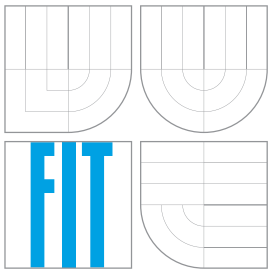
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

ODHAD EMOCÍ ŘEČNÍKA Z MLUVENÉ ŘEČI

EMOTION DETECTION FROM SPEECH

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

ANNA POPKOVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. PAVEL MATĚJKA, Ph.D.

BRNO 2016

Zadání bakalářské práce/18675/2015/xpopko00

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2015/2016

Zadání bakalářské práce

Řešitel: **Popková Anna**
Obor: Informační technologie
Téma: **Odhad emocí řečníka z mluvené řeči**
Emotion Detection from Speech
Kategorie: Zpracování řeči a přirozeného jazyka

Pokyny:

1. Prostudujte statistické techniky pro modelování řeči - zejména neuronové sítě.
2. Seznamte se s daty ze soutěže AVEC 2015 - soutěž na rozpoznávání emocí z mluvené řeči.
3. Navrhněte topologii systému.
4. Vyhodnoťte systém na "baseline" příznacích dodaných v rámci AVEC 2015.
5. Navrhněte vlastní zpracování těchto příznaků pro zlepšení úspěšnosti - transformace, časový kontext, atd.
6. Porovnejte úspěšnost s jinými příznaky.
7. Porovnejte úspěšnost s jiným teamem v této soutěži.

Literatura:

- http://sspnet.eu/avec2015/
- Fabien Ringeval et al. "*The AV+EC 2015 Multimodal Affect Recognition Challenge:Bridging Across Audio, Video, and Physiological Data*", http://sspnet.eu/wp-content/uploads/2015/03/avec2015.pdf

Pro udělení zápočtu za první semestr je požadováno:

- Body 1 až 4

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Matějka Pavel, Ing., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2015
Datum odevzdání: 18. května 2016

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

L.S.

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Tato bakalářská práce se zabývá výzkumem v oblasti rozpoznávání emocí z řeči a okrajově i z dalších modalit (video a fyziologické záznamy). Popisuje topologii systémů, které byly pro tento výzkum postaveny. Dále popisuje experimenty s těmito systémy vedoucí k optimálnímu předzpracování, trénování a po–zpracování dat. K výzkumu jsou použita data z evaluace AV+EC 2015, do níž byly zaslány výsledky fúzních systému produkujících nejpřesnější predikci. Nově jsou v oblasti rozpoznávání emocí z řeči vyzkoušeny Bottle–Neck příznaky. Jsou použity spolu s běžně používanými eGeMAPS příznaky ve fúzním systému rozpoznávající emoční dimenzi arousal. Emoční dimenze valence je pak rozpoznávána dvojicí video příznaků. Multi–task systém (rozpoznávající valenci i arousal) používající Bottle–Neck příznaky produkuje výslekdy pouze o 13 % relativně horší, než zmíněný fúzní systém, což apeluje hlavně na situace, kde jsou dostupná pouze audio data.

Abstract

This Bachelor Thesis deals with research in the field of emotion recognition mainly from speech and marginally from other modalities (video and physiological data). It closely describes the topology of the systems built specifically for the subject of this work. Moreover, it describes experiments leading to optimized pre–processing, regressor training and post–processing. Data used for these research origins from evaluation AV+EC 2015. Results of fusion systems producing the most precise prediction were sent to this evaluation. The Bottle–Neck features are newly tested and combined favorably with commonly used eGeMAPS features for the recognition of arousal. For valence, two kinds of video features are used. Muli–task system (recognizing both valence and arousal) using Bottle–Neck features produces competitive results and is only 13 % relatively behind the mentioned fusion system. This is especially appealing for applications where only audio is available.

Klíčová slova

Detekce emocí, audio, fúze, kontext, Bottle-Neck příznaky.

Keywords

Emotion recognition, speech, fusion, context, Bottle-Neck features.

Citace

POPKOVÁ, Anna. *Odhad emocí řečníka z mluvené řeči.* Brno, 2016. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Matějka Pavel.

Odhad emocí řečníka z mluvené řeči

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením pana Ing. Pavla Matějky, Ph.D. Další informace mi poskytli Ing. Ondřej Glembek, Ph.D. a Ing. František Grézl, Ph.D. Uvedla jsem všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Anna Popková
16. května 2016

Poděkování

Děkuji rodičům za podporu během všech třech let mého studia a děkuji celé Speech@FIT skupině za poskytnutí výborného prostředí pro tvorbu mé bakalářské práce.

Obsah

[1]multi–task je v tomto smyslu myšleno jako rozpoznávající obě emoční dimenze najednou, v jednom systému. Single–task je potom chápán jako systém rozpoznávající pouze jednu emoční dimenzi.

Kapitola 1

Úvod

Pro dnešní dobu je typické neustálé zjednodušování komunikace. Tento trend se dotýká i služeb, na které jsou kladeny čím dál vyšší nároky a komunikace je toho součástí. Má–li dnes zákazník problém, nebo potřebuje radu, nestartuje auto ani nepíše dopis. Pouze zvedne telefon, vytočí číslo a předpokládá rychlé a plynulé vyřízení jeho problému. Souvisle s tím očekává, že se v telefonu ozve milý a přívětivý hlas, otevřený naslouchat. Realita však vždy nemusí být taková — to je jedním z důvodů, proč začalo být téma rozpoznávání emocí z hlasu atraktivní. Nejen zaměstnavatelé call center, ale i firem, součástí jejichž služeb je komunikace se zákazníkem, chtějí mít správně reprezentující mluvčí. Právě k dosažení této skutečnosti by jim mohl dopomáhat automatizovaný přehled o tom, jak si v komunikaci vedou jejich zaměstnanci, a také jak na tyto zaměstnance reagují zákazníci.

Bakalářská práce, kterou právě držíte v ruce, se zabývá výzkumem v oblasti rozpoznávání emocí z řeči a okrajově i z dalších měřitelných modalit jako je video a záznamy elektrokardiografu a elektrodermální aktivity.

V práci se nejdříve zaměřím na možné způsoby jak emoce dělit a vnímat a blíže přiblížím dvoudimenzionální náhled na emoce, který bude použit pro tento výzkum. Dále čtenáře seznámím s daty, které byly pro tento výzkum použity. Poté navrhnu možnosti předzpracování, trénování a po–zpracování těchto dat. Na nich budu stavět experimenty, které následně také popíšu. Nakonec přednesu několik závěrů ohledně vykonaného výzkumu a navrhnu možné budoucí směry další práce.

