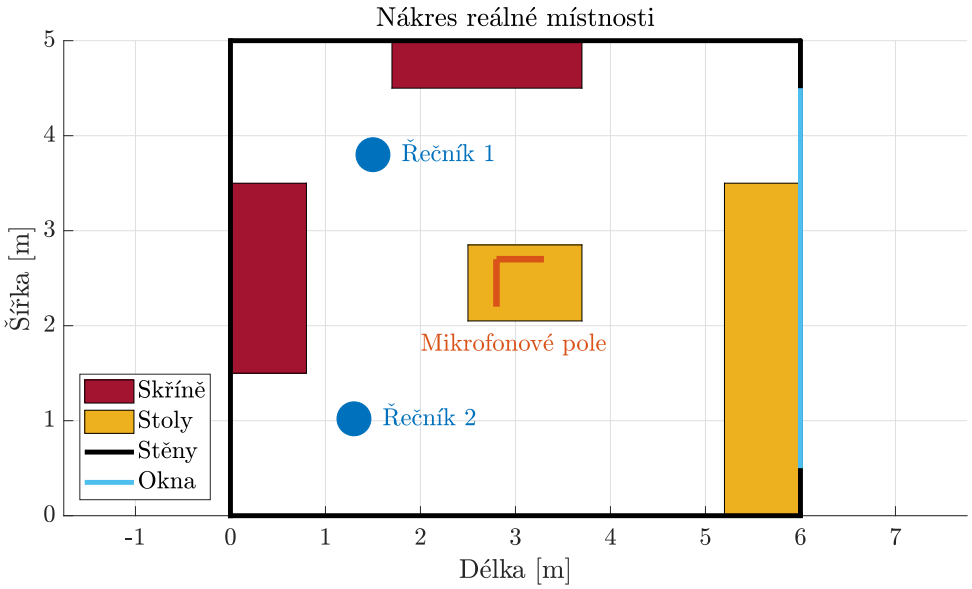
7.2 Měření v reálné místnosti

Simulovaná místnost sice díky své jednoduchosti nabídla ideální prostředí pro základní vyhodnocení separačních algoritmů, v praxi budou ale tyto algoritmy nasazovány v místnostech skutečných. Oproti simulovaným místnostem v nich budou přítomny i různé šumy a hluky a šíření zvuku bude ovlivněno například nábytkem, okny nebo i těly samotných mluvčích. Všechny tyto aspekty by sice do určité míry šly simulovat také, taková simulace by už ale byla natolik komplexní, že je výhodnější provést přímá měření v místnosti reálné, přičemž v případě této práce se měření budou účastnit 2 řečníci.

K nahrávání bylo v tomto případě použito celé mikrofonové pole se 40 takzvanými MEMS mikrofony, přičemž data z nich budou v rámci představených separačních metod pomocí PCA zredukována na počet m odpovídající počtu mluvčích, tedy $m = 2$. Celého mikrofonového pole je zde využito zejména z důvodu přítomnosti šumů a hluků okolí, jejichž přítomnost odporuje předpokladům použitých metod analýzy nezávislých komponent (popsaným v kapitole 2), což by vedlo k separaci nedokonalé nebo i nemožné. Díky využití dat z velkého množství mikrofonů a následné redukci dimenzionality pomocí PCA budou signály $\mathbf{x}$ částečně odšuměné a jejich následná separace bude díky tomu kvalitnější [5].

Na použití mikrofonového pole lze nahlížet i tak, že přítomné nežádoucí šumy a hluky je možné modelovat jako další neznámé zdrojové signály s_j. Na rozdíl od řečových signálů ale tyto signály nevychází fyzicky z jednoho místa a jejich zdroje jsou spíše rozprostřené (například vibrace strojů v budově přenášené do místnosti). Do určité míry je ale možné zdroj těchto signálů modelovat jako velké množství různě rozprostřených zdrojů s_j, čemuž odpovídá použití velkého množství mikrofonů. V rámci provedené PCA pak dojde k potlačení méně výrazných hlavních komponent zaznamenaných signálů, přičemž v ideálním případě dojde i k potlačení zaznamenaných hluků a šumů. Ponechány budou poté pouze nejsilnější komponenty odpovídající promluvám řečníků. Pokud by ale bylo odseparováno promluv více (hodnota m by při redukci dimenzionality byla zvolena větší, než byl počet řečníků), další odseparované signály by obsahovaly právě směs šumů a hluků, což ostatně bylo v rámci jednoduchých experimentů nad rámec této kapitoly ověřeno. Této vlastnosti se využívá zejména v případě takzvané extrakce signálů (ICE nebo IVE), kdy je pomocí velkého mikrofonového pole separována (extrahována) pouze jedna promluva, čímž dojde k potlačení šumového pozadí.

V rámci použitého mikrofonového pole byla vzdálenost mezi mikrofony opět 5 cm, celé pole mělo tvar písmene L a nacházelo se na stole přibližně uprostřed místnosti ve výšce zhruba 1 m. Samotná místnost byla široká přibližně 5 m, dlouhá 6 m a vysoká 3 m. Řečníci stáli u sousedních rohů místnosti a byli vysocí zhruba 180 cm. V místnosti se také nacházel nábytek a jedna ze stěn je z velké části tvořena okny. Celková kompozice reálné místnosti při nahrávání je nejlépe patrná z obrázku 7.4. Všechny nahrávky byly pořízeny s vzorkovací frekvencí 40 kHz.



Obrázek 7.4: Hrubý nákres kompozice reálné místnosti při měření

Aby bylo možné objektivně určit kvalitu separace, je nutná znalost zdrojových signálů. Měření tedy probíhalo tak, že vždy mluvil pouze jeden mluvčí, jeho promluva se zaznamenala na celé mikrofonové pole a následně se stejným způsobem změřila promluva druhého mluvčího. Signál z libovolného mikrofonu pak představoval zdrojové signály $\mathbf{s}$. Pozorovaných signálů $\mathbf{x}_s$ bylo následně uměle dosaženo součtem obou záznamů z celého mikrofonového pole.

Finální experiment představoval skutečný „Cocktail Party Effect“, kdy řečníci skutečně mluvili naráz a mikrofonové pole zaznamenalo vzniklou směs. V tomto případě pochopitelně zdrojové signály k dispozici nejsou a kvalitu separace bude nutné hodnotit subjektivně jednoduchým poslechovým testem.

Oba řečníci pronášeli vždy souvislé promluvy přibližně 60 až 90 s dlouhé. Během měření občas docházelo k poryvům větru, tlučení deště o okno a ve zvukových záznamech byl také patrný poměrně značný šum mikrofonů a okolí. Cílem měření nebylo tyto vlivy potlačovat, protože dokonalé podmínky měření nastaly už při experimentech v simulované místnosti, ale naopak vlastností reálného prostředí plně využít při ověření účinnosti implementovaných separačních metod.

Při experimentech v reálné místnosti vznikly celkem 2 směsi. Ty byly separovány implementovanými metodami ICA i IVA s fixně nastavenými 200 iteracemi. Jako použitá nelinearita byla opět zvolena funkce daná rovnicí (7.1). Separace byly provedeny pro vícero šířek pásma STFT, konkrétně 0,5 Hz, 1 Hz, 2 Hz a 4 Hz. Výsledky separace jsou shrnuty v kapitole 8.

Více experimentů v reálné místnosti nebylo provedeno především z důvodu velké časové náročnosti implementovaných separačních metod. Při použité vzorkovací frekvenci 40 kHz a při zvolených šířkách pásma STFT vznikne velké množství pásem k odseparování. Promluvy jsou navíc už poměrně dlouhé

a separační metody musí provést celých 200 iterací. To činí z odseparování jedné směsi pro průměrně výkonný počítač problém na desítky minut, někdy i hodinu. Výpočetní náročnost by samozřejmě šlo velmi snadno snížit, například podvzorkováním signálu nebo provedením menšího počtu iterací. Cílem experimentů v reálné místnosti ale není provést velké množství pokusů, na základě kterých by bylo možné separační metody dobře vyhodnotit. K tomu sloužila místnost simulovaná. Zde je cílem provést separaci co nejkvalitnější (proto nebylo k podvzorkování, zkrácení promluv nebo k snížení počtu iterací přistoupeno) a spíše ověřit, že implementované metody jsou funkční i v prostředí reálné místnosti. K tomu stačí provést dva experimenty, kdy jeden bude hodnocen objektivními metrikami a druhý poslechovým testem.

7.3 Objektivní metriky kvality separace

Pro objektivní vyhodnocení kvality separace budou využity metriky SNR a kepstrální vzdálenost. Obě budou odhadovat poměr množství patřičného zdrojového signálu s_j v jeho odseparovaném odhadu y_i ku množství ostatních zdrojových signálů v signálu y_i. Čím větší tento poměr bude, tím lépe se při separaci podařilo potlačit promluvy ostatních mluvčích a tím kvalitnější celá separace byla. Než ale bude možné tyto metriky aplikovat na signály, je nutné provést určité předzpracování separovaných signálů $\mathbf{y}$.

Metriky budou určovány vždy pro jednotlivé dvojice signálů s_j a y_i. Tedy například v případě separace směsi 3 mluvčích bude kvalita separace vyjádřena 3 hodnotami jedné metriky. Tento postup vyžaduje, aby signál y_i byl vždy odhadem promluvy příslušného řečníka s_j. Jak je ale popsáno v sekci 3.4, jednotlivé signály y_i jsou ve vektoru $\mathbf{y}$ setříděné náhodně a navíc jsou ztížené nejistotou v amplitudě. Amplitudová nejistota řešena nebude a metriky budou použity takovým způsobem, kdy na amplitudě signálů nezáleží. Permutační nejistotu, tedy správné setřídění signálů y_i, ale bude nutné vyřešit.

Signály y_i se třídí na základě podobnosti jejich obálek se známými zdrojovými signály s_j. Obálky jsou získávány pomocí zabudovaných funkcí Matlabu, jejich podobnost je určena korelací a ideální setřídění je nalezeno pomocí Maďarského algoritmu. Tento postup zde není popsán do úplnosti, protože se jedná o princip představený v sekci 4.3, kde bylo důkladně popsáno správné třídění frekvenčních pásem po separaci pomocí metod ICA. Výsledkem je správně setříděný vektor $\mathbf{y}$, kde jednotlivé prvky y_j představují odhady signálů s_j pro $j = 1, \ldots, m$.

Dalším problémem jsou vzájemná zpoždění signálů s_j a y_j vzniklá v důsledku různých zpoždění přenosových cest mezi řečníky a mikrofony. Pro dobré porovnání podobnosti signálů je vhodné tato zpoždění eliminovat. Hodnota zpoždění je pro každou dvojici s_j, y_j určena pomocí jejich korelační funkce. Pozice jejího globálního extrému pak odpovídá hledanému zpoždění (ve vzorcích) mezi oběma signály a signály jsou následně o toto zpoždění posunuty tak, aby mezi nimi žádné zpoždění nebylo.

Před výpočtem metrik budou také všechny signály $\mathbf{s}$ i $\mathbf{y}$ z obou stran o 0,5 s oříznuty. Děje se tak z důvodu aplikování ISTFT při syntéze signálů $\mathbf{y}$, kvůli čemuž se na jejich okrajích projeví tvar použitého okna. Dále budou všechny signály vycentrovány a amplitudově nanormovány na přibližně stejnou úroveň hlasitosti. Centrování je provedeno pouze pro jistotu, aby střední hodnota signálů zbytečně neovlivnila použité metriky a vyrovnání hlasitostí pomáhá při případném poslechu signálů.

Jako první bude představena známá a hojně využívaná metrika SNR. V každém odseparovaném signálu y_j se nachází určitá množství zdrojového signálu s_j, ostatních (nechtěných) zdrojových signálů a také neznámého šumu a hluku okolí e_j. Signál y_j je navíc zatížen amplitudovou nejistotou, kterou lze vyjádřit neznámou konstantou d_j. To zle zapsat jako

$$y_j = d_j \left(b_{j1} s_1 + \cdots + b_{jj} s_j + \cdots + b_{jm} s_m + e_j \right), \tag{7.2}$$

kde b_{ji} jsou reálné konstanty vyjadřující množství signálu s_i v signálu y_j.

Metrika SNR určuje odstup signál-šum. V tomto případě ale šumem není myšlen signál e_j, protože cílem základní slepé separace (kdy se nejedná o úlohu extrakce nebo jiné modifikace) není potlačit signál e_j, ale promluvy s_i ostatních řečníků. Právě tyto promluvy jsou chápány jako šum a hodnota SNR má tvar

$$\mathrm{SNR}_j = 10 \log \left(\frac{b_{jj}^2}{\frac{1}{m-1} \sum_{i \neq j} b_{ji}^2} \right) [\mathrm{dB}]. \tag{7.3}$$

Hodnota SNR_j přísluší odseparované promluvě y_j a vyjadřuje, jaký je v této promluvě poměr mezi množstvím separované promluvy s_j ku průměrnému množství ostatních nežádoucích promluv.

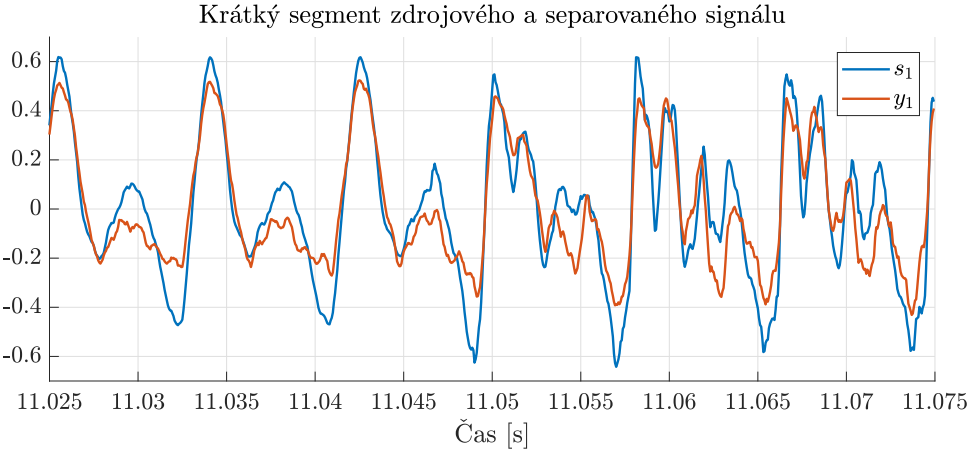
Jednotlivé koeficienty b_{ji} lze určit tak, že se rovnice (7.2) vynásobí signálem s_i a následně se na rovnici aplikuje operátor střední hodnoty E, přičemž se využije linearity tohoto operátoru a faktu, že výraz $\mathrm{E}\left[s_j s_i\right]$ nabývá pro $i \neq j$ nulové hodnoty, protože tyto signály jsou nezávislé (jedná se o různé promluvy pronášené různými řečníky). Stejně tak je předpokládána nezávislost všech promluv s_i na šumech e_j, protože ty se stávají z hluků okolí a šumu samotných mikrofonů. Poté bude platit

$$b_{ji} = \frac{\mathrm{E}\left[y_j s_i\right]}{d_j \mathrm{E}\left[s_i^2\right]}. \tag{7.4}$$

V koeficientech b_{ji} jsou sice nadále zahrnuty neznámé konstanty d_j, ty se ale po dosazení hodnot b_{ji} do vzorce (7.3) zkrátí a vznikne finální vztah pro určení hodnoty SNR_j jako

$$\mathrm{SNR}_j = 10 \log \left(\frac{\left(\frac{\mathrm{E}\left[y_j s_j\right]}{\mathrm{E}\left[s_j^2\right]} \right)^2}{\frac{1}{m-1} \sum_{i \neq j} \left(\frac{\mathrm{E}\left[y_j s_i\right]}{\mathrm{E}\left[s_i^2\right]} \right)^2} \right) [\mathrm{dB}]. \tag{7.5}$$

Metrika SNR pracuje přímo s časovými průběhy signálů, což může působit nepřesnosti. Pro představu jsou na obrázku 7.5 uvedeny krátké segmenty zdrojového a příslušného separovaného signálu v případě, kdy byla separace poslechově takřka dokonalá. Ačkoliv se jednalo o pokus v simulované místnosti, kde nebyl přítomen šum e_j, nejsou průběhy signálů úplně totožné. Tento jev bude pravděpodobně způsoben různými drobnými zkresleními metodami ICA, které například neřeší fázové posuvy jednotlivých pásem STFT, nebo v důsledku použití špatné nelinearity g. Kvůli tomu budou i hodnoty SNR vždy zatížené určitou chybou, i když z průběhů na obrázku 7.5 zároveň vyplývá, že tato chyba příliš markantní nebude a náhodné poslechové testy ukázaly, že hodnota SNR s kvalitou separace koresponduje dostatečně. Navíc je metrika SNR v praxi natolik využívána, že je možné její konkrétní výsledky dobře interpretovat, a i proto je její využití i přes drobné nepřesnosti vhodné.



Obrázek 7.5: Průběhy krátkých segmentů zdrojového a separovaného signálu v případě dokonalé separace

Jiný odhad kvality separace poskytne takzvaná kepstrální vzdálenost. Jedná se o metriku pracující s kepstry krátkodobých segmentů signálu takovým způsobem, že vyjádří podobnost spektrogramů dvou signálů [16]. Na rozdíl od SNR ale nebude tento způsob hodnocení kvality separace zatížen případnou chybou časových průběhů separovaných signálů.

Pro určení kepstrální vzdálenosti dvou řečových signálů je vhodné využít takzvané „mel-frekvenční kepstrum“ [16] pracující s melovskou bankou filtrů. Počet pásem této banky byl stanoven na 30 a bude určováno prvních 13 koeficientů kepstra. Segmentace signálů proběhne obdobným způsobem, jako tomu bylo při výpočtu STFT signálů a samotná kepstální vzdálenost mezi dvěma příslušnými segmenty bude počítána bez přítomnosti prvního kepstrálního koeficientu, který vyjadřuje pouze výkon signálu [16] a ten je kvůli amplitudové nejistotě separovaných signálů neznámý. Výsledná kepstrální vzdálenost $\mathrm{C_D}\,[y_j, s_i]$ signálů y_j a s_i bude průměrem kepstrálních vzdáleností všech příslušných segmentů obou signálů. Pro výpočet konkrétních hodnot budou přímo využity veřejně nedostupné podpůrné funkce k přednášce [16].