1.1 Dosavadní výzkum

Myšlenka na vytvoření nástroje pro automatické rozpoznávání emocí rozhodně není čerstvá. První výzkumy v této oblasti se datují až k roku 1985. Jedny z prvotních systémů pro rozpoznávání emocí byly trénovány v kabinách letadel na pilotech, u kterých se rozpoznávalo, zda mluví pod stresem nebo mluví klidně. Akustické příznaky pro klasifikaci byly v té době extrahovány precizně pomocí iterativních algoritmů [18].

Pro další výzkum bylo obecně zapotřebí množství databází, na kterých by se systémy trénovaly. Proto jich během dalších let bylo několik vytvořeno. Obsahovaly jak spontánní, tak hrané emoce. Tyto databáze klasifikovaly různé druhy a počty emocí, kde každá z nich patřila k jedné celé nahrávce, nebo k její velké části. Rozpoznávání zde bylo diskrétního charakteru — klasifikátor určoval právě jednu predikci pro daný projev, tudíž k jednomu projevu náležela pouze jedna emoce. Souhrn mnoha takových databází je možno nalézt v [18]. Jejich hlavním nedostatkem byl nízký počet promluv, a tedy malý počet příkladu

konkrétní emoce.

V nedávné době se objevily databáze s časově kontinuálními průběhy emocí — například databáze HUMAINE [2], spojující několik databází dohromady za účelem sjednocení jejich forem a také jejich referenčních hodnot. Referenční hodnoty se zde skládají z tří–dimenzionálního náhledu na emoce. Většina databází, patřící pod HUMAINE, je spontánního charakteru, je zde ale i řada databází se snahou emoci navodit. Sadu doplňuje databáze, která obsahuje pouze hrané emoce. Novější databází je databáze SEMAINE [11], založená na SAL (Sensitive Artificial Listeneru), což je stroj, nebo stroj ovládaný člověkem. Databáze obsahuje interakce těchto strojů a lidí. SALs vyskytující se v této databázi mají 4 různé charaktery: Spike, který se snaží člověka naštvat, Poppy, snažící se jej rozveselit, Obadiah chce člověka rozesmutnit a Prudence v něm vyvolat rozumnost. Databáze SEMAINE rozlišuje 5 emočních dimenzí a je multimodální[1], stejně jako výše zmíněná HUMAINE. Další multimodální databáze jsou probírány v článku [3], kde je i popsáno, jakým způsobem takové databáze vznikají.

Právě databáze s kontinuálními průběhy emocí způsobily značný posun v rozpoznávání emocí. Zaprvé přechod od klasifikace k regresi a zadruhé přechod z rozpoznávání na úrovni promluvy k rozpoznávání kontinuálnímu. Právě automatické kontinuální rozpoznávání emocí apeluje na několik diskutabilních skutečností — inspirace v článku [13]:

1. Stanovení přiměřené délky časového kontextu, která je závislá jak na modalitě, tak na druhu rozpoznávané emoce. V literatuře není uvedena jednoznačná shoda co se týče nejlepší délky časového kontextu pro určitou modalitu a emoci. Už totiž samotné trvání emoce kolísá mezi 0.5 až 4 sekundami. Rychlost změny emoce je velmi rozdílná u různých modalit, což znamená, že i velikost rozpoznávaného úseku by měla být pro různé modality různá. U audia se emoce mění rychleji, u videa o něco pomaleji a u fyziologických projevů —jako je bušení srdce či pocení — úplně nejpomaleji.

2. Přístup k multimodální fúzi — spojení většího množství informací o projevu pro dosažení přesnější predikce emocí. V literatuře se objevují dva základní přístupy k tvorbě fúze — na úrovni příznaků a na úrovni rozhodování. Na úrovní příznaků se postupuje tak, že se příznaky z různých modalit spojí do jednoho velkého příznakového vektoru a nad ním je systém trénován. Naopak na úrovni rozhodování jsou trénovány systémy pro jednotlivé modality odděleně a tyto výsledky potom tvoří základ pro systém druhý.

3. Pro správnou definici referenčních výstupů je zapotřebí více anotátorů. Jednotliví anotátoři totiž mohou dělat chyby, být nekonzistentní ve svém úsudku a mít reakční prodlevu. Pro eliminaci inkonzistence jsou referenční výstupy vyvozeny ze znormalizovaných a následně zprůměrovaných anotací. Právě tyto referenční výstupy jsou v této práci brány jako vzor a označovány jako Gold standard. Přístup k odhadu reakční prodlevy bývá různý a jedna z metod bude probírána v kapitole 5.2.1.

Na všechna zmíněná témata se ve své práci zaměřuji, experimentuji s nimi a následně o nich podávám informace a závěry. Otázku více anotátorů dále řeším nalezením nejvhodnějších jedinců pro trénování, abych se co nejvíce přiblížila Gold standardu.

[1] Obsahující více modalit, kde jako modalita je myšlena například řeč, video atd.

Kapitola 2

Klasifikace emocí

Emoce jsou velmi komplexní jevy, jejichž charakteristickým rysem je velká citlivost a proměnlivost. Tato citlivost a proměnlivost se projevuje tím, že za určité situace může být emoce pro člověka vyvolána, ale jindy, u objektivně stejné situace, být vyvolána nemusí. To, zda–li je nebo není vyvolána se řídí subjektivním vnímáním situace jedincem, o jehož emoci se jedná. I když i pro jiné duševní procesy, jako je například paměť či myšlení, platí, že je jejich průběh a projev závislý na situačních předpokladech, emoce jsou mnohem citlivější [7].

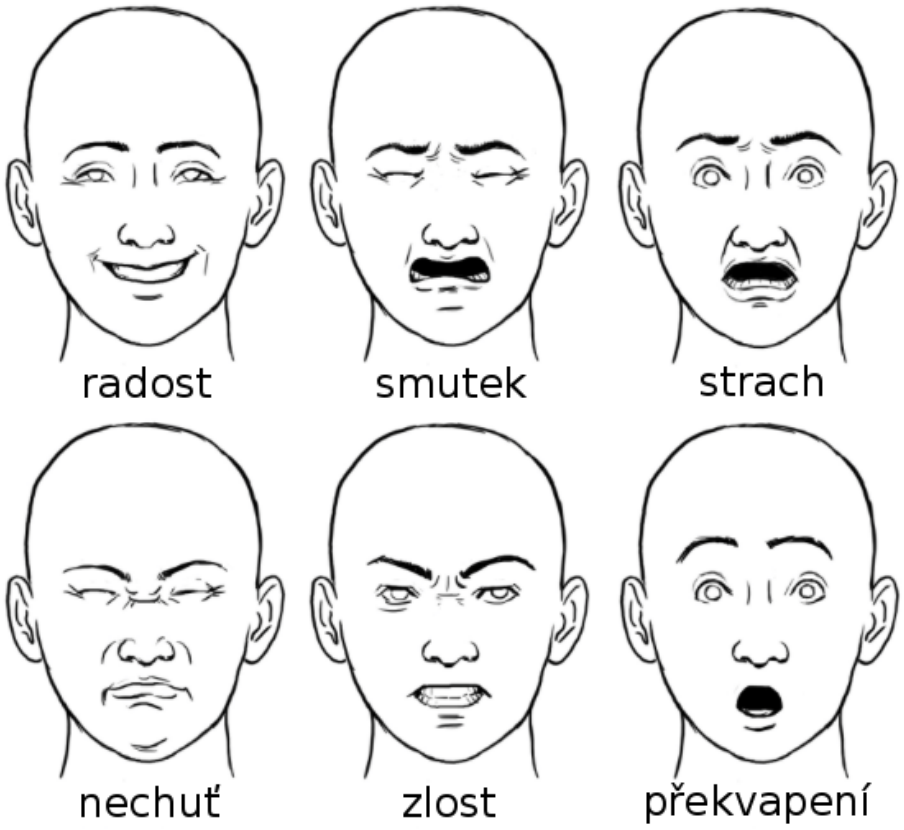
Mnoho výzkumníků se shoduje na tom, že emoce je složena z několika komponent, jmenovitě: Kognitivní odhad, subjektivní pocit, fyziologické vzrušení, výraz, tendence k činu a regulace. Na klasifikaci emocí neboli na způsobu, jak se liší jedna emoce od druhé však už existují pohledy různé.

V následujících dvou podkapitolách blíže popíšu dva v poslední době výrazně se profilující náhledy na emoce. První je teorie diskrétních emocí, vycházející z popisu tzv. primárních — základních, vrozených — emocí, které tvoří základ pro emoce sekundární — odvozené. Druhá je teorie dimenzionální, kde každá dimenze tvoří jistou složku emoce a dohromady tvoří emoční směs, kterou lze zařadit do konkrétního místa ve spojitém prostoru o určité dimensionalitě.

Popis obou zmíněných přístupů má v práci své opodstatnění, protože silně ovlivnily výzkum v oblasti řečového projevu. Zvláštní důraz je u nich kladen na fyziologickou složku emoce, neboť se všeobecně předpokládá, že fyziologické proměnné mají významný vliv na akustické charakteristiky hlasových projevů [9].

2.1 Teorie diskrétních emocí

Každá diskrétní emoce je chápána jako mající svoje vlastní jedinečné schéma kognitivního odhadu, subjektivního pocitu, fyziologického vzrušení, výrazu, tendenci k činu a regulace. V teorii diskrétních emocí se setkáváme s pojmem primární (základní, vrozené) emoce a existencí jistého setu těchto primárních emocí, které jsou rozeznatelné napříč kulturním odlišnostem. Z těchto primárních emocí jsou potom dále tvořené emoce sekundární. Zatímco primární emoce jsou záležitostí pudovou, sekundární záležitostí kognitivní. Teoretici polemizují nad tím, jaké emoce lze považovat za primární. Oblíbený je pohled Paula Ekmana, který uvažuje šest primárních emocí: hněv, znechucení, strach, štěstí, smutek a překvapení. Tento názor je nejvíce podpořen tradiční studií emoční komunikace výrazů v obličeji, které jsou všeobecně vyjádřitelné a rozpoznatelné, viz 2.1.



Obrázek 2.1: Ukázka šesti základních emocí, všeobecně vyjádřitelných výrazem v obličeji.

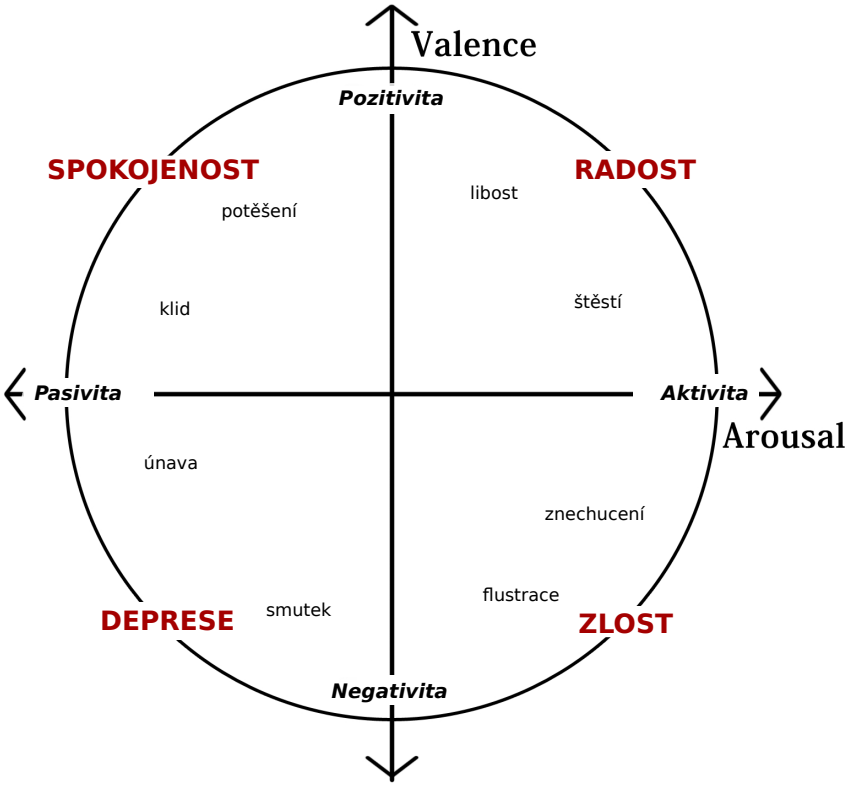
2.2 Teorie dimenzionálních emocí

Dimenzionální pojetí emocí je do značné míry soustředěno pouze na jednu složku emoce — na subjektivní pocit. Identifikace emoce je založena na umístění v prostoru s malým počtem základních dimenzí. Známým se stalo vymezení Wilhelma Maxe Wundta, otce moderní psychologie. Navrhl, že emoce může být vyjádřena třemi charakteristikami: Libost–nelibost, úroveň vzrušení a zážitek napětí–uvolnění [20].