Pro vyhodnocení kvality separace se nejprve určí kepstrální vzdálenosti $\mathrm{C_D}\,[y_j, s_i]$ pro všechny dvojice zdrojových a separovaných signálů. Opět bude využit model popsaný rovnicí (7.2), kdy se v každém separovaném signálu y_j očekává nejen určité množství odpovídajícího signálu s_j, ale i nějaké množství ostatních zdrojových signálů a šumu e_j. V tomto případě bude ono „množství signálu“ vyjádřeno podobností dané dvojice signálů, tedy jejich kepstrální vzdáleností. Přítomnost šumu e_j nebude kvantifikována, a to ze stejných důvodů, jako u metriky SNR.

Cílem bude pomocí kepstrální vzdálenosti vytvořit takovou metriku, která by vyjadřovala poměr podobnosti separované promluvy s odpovídající promluvou zdrojovou ku průměrné podobnosti s ostatními zdroji, přičemž malá hodnota kepstrální vzdálenosti odpovídá velké podobnosti signálů (platí zde nepřímá úměra). Proto byl uměle vytvořen vzorec vyjadřující kvalitu separace signálu y_j parametrem Keps_j jako

$$\mathrm{Keps}_j = \frac{\frac{1}{m-1}\sum_{i\neq j}\mathrm{C_D}\,[y_j, s_i]}{\mathrm{C_D}\,[y_j, s_j]}. \tag{7.6}$$

Na rozdíl od SNR se takto určený parametr běžně nepoužívá a pro jeho snadnou interpretaci je nutné okomentovat význam hodnot, kterých může nabývat. Hodnota $\mathrm{Keps}_j = 1$ odpovídá nulové separaci (separovaný signál si je v průměru podobný se všemi zdroji stejně). Hodnota $\mathrm{Keps}_j < 1$ by odpovídala kontraproduktivní separaci, která by naopak promluvy okolních řečníků zvýraznila. Tato hodnota by ale mohla vyjít například v důsledku špatného setřídění signálů $\mathbf{y}$. V praxi budou hodnoty vždy $\mathrm{Keps}_j > 1$. Poslechově dostatečná separace (promluvy okolních řečníků jsou slyšitelné, ale potlačené) odpovídá přibližně hodnotám $\mathrm{Keps}_j > 1{,}25$. Promluvy s hodnotami $\mathrm{Keps}_j > 2$ se poslechově projevily jako odseparované takřka dokonale.

7.4 Návrh poslechového testu

Pro vyhodnocení finálního experimentu v reálné místnosti, kdy jsou zdrojové promluvy neznámé, bude navrhnut jednoduchý poslechový test. Jeho cílem bude primárně vyhodnotit míru separace, tedy posoudit, nakolik je v odseparované promluvě potlačená promluva druhého mluvčího. Separační metody budou ale odseparované promluvy i částečně zkreslovat. Děje se tak například v důsledku špatného třídění některých frekvenčních pásem při separaci metodou ICA nebo při rozkladu signálů do pásem příliš úzkých. Míra tohoto zkreslení je z pochopitelných důvodů taktéž podstatným parametrem a i ona bude poslechovým testem hodnocena.

Směs 2 promluv byla separována metodami ICA i IVA, navíc obě byly použity se 4 různými šířkami pásma STFT, konkrétně 0,5 Hz, 1 Hz, 2 Hz a 4 Hz. Výsledkem je tedy 16 akustických signálů. Protože je každá nahrávka přes minutu dlouhá a hodnotí se míra separace i zkreslení nahrávek, byl by takový poslechový test příliš náročný.

Po provedení separace provedl autor této práce rychlý poslech odseparovaných promluv a okamžitě se ukázalo, že míra separace i poslechová kvalita jsou šířkou pásma STFT značně ovlivněny. S klesající šířkou pásma sice (do určité hodnoty) roste kvalita separace, zároveň s ní ale roste i zkreslení promluv. Větší šířka pásma naopak způsobí zkreslení nízké vykoupené špatnou separací. Tento efekt byl natolik jednoznačný, že promluvy separované při šířkách pásma 0,5 Hz a 4 Hz byly z poslechového testu rovnou vyřazeny z důvodu přílišného zkreslení, respektive nedostatečné separace. V rámci poslechového testu tedy budou hodnoceny promluvy separované při šířkách pásem 1 Hz a 2 Hz, díky čemuž došlo ke snížení počtu testovaných nahrávek z 16 na 8.

Odseparované signály byly uloženy jako mono stopy do souboru `*.wav` s vzorkovací frekvencí 40 kHz. Před uložením byly signály vycentrovány a amplitudově nanormovány, aby měly přibližně stejnou úroveň hlasitosti a při přehrávání nedocházelo k saturaci reproduktorů. Každý účastník testu poslouchal uložené nahrávky na stejných sluchátkách s předem nastavenou fixní hlasitostí, aby se co nejvíce eliminoval vliv audiosoustavy na rozdílné vnímání nahrávek jednotlivými účastníky testu.

K 8 odseparovaným nahrávkám byl přidán i záznam směsi promluv před separací, který sloužil jako určitý příklad nulové separace a nulového zkreslení signálu separačními metodami. Tento záznam byl jako jediný zřetelně označen, ale ostatních 8 nahrávek bylo pojmenováno náhodnými písmeny, aby účastníci experimentu nevěděli, za jakých podmínek byly separovány.

Poslechového testu se postupně účastnilo 6 lidí. Jednotlivé nahrávky si pouštěli sami, takže si mohli libovolně kontrolovat jejich trvání, pořadí a případná opakování. Úkolem bylo si všechny nahrávky dostatečně poslechnout a při tom je ohodnotit z hlediska míry separace a zkreslení.

Hodnocení bylo vyjádřeno procenty. V případě kvality separace odpovídala hodnota 0 % nulové separaci (na úrovni kontrolní směsi před separací) a 100 % představovalo separaci dokonalou (druhý mluvčí není slyšitelný vůbec). Při hodnocení míry zkreslení odpovídala hodnota 100 % zkreslení nulovému (na úrovni kontrolní směsi) a hodnota 0 % by pak teoreticky odpovídalo maximálnímu zkreslení, kdy by promluvy nebyly vůbec srozumitelné. Zde se pochopitelně projeví různé měřítko míry zkreslení mezi účastníky testu, a proto budou v této kategorii spíš než konkrétní procentuální hodnoty zajímavé vzájemné pořadí a odstupy jednotlivých promluv. Výsledky testu jsou shrnuty v kapitole 8.

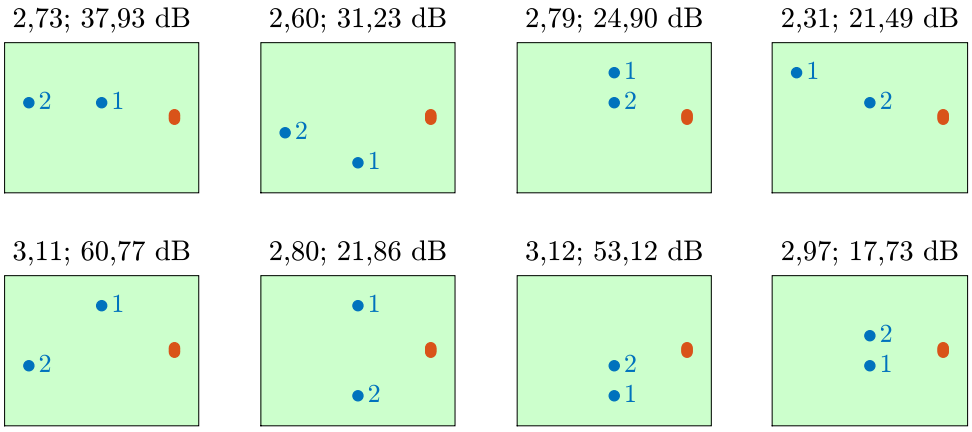
Kapitola 8

Výsledky

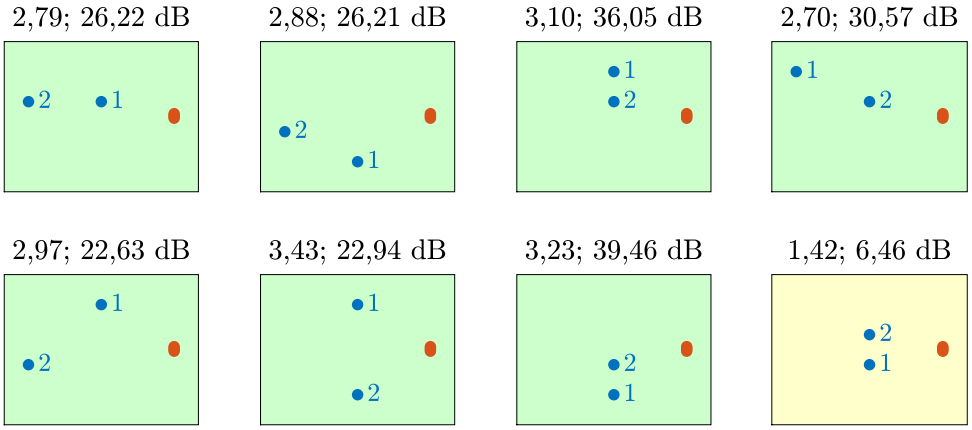
8.1 Výsledky experimentů v simulované místnosti

Simulací malé a velké místnosti vzniklo vždy po 48 směsích. Ty byly odseparovány za použití metody ICA i IVA, jak je popsáno v kapitole 7. Výsledkem separace je vždy takové množství promluv, jaké se v místnosti nacházelo simulovaných řečníků. Pro každou odseparovanou promluvu y_j jsou vypočítána měřítka kvality Keps_j a SNR_j. Ty jsou následně v rámci jedné modelové situace (konkrétní uspořádání lidí v místnosti) mezi jednotlivými promluvami zprůměrovány. Výsledné hodnoty Keps a SNR jsou (v tomto pořadí) spolu s nákresem příslušné modelové situace vykresleny v následujících 12 obrázcích.