Harold Schlosberg pojmenoval tyto tři emoční dimenze jako: Libost–nelibost, pozornost–odmítnutí a úroveň aktivace [15]. Dále navrhl, že základní strukturu emočních prožitků lze charakterizovat jako uspořádání emocionálních stavů po obvodu kruhu. Tento model, nyní běžně známý jako „circumplex model“, se ukázal jako velmi vlivný. Zastánci tohoto modelu se dnes shodují na tom, že základem kruhového uspořádání jsou dvě ortogonální osy. Mnoho autorů předpokládá existenci dimenze arousal — aktivační dimenze, odrážející energii nebo subjektivní pocit uvolnění — a dimenzi valence — popisující subjektivní úroveň prožitku, měřený od pozitivního nebo příjemného k negativnímu či nepříjemnému. Tento přístup je považován za relevantní při odhadování pocitů, které mohou být vyhýbavé nebo přijímací. Proto je vhodný i pro tento výzkum, který je směřovaný pro použití v call centrech, kde nás zajímá právě postoj k nabídnutí služby či návrhu smlouvy — tento postoj může být přijímací či odmítavý. Pro představu uspořádání emocí dle dvoudimenzionálního modelu do spojitého prostoru viz 2.2.

Použití pouze dvou dimenzí k identifikaci emocí bývá kritizováno tím, že v něm není

možné oddělit určité emocionální stavy. Například strach a zlost. Obě tyto emoce jsou totiž považovány za negativní a nepříjemné a zároveň skrývající v sobě dost energie. Z tohoto důvodů bývá někdy přidávána dimenze další — dominance neboli síla či moc. Ta vyjadřuje kognitivní posouzení potenciálu pro dominanci v určité situaci. Další přidávanou dimenzí bývá intenzita, která dokáže odlišit i kvantitu emoce[9].



Obrázek 2.2: Arousal-valence model pro klasifikaci emocí.

Kapitola 3

Experimentální setup

V této kapitole čtenáře seznámím s daty, která byla použita při tvorbě systémů a také pří následném experimentováním se systémy. Jelikož byla pro účely tohoto výzkumu použita data z evaluace AV+EC 2015 (Audio–Visual Emotion Recognition Challenge), bude nejprve čtenáři sděleno, o jakou soutěž se jedná, a bude představena databáze RECOLA, z níž soutěž čerpala. Následně budou popsána data. Soutěž poskytla audio, video a fyziologické záznamy, z nich vygenerované základní příznakové sady, a také anotace emocí od šesti různých anotátorů včetně Gold standardu — referenčního výstupu, který udává, kde je v promluvě zastoupena jaká emoce. K příznakovým sadám evaluace AV+EC 2015 byl přidán další audio příznakový set — Bottle–Neck příznaky, vygenerované na Fakultě Informačních Technologií VUT v Brně. Nakonec bude představen přístup k posuzování správnosti predikce vzhledem ke skutečným hodnotám emocí, určených Gold standardem.

3.1 AV+EC 2015

Audio–Visual Emotion Recognition Challenge [17] je soutěž v rozpoznávání emocí ze zvukových a vizuálních záznamů. Letošní ročník zahrnoval nově i záznamy fyziologické. Soutěžící zde mají k dispozici několik základních — dále budou v práci označovány jako *baseline* — sad příznaků, vygenerovaných z původních dat; i samotná původní nezpracovaná data, kde se nabízí možnost vygenerování vlastních příznaků. Všechna tato data jsou rozdělena do tří skupin — do trénovací, vývojové a testovací. Pro trénovací a vývojovou skupinu jsou k disposici referenční výstupy — ohodnocení emocí v čase, označované jako *labely*. Úkolem účastníků je postavit a natrénovat systém, na něm vyhodnotit data testovací a tuto predikci následně odeslat.

3.2 Databáze RECOLA

Původní data pochází z multimodální databáze RECOLA (Remote COLaborative and Affective interactions), představená Ringevalem a spol., viz [16]. Obsahem jsou dialogy lidí, mající mezi sebou určitý blízký vztah, kde snímaný je vždy jen jeden z nich. Jedná se o spontánní, nehranou komunikaci, kdy rozhovor je realizován skrze videokonferenci a dvojice společně řeší úlohu spolupráce, konkrétně „Winter survival task“. Členové dialogu řeší situaci, kdy jejich letadlo ztroskotalo v divočině, okolím je hustý les a teplota v noci dosahuje -40° C. Jejich snahou je domluvit se spolu na věcech, které by byly pro záchranu jejich

života v této situaci nezbytné. V této interakci lze tudíž očekávat objevení přijímacích či odmítavých postojů, které jsou pro potřeby tohoto výzkumu relevantní.

Pro vědecké účely je tato databáze zdarma k dispozici na https://diuf.unifr.ch/diva/recola/. Obsahuje 27 dialogů ve francouzštině, všichni účastníci mluví plynule francouzsky, i když rodným jazykem čtyř z nich je italština a dvou z nich němčina. Databáze je věkově i genderově vyvážená. Věk je v rozmezí 19–24 let. Následné rozdělení dat do trénovací, vývojové a testovací sady bylo učiněno tak, aby tato právě zmíněná vyváženost byla zachována.

Anotace pro data vytvořilo šest francouzsky mluvících asistentů. Jednalo se o dvě spojité hodnoty — arousal a valence — pro každých 40 ms z prvních 5–ti minut nahrávky, jelikož právě na začátku dvojice mluvila o strategii a vykazovala emoce. Správnost jednotlivých evaluací byla důkladně ověřována a následně shledána jako navzájem si odpovídající.

3.3 Příznakové sady

V této podkapitole budou popsány jednotlivé příznakové sady, které byly použity ve vytvořených systémech. Bylo použito pět základních — baseline — příznakových sad, které poskytla soutěž AVEC: 1krát audio, 2krát video, 1krát elektrokardiograf a 1krát elektrodermální aktivita. Následně byla přidána navíc další audio sada — příznaky odebrané z úzkého hrdla neuronové sítě.