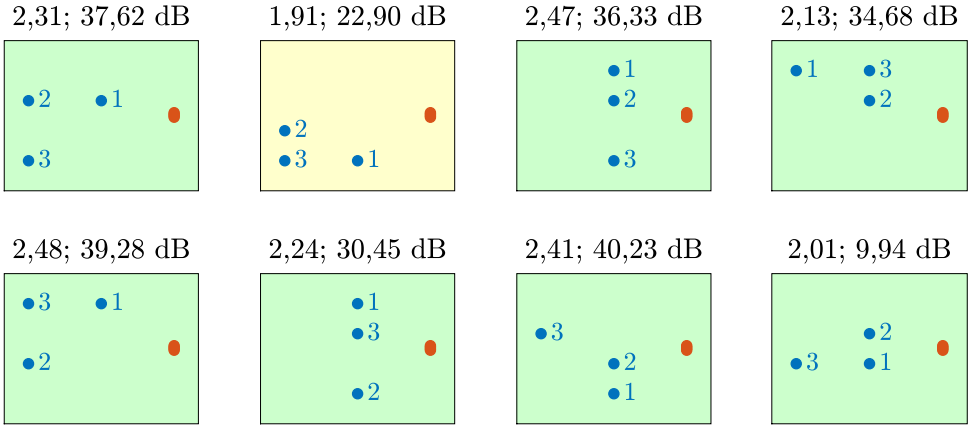
Barevné pozadí nákresů jednotlivých situací se řídí danou hodnotou Keps a je rozděleno do tří kategorií. Zelené pozadí odpovídá poslechově takřka dokonalé separaci, kdy byly Keps > 2. Žluté pozadí je v případě poslechově dostatečné separace, tedy při Keps $> 1{,}25$. Nedostatečná separace je vyobrazena červeně.



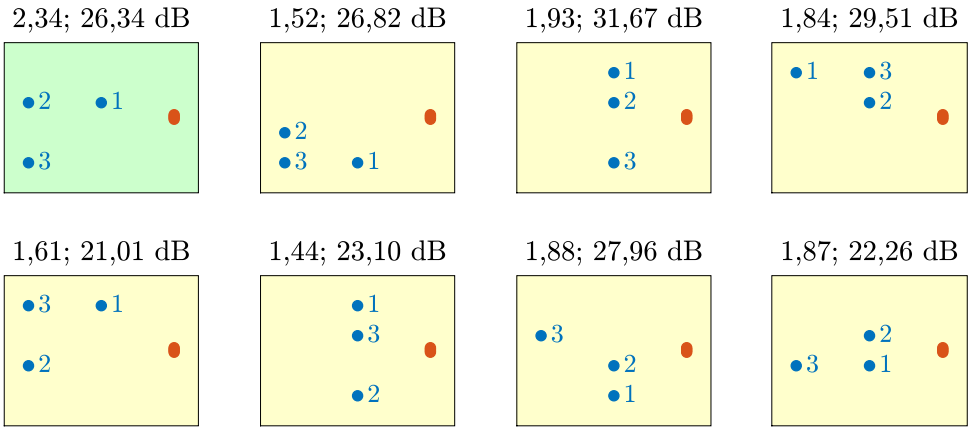
Obrázek 8.1: Výsledky v simulované malé místnosti s 2 řečníky separovanými metodou ICA



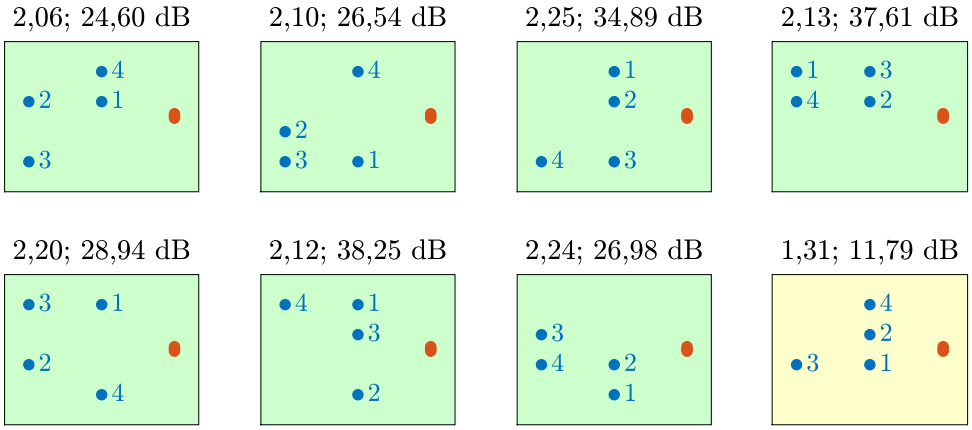
Obrázek 8.2: Výsledky v simulované malé místnosti s 2 řečníky separovanými metodou IVA



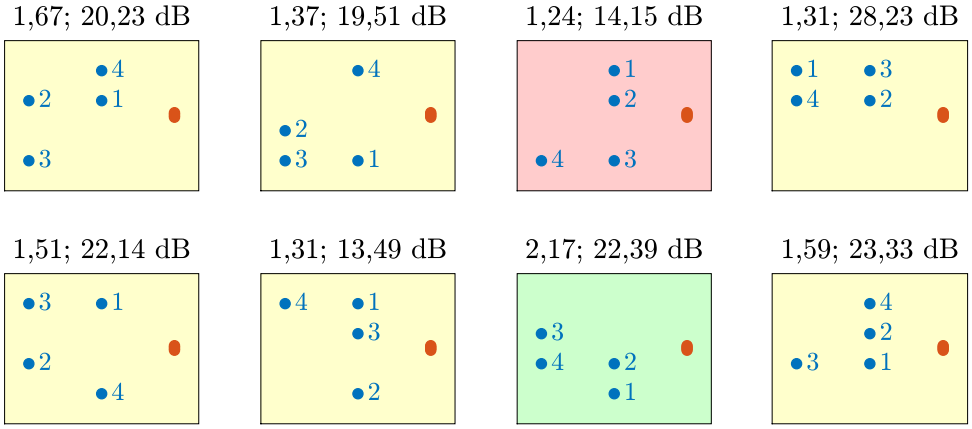
Obrázek 8.3: Výsledky v simulované malé místnosti s 3 řečníky separovanými metodou ICA



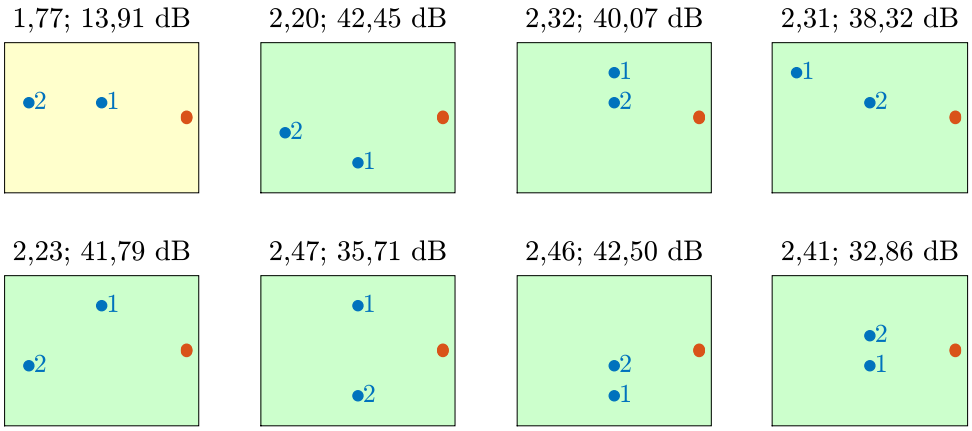
Obrázek 8.4: Výsledky v simulované malé místnosti s 3 řečníky separovanými metodou IVA