3.3.1 Audio baseline příznaky

Jako baseline audio příznaky byl použit rozšířený minimalistický akustický set parametrů eGeMAPS (extended Geneva Acoustic Parameter Set) [4]. Tento standardní příznakový set byl vytvořen pro podporu výzkumu rozpoznávání emocí a následně navrhnut jako výchozí (základní) set, ke kterému mají možnost výzkumníci nebo laboratoře přidat jakýkoli další specifický příznakový set. Tento přístup s sebou přináší replikaci poznatků, srovnání mezi studiemi v různých výzkumných ústavech a také lepší náhled do specifických snah jednotlivých laboratoří právě přidáním konkrétního příznakového setu. Velké příznakové sady jsou známé tím, že u nich často dochází k přetrénování na trénovací sadě dat a s tím souvisle znemožnění generalizace, která je naprosté jádro jakékoli snahy o rozpoznávání. Tohle nebezpečí by měl právě zmiňovaný minimalistický set redukovat. Dřívější i nedávné výzkumy potvrzují, že **intenzita** neboli **hlasitost**, **základní frekvence F0**, její **střední hodnota**, **variabilita** a **rozsah**, stejně jako **vysokofrekvenční obsah** neboli **energie mluveného signálu** ukazují přítomnost korelace s prototypickými hlasovými projevy emočních stavů. Například stres ovlivňuje hodnoty intenzity a střední hodnoty základní frekvence F0, hněv či smutek ovlivňuje hodnoty všech parametrů a pocit nudy ovlivňuje variabilitu a rozsah základní frekvence F0. Doporučení pro výběr těchto parametrů bylo zformulováno v Ženevě a dále vyvíjeno na Technické Univerzitě v Mnichově. Výběr byl uskutečňován na základě tří hlavních kritérií:

1. Potenciál onoho příznaku při indikaci fyziologických změn v hlasovém projevu v průběhu afektivních procesů.
2. Četnost použití a množství úspěchů onoho příznaku zmíněných v dřívější literatuře.
3. Teoretický význam onoho příznaku.

V použitém rozšířeném minimalistickém setu jsou obsaženy parametry založené na základní frekvenci F0, amplitudě, spektrální parametry, kepstrální MFCC parametry a parametry časově závislé (temporální). Tyto charakteristiky potom dohromady dávají 102–dimenzionální příznakový set, kde časové okno je dlouhé 3 s a posouvané o 40 ms.

3.3.2 Video baseline příznaky

Výrazy v obličeji hrají velmi důležitou roli ve vyjadřování emocí. Většinou jsou kvantifikovány dvěma typy popisů — vzhledem obličeje a geometrií obličeje. AV+EC poskytla oba tyto popisy v podobě dvou příznakových sad.

Pro extrakci příznaků první sady byly použity Lokální binární vzory, konkrétně Gabor Binary Patterns ze tří kolmých rovin [1], kde v každé rovině je obraz reprezentován histogramem. Výhodou této metody je rotační invariance, tudíž nezávislost na natočení klasifikovaného obrazu. Takto vznikla příznaková sada o výsledné dimensionalitě 84. Druhá příznaková sada, tentokrát 316–ti dimenzionální, byla vygenerována za pomoci orientačních bodů v obličeji neboli *facial landmarks*. Stejně jako u základních audio příznaků byly použity překrývající se segmenty o délce 3 s se vzájemným posunem o 40 ms.

Dodané video příznaky s sebou ale bohužel nesly problém v podobě nerozpoznané tváře v určitých okamžicích nahrávání. Například když snímaná osoba měla skloněnou hlavu k textu. V určitých nahrávkách dosahovalo množství nerozpoznaných rámců až k 40 %, což značně ovlivnilo kvalitu rozpoznávání. Tomuto problému a jeho řešení se budu věnovat v kapitole 5.6.

3.3.3 Fyziologické baseline příznaky

Fyziologické signály jsou také velmi silně korelovány s emocemi, i když nejsou tak dobře znatelné jako ty audio–vizuální. Jelikož existují jisté spory o souvislosti periferní fyziologie a emocích, je vhodné tato samotná měření kombinovat právě se zmíněnými audio–vizuálními.

AV+EC poskytla dvě sady fyziologických příznaků — záznamy elektrokardiografu (ECG) a elektrodermální aktivity (EDA), obě tyto sady byly extrahovány z 4 s dlouhých překrývajících se segmentů, posouvaných vždy o 40 ms. U signálu elektrokardiografu se v závěru jednalo o 54 příznaků, zahrnující srdeční frekvenci, její variabilitu, frekvenci průchodů signálu nulovou úrovní, první 4 statistické momenty, spektrální entropii, střední frekvenci F0 a další charakteristiky a také jejich první derivace.

Záznam elektrodermální aktivity zahrnoval odezvu vodivosti kůže a úroveň vodivosti kůže. Z těchto signálů byly vypočítány obdobné charakteristiky jako ze záznamů elektrokardiografu, což dohromady dalo 60–ti dimenzionální příznakový set. Detailnější informace o baseline příznacích je možné nalézt v článku, zveřejněném jako vodítko pro evaluaci AV+EC 2015, viz [14].

3.3.4 Bottle–Neck příznaky

K základnímu příznakovému setu, který poskytla AV+EC, byl přidán skládaný, tzv. *stacked*, Bottle–Neck příznakový set. Pro extrakci příznaků tohoto typu je zapotřebí architektura sestávající se ze dvou neuronových sítí, kde je na výstupy z první neuronové sítě aplikován časový kontext, tudíž se stává kontextově závislým příznakovým setem pro neuronovou síť druhou — odtud pojmenování *stacked* [10].

Vygenerované příznaky pro vstup do první neuronové vrstvy jsou celkem 222–rozměrné. V základu se jedná o 24 příznaků z melovských filtrů pásové energie a 13–ti příznaků odvozených ze základní frekvence F0. Na tento vektor o velikosti 37 prvků je poté aplikován odpočet středních hodnot a dále Hammingovo okno následované diskrétní kosinovou transformací, uchovávající prvních šest bází, aplikovanou v časové oblasti pro každý z příznaků: (24 + 13) * 6 = 222.

Architektura použitých dvou neuronových sítí je ekvivalentní. Obě mají 4 skryté vrstvy, z nichž každá kromě třetí, nazývané úzké hrdlo (Bottle–Neck), obsahuje 1500 buněk. Vrstva úzkého hrdla má prvků pouze 80. Jako konečný Bottle–Neck příznakový set jsou považovány hodnoty výstupů z neuronů vrstvy úzkého hrdla druhé neuronové sítě. Obě sítě jsou trénovány na kontextově závislých fonémech z 11–ti různých jazyků. Bližší a detailnější popis těchto příznaků viz [5].