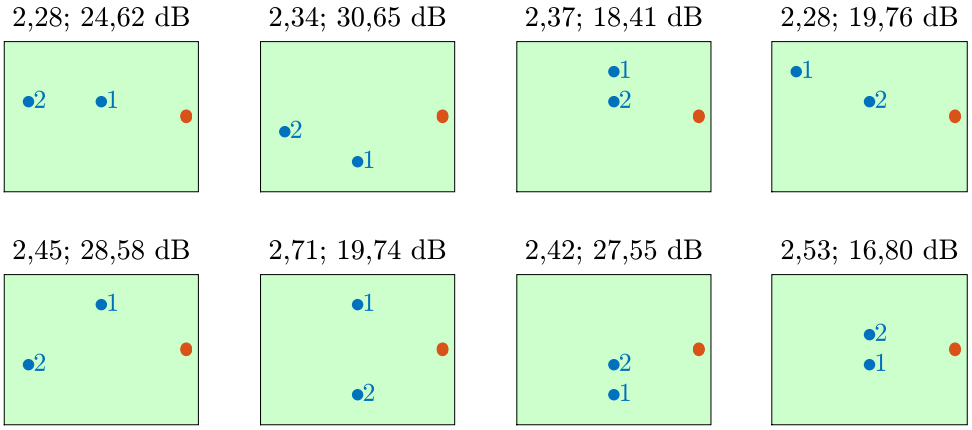
Obrázek 8.5: Výsledky v simulované malé místnosti s 4 řečníky separovanými metodou ICA



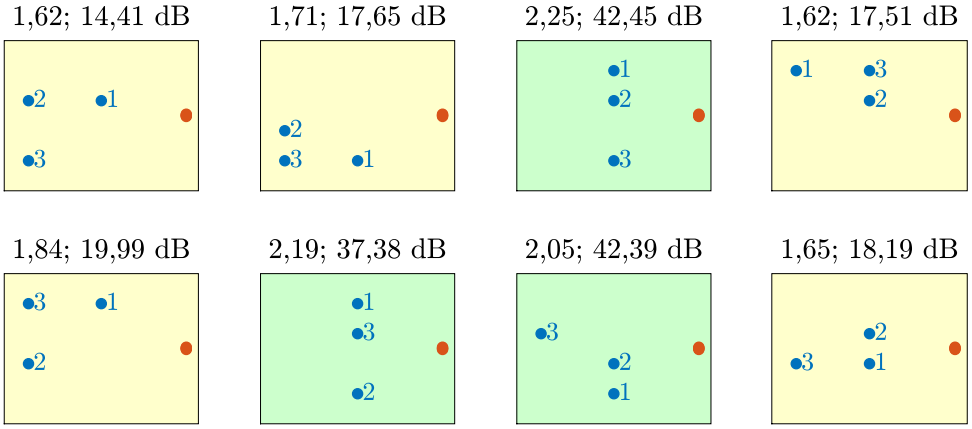
Obrázek 8.6: Výsledky v simulované malé místnosti s 4 řečníky separovanými metodou IVA



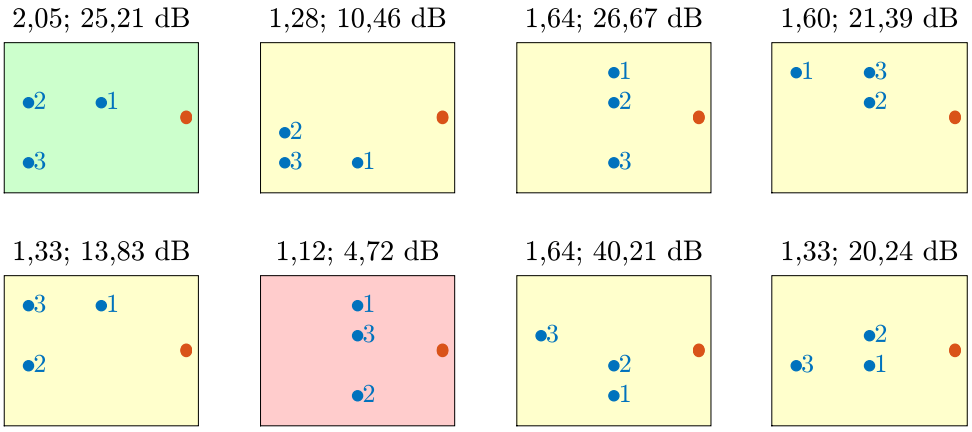
Obrázek 8.7: Výsledky v simulované velké místnosti s 2 řečníky separovanými metodou ICA



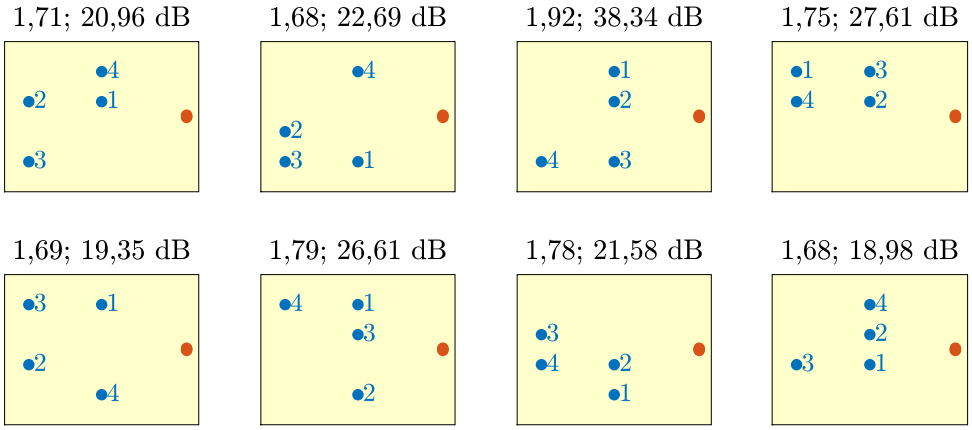
Obrázek 8.8: Výsledky v simulované velké místnosti s 2 řečníky separovanými metodou IVA



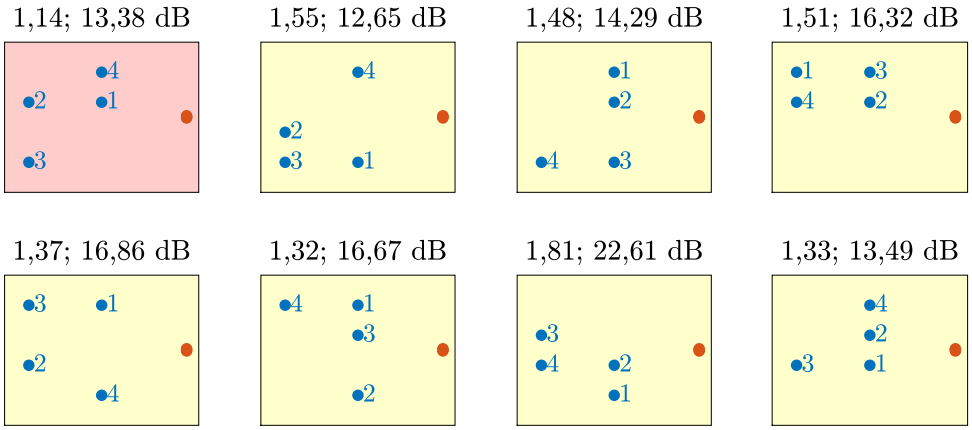
Obrázek 8.9: Výsledky v simulované velké místnosti s 3 řečníky separovanými metodou ICA



Obrázek 8.10: Výsledky v simulované velké místnosti s 3 řečníky separovanými metodou IVA



Obrázek 8.11: Výsledky v simulované velké místnosti s 4 řečníky separovanými metodou ICA



Obrázek 8.12: Výsledky v simulované velké místnosti s 4 řečníky separovanými metodou IVA

Finální zprůměrované hodnoty vyjadřující kvalitu separace v simulované místnosti jsou shrnuty v tabulkách 8.1 a 8.2. Celkově vzato, dosáhly obě metody v simulované místnosti velmi dobrých výsledků. Výsledné průměrné hodnoty SNR jsou až na jeden případ vyšší než 20 dB, což je solidní výsledek, který je o trochu lepší než výsledky jiné práce [2], kde ale byly hodnoty SNR odhadovány jiným způsobem. Hodnoty Keps jsou v průměru blízké 2, což odpovídá separaci takřka dokonalé.

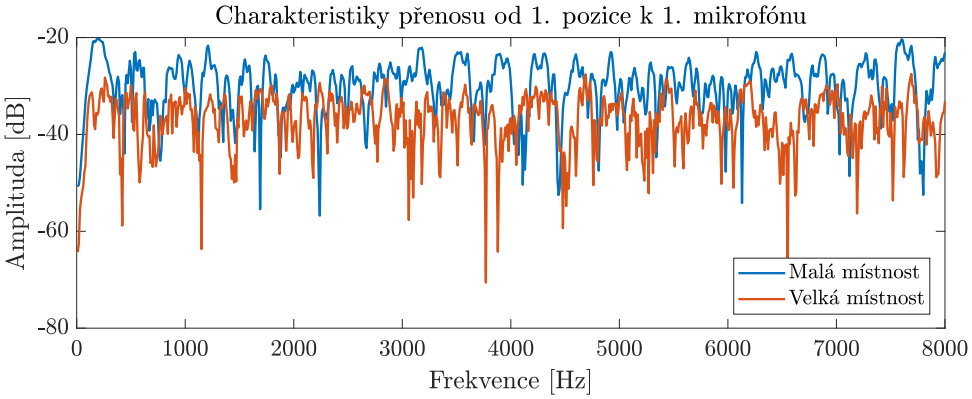
Z vypočítaných hodnot je jednoznačně patrné, že v menší místnosti byla separace o něco úspěšnější. Vzhledem k tomu, že v obou místnostech byly řečníci i mikrofony rozmístěny stejným způsobem, bude tento rozdíl kvality separace způsoben pravděpodobně akustickými specifiky obou místností. V tomto ohledu se ale větší místnost od menší liší pouze v delší době mezi odrazy signálů ode zdí (vlivem větších vzdáleností v místnosti). To pochopitelně výrazným způsobem změní i frekvenční charakteristiku přenosu, jak je ukázáno v grafu na obrázku 8.13.

počet řečníků		2	3	4	$\mu \pm \sigma$
malá místnost	ICA	33,63	31,43	28,70	**31,25 ± 1,19**
	IVA	26,32	26,08	20,43	**24,28 ± 0,72**
velká místnost	ICA	35,95	26,25	24,52	**28,90 ± 1,08**
	IVA	23,26	20,34	15,78	**19,80 ± 0,78**

Tabulka 8.1: Finální výsledky SNR [dB] pro simulovanou místnost (tvořené průměry pro jednotlivé počty řečníků a střední hodnotou μ se směrodatnou odchylkou σ určenými ze všech hodnot SNR)

počet řečníků		2	3	4	$\mu \pm \sigma$
malá místnost	ICA	2,80	2,24	2,05	**2,37 ± 0,04**
	IVA	2,82	1,80	1,52	**2,05 ± 0,07**
velká místnost	ICA	2,27	1,87	1,75	**1,96 ± 0,03**
	IVA	2,42	1,50	1,43	**1,79 ± 0,05**

Tabulka 8.2: Finální výsledky Keps pro simulovanou místnost (tvořené průměry pro jednotlivé počty řečníků a střední hodnotou μ se směrodatnou odchylkou σ určenými ze všech hodnot Keps)



Obrázek 8.13: Srovnání frekvenčních charakteristik přenosů mezi první pozicí a mikrofonem v malé i velké místnosti

Po prozkoumání tohoto grafu lze prohlásit (jedná se ale pouze o subjektivní hodnocení), že modul frekvenční charakteristiky daného přenosu je ve větší místnosti o něco proměnlivější, než v místnosti malé. A právě to by mělo na kvalitu separace negativní vliv. Použité metody totiž separované signály pomocí STFT rozkládají do úzkých frekvenčních pásem, ve kterých je s určitou nepřesností předpokládán lineární model směsí zdrojových signálů, jak je ukázáno v kapitole 3. Lineární model se zároveň vyznačuje konstantním průběhem modulu frekvenční charakteristiky v daném pásmu (mixážní matice $\mathbf{A}$ je v tomto pásmu frekvenčně nezávislá). Čím je ale modul frekvenční charakteristiky proměnlivější, tím méně se v daných pásmech blíží konstantnímu průběhu a tím horší jsou podmínky pro použité separační metody. Tuto vlastnost lze samozřejmě řešit volbou užšího pásma při aplikaci STFT, to ale nicméně v tomto případě vedlo na výsledky jednoznačně horší.

Z výsledků také vyplývá, že metoda ICA dosahovala vyšší úspěšnosti, než metoda IVA. To je sice v rozporu s dostupnou literaturou, kde lepších výsledků dosáhla metoda IVA [17], jednalo se ale o situaci, kdy byly separovány uměle vytvořené signály neakustického charakteru. I tak byl ale v rámci této práce rozdíl mezi výsledky obou metod signifikantní a bude snaha kvalitu separace metody IVA vylepšit. Na začátek byly použity jiné nelinearity g, větší počet iterací a jiné šířky pásma STFT, ale nic z toho k lepším výsledkům nevedlo.

Zásadní rozdíl mezi oběma metodami tkví v tom, že zatímco metoda ICA se během separování soustředí pouze na jedno konkrétní frekvenční pásmo, metoda IVA pracuje se všemi pásmy zároveň. Metoda IVA proto během separace pracuje i s pásmy, která sice nejsou dobře slyšitelná, mohou ale obsahovat velké množství výkonu, čímž by výsledky metody IVA mohly být značně ovlivněny. Jedná se zejména o nízkofrekvenční pásma (zhruba do 100 Hz), ve kterých je soustředěn takzvaný „1/f šum“ představující přirozený šum mikrofonů [18]. Byly tedy učiněny experimenty, během kterých se nahrané signály $\mathbf{x}_s$ ještě před samotnou separací podrobily filtraci horní propustí, která zmíněná nízkofrekvenční pásma dostatečně potlačila. Tento způsob ale vedl na takřka totožné výsledky. Jak totiž ukázala spektrální analýza směsí $\mathbf{x}_s$, tato pásma v nich již potlačena byla. Směsi totiž při simulaci vznikly smíšením promluv z databází Českého rozhlasu, které byly nahrány ve studiové kvalitě a evidentně i nějakým způsobem předzpracovány. Filtrování signálů $\mathbf{x}_s$ proto v tomto případě nakonec provedeno nebylo.

Ačkoliv kvalita separace metodou IVA byla v případě simulované místnosti o něco horší a tento fakt se nepodařilo vysvětlit, náhodné poslechové testy odhalily, že metoda IVA naopak způsobuje znatelně menší zkreslení separovaných promluv, než ICA. Pro finální srovnání obou metod proto budou důležité závěry poslechového testu finálního experimentu v reálné místnosti, který ono zkreslení promluv hodnotí.

Ve 3 případech (na obrázcích s výsledky zobrazeny červeně) dopadla separace nedostatečně. Poslech separovaných promluv tuto skutečnost potvrdil, takže se nejednalo o špatné vyhodnocení kvality separace například v důsledku chybného setřídění signálů. V těchto konkrétních případech bylo experimentováno s jiným umístěním mikrofonového pole, což kvalitu separace okamžitě zvedlo na vynikající hodnoty. Pokud ale bylo pozměněné umístění mikrofonů aplikováno i při ostatních modelových situacích, nastaly pro změnu jiné jednotky případů s nedostatečnou separací a výsledné průměrné hodnoty kvality separace zůstaly takřka beze změny. To ukázalo, že rozmístění zdrojů má na kvalitu separace v dané místnosti velmi výrazný vliv a pro každé uspořádání mluvčích v simulované místnosti je možné najít takové umístění mikrofonů, které směs separuje takřka dokonale.

Z hodnot Keps i SNR je ve finálním zprůměrování možné vysledovat efekt počtu lidí v místnosti. Totiž, že s rostoucím počtem lidí se kvalita separace snižovala, což je závěr, který byl kvůli povaze úlohy očekáván. Je zřejmé, že čím více mluvčích v místnosti bude, tím bude směs komplexnější a její separace bude náročnější.

Na výsledcích jednotlivých separovaných modelových situacích je občas patrné, že hodnoty Keps a SNR spolu příliš nekorespondují. Tato jejich souvislost byla vyšetřena pomocí korelace s výsledkem

$$\mathrm{r}(94) = 0{,}58; p < 0{,}001.$$

Ten ukazuje, že nějaká souvislost mezi oběma parametry spolehlivě existuje, je ale na průměrné úrovni a přesvědčivých hodnot nedosahuje. Tato skutečnost bude pravděpodobně důsledkem faktu, že oba parametry hodnotí kvalitu separace úplně jiným způsobem. Náhodné poslechové testy provedené autorem práce ukázaly, že parametr Keps odráží kvalitu separace o něco lépe. Děje se tak pravděpodobně z toho důvodu, že hodnoty SNR jsou zatíženy chybou v časových průbězích separovaných signálů, jak je popsáno v sekci 7.3. Tato chyba je ale průměrováním hodnot SNR do určité míry potlačena a na výsledné hodnoty se tedy spolehnout lze.

8.2 Výsledky experimentu v reálné místnosti

	Šířka pásma [Hz]	**4**	**2**	**1**	**0,5**
SNR [dB]	ICA	14,52	21,51	**29,49**	8,56
	IVA	14,89	17,76	**19,81**	12,60
Keps [-]	ICA	1,43	1,89	**2,03**	1,49
	IVA	1,67	1,96	**2,02**	1,47

Tabulka 8.3: Výsledky separace v reálné místnosti se známými zdrojovými promluvami (maxima vyznačena tučně)

Výsledky objektivních metrik při experimentu v reálné místnosti jsou shrnuty v tabulce 8.3. V tomto případě dosahují obě metody poměrně podobných výsledků, které jsou ale v porovnání se simulovanou místností o něco horší. Tento efekt byl očekáván, neboť při měření v reálné místnosti byl přítomen skutečně velký šum a hluk okolí. I tak se ale jedná o solidní výsledky a hodnoty Keps ≈ 2 by měly odpovídat hraničním hodnotám pro dokonalou separaci. Zběžně byl proveden i poslech nahrávek, který subjektivně vyhodnotil separaci jako velmi kvalitní, ale ne úplně dokonalou, protože v odseparovaných nahrávkách byl druhý řečník stále, i když opravdu slabě, slyšitelný.