3.4 Evaluační metrika — koeficient korelační shody

Anglicky Concordance Correlation Coefficient (CCC), slouží k měření korelace mezi predikcí a skutečnými hodnotami emocí (Gold standardem). Kombinuje v sobě Pearsonův korelační koeficient ze dvou časových řad se střední kvadratickou odchylkou:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2\sigma_y^2 + (\mu_x - \mu_x)^2}. \tag{3.1}$$

Získané hodnoty se pohybují mezi -1 a 1, kde 1 značí absolutní shodu, -1 opačnou korelaci a 0 žádnou korelaci. Tento koeficient je použit v celé práci jako měřítko úspěšnosti jednotlivých systémů a zároveň pro možnost porovnání s organizátory, kteří poskytli výsledky vyhodnocené na jejich vlastním systému s baseline příznaky, viz [14].

Kapitola 4

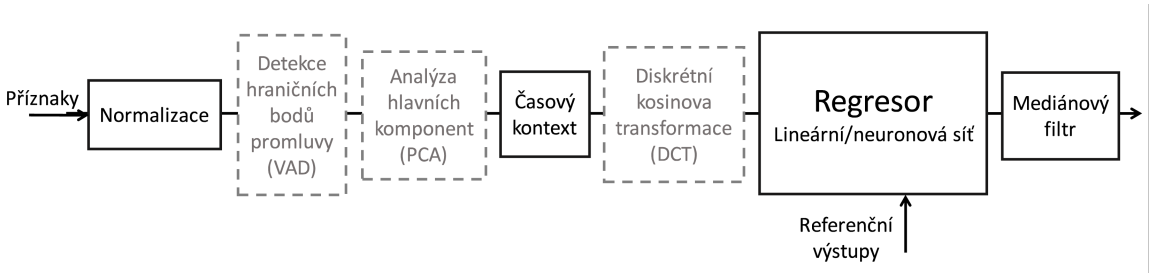
Popis systémů

Vzhledem k rozmanitosti dat ke zpracovávání bylo vhodné zvolit přístup konstrukce několika samostatných systémů (pro každou příznakovou sadu zvlášť) a následně nejlepší z nich spojit do fúze. Několik prvotních experimentů ukázalo, že je vhodnější mít i pro každou složku emoce systémy zvlášť. Výjimka byla pouze u systémů používajících Bottle–Neck příznaky — tam byl více zkoumán i systém rozpoznávající obě emoční dimenze. Celkem tedy bylo vytvořeno $6 + 6 + 1 = 13$ systémů, každý používající vždy jednu z modalit, a následně 2 systémy fúzní, používající 2 sady výstupů ze samostatných systémů. Základní schéma samostatného systému je vidět na obrázku 4.1. Detailnější popis fúzního systému bude popsán později.

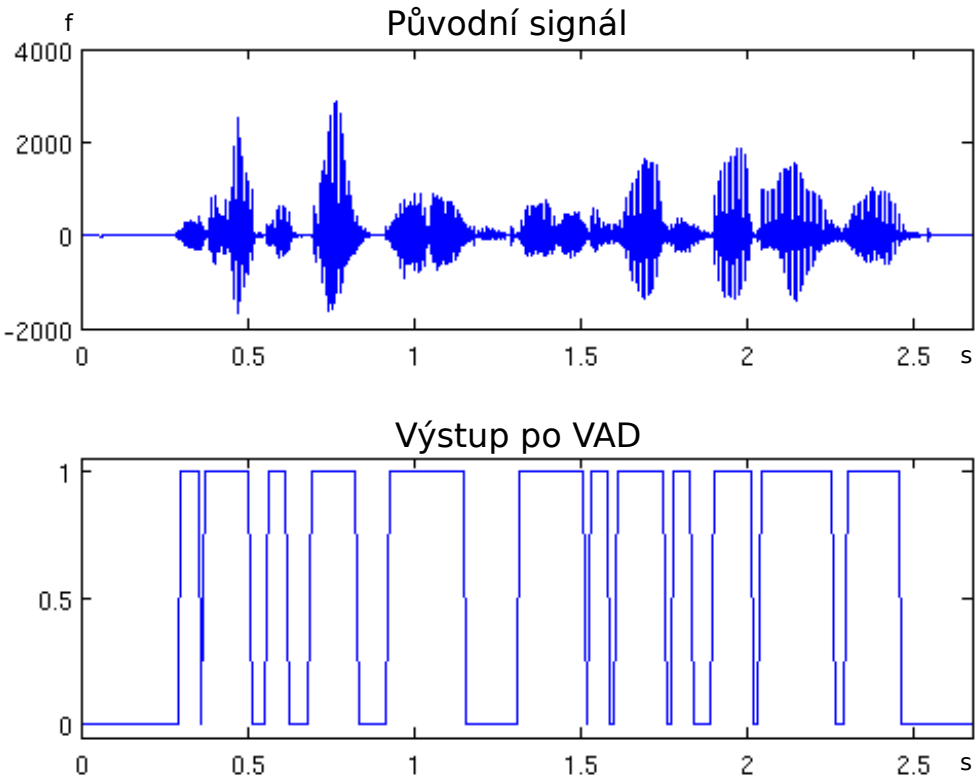
V této kapitole se budu zabývat popisem systémů tak, že představím metody předzpracování, trénování a po–zpracování dat. Ty složitější z nich popíšu podrobněji a naznačím jejich realizaci.

4.1 Předzpracování dat

Jistá úprava vyextrahovaných příznaků z původních audio, video či fyziologických dat před tím, než se trénují je nezbytná pro dosažení reprezentativních výsledků. Na tuto část by tudíž měl být kladen velký důraz a mělo by být vyzkoušeno více přístupů a vybrat z nich ty nejvodnější pro daný typ úlohy. Často používané metody jsou normalizace, určování hraničních bodů promluvy, analýza hlavních komponent a modelování delšího časového průběhu signálu — aplikace časového kontextu. Právě tyto metody budou nyní popsány



Obrázek 4.1: Schéma samostatného systému pro rozpoznávání emocí. Metody vyobrazené tučně jsou používány vždy, ty světlejší pouze v konkrétních případech.



Obrázek 4.2: Ukázka aplikace detekce řečové aktivity. Nahoře původní signál, dole příslušné hodnoty detekce řečové aktivity.[1]

a v kapitole 5 následně aplikovány na jednotlivé systémy.

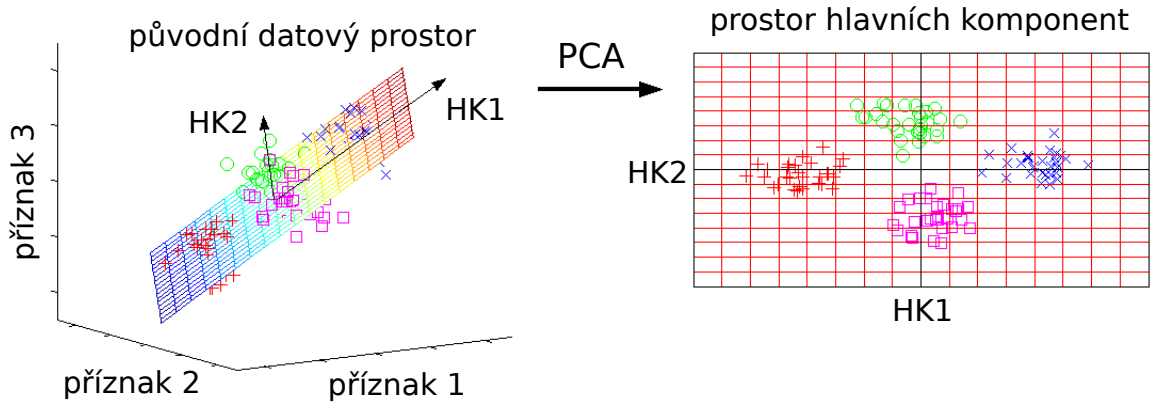
4.1.1 Normalizace

Rozsahy hodnot jednotlivých příznakových dimenzí jsou přímo po extrakci různé. Dohromady jsou tudíž špatně vypovídající. K odstranění tohoto problému slouží normalizace. Nejčastěji používaná a také ta, která je použita pro příznaky v této práci, je normalizace posunutím středních hodnot jednotlivých dimenzí na nulu a dosažením jednotkové variance (zero mean, unit variance). Takto normalizované hodnoty lze získat vypočtením středních hodnot jednotlivých dimenzí skrze celý data-set a následným odečtením těchto hodnot od konkrétních hodnot v dané dimenzi. Tím získáme onu střední hodnotu rovnu nule. Jednotkovou varianci poté získáme vydělením všech dimenzí standardní odchylkou ze všech dat.

4.1.2 Detekce řečové aktivity

V angličtině je tato technika známá pod názvem Voice Activity Detection (VAD). Jedná se o techniku, používanou pro rozlišení mluvených úseků a úseků ticha v nahrávce. Má

[1] Zdroj: http://practicalcryptography.com/miscellaneous/machine-learning/voice-activity-detection-vad-tutorial/.



Obrázek 4.3: Ukázka promítnutí dat do jiného souřadného systému pomocí PCA, vedoucí k redukci původní dimensionality. [2]

široké využití ve zpracovávání řečových signálů, například u Voice over IP (VoIP) pro redukci přenášených dat — přenášet ticho je zbytečné. V této konkrétní úloze bylo určování hraničních bodů promluvy vyzkoušeno z důvodu, že zpracovávaná data jsou dialogy osob, kde nás zajímá hlas pouze jedné z nich. V pozadí je ale slyšet i hlas druhého účastníka rozhovoru. Cílem tedy bylo tato místa, kde promlouvá vedlejší účastník dialogu, odstranit. Úseky se slabým zachycením zvuku byly odstraněny nastavením vhodného prahu hodnoty nultého MFCC příznaku, který koreluje s energií signálu. Výsledkem byl vektor jedniček a nul, kde nula symbolizovala ticho a jednička hlas. Ukázka podobného typu detektoru je na obrázku 4.2.

4.1.3 Redukce dimenzí pomocí analýzy hlavních komponent

V angličtině je tato metoda nazývána jako Principal Conponent Analysis (PCA). Jde o metodu, promítající data do jiného souřadného systému. Slouží k dekorelaci příznaků, redukci dimensionality a maximalizaci variability. PCA je počítána určením vlastních vektorů a vlastních čísel kovarianční matice [8]. Ukázka této metody viz obrázek 4.3.

Kovarianční matice je symetrická, čtvercová matice, o počtu dimenzí původní matice příznaků. Na její diagonále leží variance — kvadráty směrodatných odchylek jednotlivých dimenzí. Zbytek matice je vyplněn kovariancemi — ty určují, jaká kovariance (závislost) je mezi jednotlivými dimenzemi příznaků. Snahou analýzy hlavních komponent je snížit dimensionalitu minimalizací kovariance a maximalizovat varianci.

Vlastní vektory kovarianční matice se seřadí podle hodnot odpovídajících vlastních čísel od nejvyšší k nejnižší. Několik prvních vlastních vektorů (hlavních komponent) tvoří poté osy nového souřadného systému. Redukce dimensionality musí být ovšem pouze taková, aby vzniklá chyba neměla vliv na kvalitu rozpoznávání. Obvykle je použito následující kritérium: $\frac{\sum_{i=1}^{K}\lambda_i}{\sum_{i=1}^{N}\lambda_i} > \theta$, kde K je počet dimenzí, na které chceme zredukovat původní prostor o dimensionalitě N, λ_i je konkrétní vlastní číslo a θ označuje spolehlivost nového datového setu. Jako θ se používá často hodnota 0.9 nebo 0.95, což znamená, že ztráta variability a zbytková kovariance oproti původnímu datovému setu nepřesáhne 10 % [8].

[2] Zdroj: http://www.nlpca.org/pca_principal_component_analysis.html.

4.1.4 Aplikace časového kontextu

S vědomostí toho, že samotné trvání emoce kolísá mezi 0.5 až 4 sekundami, bychom intuitivně měli v době zpracovávání jednoho rámce mít i informace o okolních rámcích. K tomu slouží metoda aplikace časového kontextu. Přidání kontextové informace se dá řešit různými způsoby:

- Recurrent Neural Networks (RNN), popř. Long Short Term Memory Networks (LSTMs) — druhy neuronových sítí, obsahující v sobě cykly, jimiž lze uchovávat informace o již přečtených datech v době, kdy se zpracovávají data pozdější. Kontextová informace je tedy uchovávána na úrovni trénovacího modelu. LSTMs mají poněkud složitější strukturu a dají se využít i při problému dlouhodobější kontextové závislosti.
- Seskupení rámců do jednoho vektoru na úrovni příznaků, tedy ve fázi předzpracování. Tato varianta vede ke zvyšování dimensionality dat.
- Seskupení rámců na úrovni příznaků spolu s použitím diskrétní kosinové transformace. Tento přístup je použit v této bakalářské práci a bude později důkladněji vysvětlen. Zvolen byl z důvodu úspěchu této metody, prezentované v článku [5], při použití Bottle–Neck příznaků na rozpoznávání řeči.

4.1.5 Diskrétní kosinová transformace

Diskrétní kosinová transformace, známá pod označením DCT (Discrete Cosine transform), nalézá uplatnění hlavně při zpracovávání obrazu — konkrétně je využita ve ztrátově kompresním formátu JPEG. Využitelná je však i ve zpracovávání řeči [12, 5].