Z výsledků je především patrný vliv šířky pásma STFT na kvalitu separace. Skutečně je možné nalézt určitou šířku pásma, při které dopadla separace nejlépe. V tomto případě šlo zcela jednoznačně o hodnotu 1 Hz. Může se zdát trochu překvapivé, že v simulované místnosti byla tato hodnota 4 Hz. Pravděpodobně se jedná o důsledek faktu, že v simulované místnosti nebyly simulovány všechny aspekty šíření zvuku místností a frekvenční charakteristiky přenosů tak nejspíš byly vyrovnanější, což umožnilo zvolit větší šířku pásma, jak již bylo diskutováno v předchozích částech práce.

Vliv šířky pásma byl patrný i v rámci porovnání obou použitých metod. Zejména na parametru SNR (ale z části i na parametru Keps) je možné sledovat, že výsledky metody ICA se s použitu šířkou pásma mění více, než je tomu v případě metody IVA. Tento jev by mohl být způsobený tím, že metoda IVA při separaci pracuje vždy se všemi pásmy zároveň, díky čemuž by mohla být konkrétní hodnotou šířky pásma méně ovlivněná.

Šířka pásma ale nebude mít vliv pouze na kvalitu separace, ale i na zkreslení odseparovaných promluv. Tento vliv byl sledován poslechovým testem, jehož výsledky se nacházejí v následující sekci.

8.3 Výsledky poslechového testu

Poslechového testu se účastnilo 6 lidí. K dispozici měli jednu referenční nahrávku a 8 odseparovaných, které měli za úkol ohodnotit z hlediska kvality separace a míry jejich zkreslení. Celý poslechový test je popsán v sekci 7.4.

Před shrnutím a vyhodnocením výsledných hodnot poslechového testu bylo vyšetřeno, zda jsou hodnoty uvedené účastníky testu v rámci jednotlivých promluv konzistentní. Tedy, zda jsou odpovědi účastníků podobné a nejsou pouze náhodné (například v důsledku nízké rozlišitelnosti hodnocených promluv). Tato podoba byla určena vzájemnou korelací uvedených odpovědí ve všech možných dvojicích řečníků. Hodnoty korelací dosahovaly vysokých hodnot a vždy platilo

$$\mathrm{r}(14) \geq 0{,}8; p < 0{,}01.$$

Na základě toho lze říci, že výpovědi nejsou náhodné, ale naopak silně konzistentní a díky tomu je z nich možné vyvozovat nějaké úsudky.

Při návrhu poslechového testu také byla obava, že jeho účastníkům bude při hodnocení chybět nějaké měřítko, aby věděli, jaké konkrétní hodnoty mají pro ohodnocení jednotlivých nahrávek zvolit. To sice měla do určité míry řešit přítomnost referenční nahrávky, ale například měřítko maximálního možného zkreslení i tak nebylo k dispozici. Účastníci proto měli za úkol jednotlivé hodnoty vyplnit spíše tak, aby dobře vystihli vzájemnou podobnost nebo naopak rozdílnost jednotlivých promluv v obou testovaných parametrech. Vyplněné hodnoty poté bylo v plánu škálovat do podobných rozsahů. Účastníci testu ale nakonec nezávisle na sobě provedli hodnocení v podstatě stejných rozmezích (konkrétně 45 % až 95 %) a toto škálování nebylo nutné provádět.

Pro každou promluvu byly hodnoty uvedené jednotlivými účastníky zprůměrovány a uvedeny v tabulkách 8.4 a 8.5. Vzhledem ke zmíněnému způsobu ohodnocení těchto nahrávek jsou pro vyhodnocení poslechového testu spíše než konkrétní hodnoty zajímavé jejich vzájemné rozdíly.

Na základě výsledků je možné prohlásit, že promluva prvního řečníka byla separována o něco lépe. Důvod tkví pravděpodobně v tom, že se první řečník během nahrávání nacházel o trochu blíže mikrofonovému poli a mluvil o něco hlasitěji. V nahrané směsi je proto jeho promluva dominantnější, což je ostatně

	Šířka pásma [Hz]	1	2	$\mu \pm \sigma$
ICA	promluva 1	61 %	68 %	**63,3 ± 2,3 %**
	promluva 2	59 %	65 %	
IVA	promluva 1	83 %	84 %	**82,1 ± 1,7 %**
	promluva 2	81 %	80 %	

Tabulka 8.4: Zprůměrované výsledky poslechového testu hodnotícího kvalitu separace včetně celkových průměrů μ a směrodatných odchylek σ (100% odpovídá dokonale separované promluvě, kdy druhý řečník není vůbec slyšet)

	Šířka pásma [Hz]	1	2	$\mu \pm \sigma$
ICA	promluva 1	66 %	75 %	**68,6 ± 1,5 %**
	promluva 2	63 %	71 %	
IVA	promluva 1	86 %	91 %	**86,7 ± 0,9 %**
	promluva 2	85 %	86 %	

Tabulka 8.5: Zprůměrované výsledky poslechového testu hodnotícího zkreslení promluv včetně celkových průměrů μ a směrodatných odchylek σ (100% odpovídá nezkreslené promluvě)

na referenční nahrávce zřetelně slyšet. Ačkoliv je rozdíl v separaci mezi řečníky ve výsledcích poslechového testu systematický, není příliš markantní.

Na výsledné hodnoty poslechového testu měla zajímavý efekt použitá šířka pásma STFT při separaci. Prakticky vždy byla separace úspěšnější při šířce pásma 2 Hz a to jak z hlediska kvality separace, tak míry zkreslení promluv. V případě zkreslení promluv je tento efekt poměrně velký a jen potvrzuje domněnku, že s rostoucí šířkou pásma se snižuje zkreslení signálu použitými separačními metodami. V případě kvality separace se ale jedná o výsledek trochu překvapivý, neboť při předchozím experimentu v reálné místnosti (který byl hodnocen objektivními metrikami) se ukázala být lepší separace při šířce pásma 1 Hz, jak je patrné z tabulky 8.3. Oba experimenty ale byly uskutečněny ve stejné místnosti, se stejnými a stejně rozestavěnými řečníky, kteří pronášeli stejné promluvy. Nicméně trochu rozdílné výsledky v hodnocení subjektivními a objektivními metrikami jsou očekávatelné, navíc v případě poslechového testu není efekt šířky pásma na kvalitu separace tak velký.

Velkého efektu si naopak lze všimnout při porovnání úspěšnosti metod ICA a IVA, kdy metoda IVA dosáhla jednoznačně lepších výsledků při hodnocení kvality separace i míry zkreslení. Mimo to je patrné, že na její výsledky měla použitá šířka pásma STFT znatelně menší vliv, než tomu bylo v případě metody ICA, což bylo sledováno i v případě vyhodnocení experimentu v reálné místnosti objektivními metrikami. Tato vlastnost metody IVA je spíše pozitivní a na základě dostupných výsledků z reálné místnosti lze prohlásit, že se v ní metoda IVA chovala lépe, než ICA. Toto ale nelze prohlásit obecně, protože nebylo uskutečněno větší množství pokusů, které by obě metody prověřily lépe, jako tomu bylo v případě místnosti simulované, kde o něco lepších výsledků dosáhla naopak metoda ICA.

Kapitola 9

Závěr

Cílem této práce bylo popsat, implementovat a následně otestovat vybrané metody slepé separace řečových signálů pomocí analýzy nezávislých komponent (ICA).

Na začátku práce byla nejprve definována úloha slepé separace spolu s jednoduchým, takzvaně lineárním, modelem směsí signálů. Představeny byly základní metody ICA schopné tyto lineární směsi za určitých předpokladů separovat. Kromě těchto metod byly uvedeny i jejich klíčové vlastnosti, zejména nejistota v pořadí a amplitudě, se kterými se v případě separace řečových signálů bylo nutné vypořádat.

Cílem práce je slepě separovat směsi řečových signálů. Tyto směsi ale jednoduchým lineárním modelem s dostatečnou přesností popsat nelze, a proto by ani základní představené metody ICA při separaci těchto směsí nebyly účinné. V práci je proto zaveden komplexnější konvoluční model směsi signálů, který již způsob mísení řeči popisuje s dostatečnou přesností. Při separaci konvolučních směsí řečových signálů se postupuje takovým způsobem, že jsou zaznamenané směsi nejprve pomocí krátkodobé Fourierovy transformace (STFT) rozloženy do úzkých frekvenčních pásmech, ve kterých je již uvažován lineární model směsi signálů. V těchto pásmech jsou po drobných změnách aplikovány základní metody ICA. Odseparovaná pásma jsou pak pomocí inverzní krátkodobé Fourierovy transformace převedeny zpět do časové oblasti, čímž vzniknou výsledné odseparované promluvy.

Tento způsob byl v rámci práce detailně popsán a to včetně změn metod ICA, které bylo nutné provést. Jednalo se o uzpůsobení metod k separaci komplexních signálů (kvůli aplikaci STFT) a vyřešení nejistot v pořadí a amplitudě separovaných signálů metodami ICA. Tyto nejistoty se totiž v každém separovaném frekvenčním pásmu projeví jiným způsobem, a pokud by takto odseparovaná pásma byla naivně převedena zpět do časové oblasti, výsledné signály by byly nesrozumitelné. Metody ICA také bylo žádoucí uzpůsobit separaci několika promluv na základě dat z mikrofonových polí o desítkách mikrofonů, což je postup, který značně pomůže při separování promluv smíšených v reálné místnosti, kde jsou kromě zdrojových signálů přítomny i šumy a hluky okolí.

V rámci této práce byly vybrány dvě konkrétní metody, které všechny zmíněné problémy řeší, a které jsou tedy pro separaci řečových signálů použitelné. Jedná se o metody Fast fixed-point ICA a Fast fixed-point IVA. Ty byly implementovány v prostředí Matlab a ve stejném prostředí byly následně připraveny experimenty se simulovanou místností za účelem důkladného prověření účinnosti těchto metod.

Simulovány byly místnosti dvou velikostí s různě rozmístěnými řečníky o počtu 2 až 4 osob. Takto bylo simulováno celkem 48 směsí k odseparování. Separace proběhla vždy pomocí obou metod a následně byla objektivně vyhodnocena její kvalita (úroveň potlačení promluv ostatních řečníků) pomocí parametrů založených na metrikách SNR a kepstrální vzdálenost.

Dva experimenty také byly provedeny v místnosti reálné. V ní se fyzicky nacházeli dva řečníci, kteří současně pronášeli souvislé promluvy. Protože bylo při měření v místnosti přítomno i velké množství šumu a hluku, byly promluvy zaznamenány pomocí 40 mikrofonů sdružených v poli. Vzniklé směsi byly odseparovány opět pomocí obou metod, ale s různými šířkami pásma použitými při rozkladu signálu pomocí STFT. Při prvním z těchto dvou experimentů byly k dispozici i signály zdrojové a kvalita separace díky tomu byla opět vyhodnocena objektivními metrikami. Finální experiment byl podroben jednoduchému poslechovému testu sledujícímu jak kvalitu separace, tak úroveň zkreslení promluv použitými metodami.

Výsledky z experimentů v simulované místnosti ukázaly, že obě metody fungují výtečně. Zpozorován byl pokles kvality separace s rostoucím počtem řečníků v místnosti. Jejich konkrétní rozmístění pak kvalitu separace taktéž velmi ovlivňovalo. Také se ukázalo, že metoda Fast fixed-point ICA poskytovala o trochu lepší výsledky.

V případě reálné místnosti sice došlo k určitému poklesu kvality separace v porovnání s místností simulovanou, i tak se ale jednalo o velmi dobré hodnoty. Tyto experimenty ale především potvrdily, že implementované metody jsou schopné promluvy separovat i v reálném prostředí plném šumů a hluků. Poslechové testy také odhalily zajímavý efekt šířky pásma STFT na výsledné odseparované promluvy. S klesající šířkou pásma rostla nejen kvalita separace, ale rostlo i zkreslení výsledných signálů. Bylo tedy možné najít určitou kompromisní šířku pásma, pro kterou byly na rozumných hodnotách oba sledované parametry. Poslechové testy také ukázaly, že v prostředí reálné místnosti dosahuje o něco lepších výsledků naopak metoda Fast fixed-point IVA.

Výsledkem práce jsou metody Fast fixed-point ICA a Fast fixed-point IVA pro slepou separaci řečových signálů implementované v prostředí Matlab. Tyto metody byly důkladně prověřeny a dosahují solidních výsledků.

Literatura

[1] HYVÄRINEN, Aapo a Erkki OJA. Independent component analysis: algorithms and applications. *Neural Networks* [online]. 2000, **13**(4-5), 411-430 [cit. 2022-02-10]. ISSN 08936080. Dostupné z: doi:10.1016/S0893-6080(00)00026-5

[2] MURATA, No, Shiro IKEDA a Andreas ZIEHE. An approach to blind source separation based on temporal structure ofspeech signals. *Neurocomputing* [online]. 2001, **41**(1-4), 1-24 [cit. 2022-03-05]. Dostupné z: doi:10.1016/S0925-2312(00)00345-3

[3] REDDY, Chandan K A, Anshuman GANGULY a Issa PANAHI. ICA based single microphone Blind Speech Separation technique using nonlinear estimation of speech. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* [online]. IEEE, 2017, 2017, 5570-5574 [cit. 2022-04-02]. ISBN 978-1-5090-4117-6. Dostupné z: doi:10.1109/ICASSP.2017.7953222

[4] HIMAWAN, Ivan, Iain MCCOWAN a Mike LINCOLN. Microphone Array Beamforming Approach to Blind Speech Separation. *Machine Learning for Multimodal Interaction* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, 295-305 [cit. 2022-04-02]. Lecture Notes in Computer Science. ISBN 978-3-540-78154-7. Dostupné z: doi:10.1007/978-3-540-78155-4_26

[5] Principal component analysis. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2022-05-05]. Dostupné z: https://en.wikipedia.org/wiki/Principal_component_analysis

[6] HYVARINEN, Aapo. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* [online]. 1999, **10**(3), 626-634 [cit. 2022-03-05]. ISSN 1045-9227. Dostupné z: doi:10.1109/72.761722

[7] TRUAX, Barry. Sound - Environment Interaction: Sound Propagation in Open and Enclosed Spaces. *Simon Fraser University* [on-

line]. [cit. 2022-03-07]. Dostupné z: https://www.sfu.ca/sonic-studio-webdav/cmns/Handbook%20Tutorial/Sound-Environment.html

[8] LEE, Intae, Taesu KIM a Te-Won LEE. Fast fixed-point independent vector analysis algorithms for convolutive blind source separation. *Signal Processing* [online]. 2007, **87**(8), 1859-1871 [cit. 2022-03-17]. ISSN 01651684. Dostupné z: doi:10.1016/j.sigpro.2007.01.010

[9] KIM, Taesu, Intae LEE a Te-Won LEE. Independent Vector Analysis: Definition and Algorithms. *2006 Fortieth Asilomar Conference on Signals, Systems and Computers* [online]. IEEE, 2006, 2006, 1393-1396 [cit. 2022-03-20]. ISBN 1-4244-0784-2. Dostupné z: doi:10.1109/ACSSC.2006.354986

[10] BINGHAM, ELLA a AAPO HYVÄRINEN. A FAST FIXED-POINT ALGORITHM FOR INDEPENDENT COMPONENT ANALYSIS OF COMPLEX VALUED SIGNALS. *International Journal of Neural Systems* [online]. 2012, **10**(01), 1-8 [cit. 2022-03-21]. ISSN 0129-0657. Dostupné z: doi:10.1142/S0129065700000028

[11] KIM, Taesu, Hagai ATTIAS, Soo-Young LEE a Tee-Won LEE. Blind Source Separation Exploiting Higher-Order Frequency Dependencies. *IEEE Transactions on Audio, Speech, and Language Processing* [online]. 2007, **15**(1), 70-79 [cit. 2022-03-24]. Dostupné z: doi:10.1109/TASL.2006.872618

[12] CIARAMELLA, Angelo a Roberto TAGLIAFERRI. Amplitude and permutation indeterminacies in frequency domain convolved ICA. *Proceedings of the International Joint Conference on Neural Networks* [online]. 2003, **1**(1), 708 - 713 [cit. 2022-03-24]. Dostupné z: doi:10.1109/IJCNN.2003.1223454

[13] ČMEJLA, Roman. Syntéza audio signálů, spektrální manipulace [přednáška]. Praha: ČVUT FEL, 14. 10. 2020. Dostupné z: https://moodle.fel.cvut.cz/course/view.php?id=5602

[14] LEHMANN, Eric A. a Anders M. JOHANSSON. Diffuse Reverberation Model for Efficient Image-Source Simulation of Room Impulse Responses. *IEEE Transactions on Audio, Speech, and Language Processing* [online]. 2010, **18**(6), 1429-1439 [cit. 2022-04-11]. ISSN 1558-7916. Dostupné z: doi:10.1109/TASL.2009.2035038

[15] LEHMANN, Eric A. Fast simulation of acoustic room impulse responses (image-source method). In: *MathWorks* [online]. 2022 [cit. 2022-04-11]. Dostupné z: https://www.mathworks.com/matlabcentral/fileexchange/25965-fast-simulation-of-acoustic-room-impulse-responses-image-source-method

[16] POLLÁK, Petr. Zpracování řeči, kepstrální analýza řečového signálu [přednáška]. Praha: ČVUT FEL, 1. 3. 2021. Dostupné z: https://moodle.fel.cvut.cz/course/view.php?id=5736

[17] KIM, Taesu, Torbjørn ELTOFT a Te-Won LEE. Independent Vector Analysis: An Extension of ICA to Multivariate Components. *Independent Component Analysis and Blind Signal Separation* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, 2006, 165-172 [cit. 2022-05-14]. Lecture Notes in Computer Science. ISBN 978-3-540-32630-4. Dostupné z: doi:10.1007/11679363_21

[18] Pink noise. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001- [cit. 2022-04-21]. Dostupné z: https://en.wikipedia.org/wiki/Pink_noise