Diskrétní kosinová transformace patří mezi lineární transformace příbuzné Fourierově transformaci. Oproti diskrétní Fourierově transformaci však používá pouze reálné koeficienty. Výsledkem aplikace diskrétní kosinové transformace je posloupnost součtů kosinových funkcí oscilujících na různých frekvencích. Pro kompresi dat se tedy zpravidla bere několik prvních koeficientů této posloupnosti, brané jako báze, do kterých se promítne původní prostor, jelikož malé vysokofrekvenční složky mohou být zanedbány.

Formulí pro výpočet diskrétní kosinové transformace je více, existuje i inverzní diskrétní kosinová transformace. V systémech popisovaných v této práci je použita nejznámější varianta, označována jako DCT–II nebo pouze jako „the DCT“ [19]:

$$\sum_{n=0}^{N-1} x_n cos[\frac{\pi}{N}(n+\frac{1}{2})k] \qquad k = 0, \dots N-1 \tag{4.1}$$

Diskrétní kosinová transformace má v popisovaných systémech své místo po aplikaci časového kontextu následovaným vynásobením spojených rámců Hammingovým oknem. Nad otázkou, kolik bází je potřeba zanechat pro nejlepší výsledky, jsou prováděny experimenty prezentované v kapitole 5.1.4.

4.2 Trénování a po–zpracování dat

Při trénování a následném vyhodnocování hodnot emocí je použita regrese, jelikož hlavní referenční hodnoty (Gold Standard) jsou spojitého charakteru. V některých případech je

použita neuronová síť, ale většinově pouze regrese lineární. Vzhledem k tomu, že jsou k dispozici hodnocení šesti různých anotátorů, je možné experimentovat s tím, která z těchto hodnocení jsou vhodná pro trénování systémů. Dále je možné experimentovat s velikostí posunutí těchto anotací a eliminovat tím reakční prodlevu anotátorů — referenční hodnoty jsou udávány v řádu 40 ms, což je pro člověka nerozlišitelný časový úsek, tudíž anotace nemůžou sedět naprosto přesně.

Na výstupní hodnoty z regresoru je aplikován mediánový filtr, protože predikce spojitých hodnot bývají chaotické. Velikost mediánového filtru je určována pro různé systémy opět experimentálně.

Kapitola 5

Experimenty se systémy

Tato kapitola se zabývá experimenty, které vedly k nalezení optimálních parametrů pro předzpracování, trénování a po–zpracování dat. Experimenty jsou prováděny se všemi postavenými systémy zvlášť, kde největší důraz je kladen na systémy používající řeč jako modalitu k rozpoznávání emocí. Nejčastěji jsou tudíž zmiňovány výsledky a pokroky systémů používající eGeMAPS nebo Bottle–Neck příznaky. Výstupy všech samostatných systémů lze následně seřadit dle přesnosti predikce, viz tabulka 5.2. Ty nejlepší z nich jsou použity pro finální fúzi, jejíž tvorba a cesta k nalezení správného nastavení je taktéž popsána zde. Kapitola pokračuje popisem řešení problému nerozpoznaných rámců ve video příznacích a je zakončena zmínkou o úspěšném systému rozpoznávající obě emoční dimenze používajíc Bottle–Neck příznaky.

5.1 Výsledky metod předzpracování

Jak již bylo zmíněno, výběr metod předzpracování příznaků a způsob jejich realizace před vstupem do regresoru má významný vliv na konečnou přesnost predikce. Experimenty tuto teorii potvrzují a dále ukazují odlišnosti v přístupu k těmto metodám v závislosti na zkoumané modalitě či konkrétních příznacích.

5.1.1 Detekce řečové aktivity

Detekce řečové aktivity byla vyzkoušena za účelem zlepšení predikce v souvislosti s odstraněním rámců, kde není přítomen hlas mluvčího, jehož emoce má být rozpoznávána. Tato technika se používá například u identifikace mluvčího z hlasu, kde je třeba eliminovat rámce, kde je mluvčí, kterého chceme identifikovat, ticho, aby se systém nenatrénoval na okolních vlivech jako je třeba zvuk prostředí, ve kterém je nahrávka pořizována. V souvislosti s touto znalostí mělo význam vyzkoušet detekci řečové aktivity promluvy i pro problém rozpoznávání emocí.

Postup po detekci řečové aktivity byl následující: Systém byl trénován na datech, ze kterých byly odstraněny rámce, identifikované jako ticho. Výstupy ze systémů byly však porovnávány s celými nahrávkami, včetně úseků, kde zkoumaná osoba mlčela. Odstraněním těchto rámců i z referenčních výstupů by sice ke zlepšení došlo, ale pro účely evaluace tento přístup možný nebyl.

Detekce řečové aktivity v promluvě nakonec tedy nevedla k zlepšení predikce. Metoda byla vyzkoušena na samostatných systémech používajících dva typy audio příznaků — eGeMAPS a Bottle–Neck. V porovnání se systémem nepoužívající tuto metodu, je systém tré-

Arousal

	Audio		Video		Fyziologie	
	eGeMAPS	Bottle–Neck	geometrie	vzhled	EDA	ECG
Analýza hlavních komponent	z 102 na 13	80*	316*	z 84 na 80	54*	60*
Velikost kontextové informace	141	161	3	3	–	–
Podvzorkování	–	každý 5. rámec	–	–	–	–
Báze diskrétní kosinové transformace	–	7	–	–	–	–
Velikost mediánového filtru	201	121	101	181	1	181
Anotátoři pro trénování	1 + 3	1 + 2 + 3	1 + 2 + 3 + GS**	3 + GS**	všichni	1
Váhy anotátorů pro konečný výstup	0.875 : 0.125	0.34 : 0.16 : 0.5	0.1 : 0.1 : 0.7 : 0.1	0.75 : 0.25	průměr	1

Valence

	Audio		Video		Fyziologie	
	eGeMAPS	Bottle–Neck	geometrie	vzhled	EDA	ECG
Analýza hlavních komponent	z 102 na 24	z 80 na 13	316*	z 84 na 70	54*	z 60 na 30
Velikost kontextové informace	181	181	3	101	11	11
Podvzorkování	–	každý 4. rámec	–	každý 2. rámec	–	–
Báze diskrétní kosinové transformace	–	30	–	3	1	2
Velikost mediánového filtru	31	81	41	41	41	61
Anotátoři pro trénování	1 + 2 + 3	1 + 2	1 + 3	1 + 2	1 + 2 + 3	všichni
Váhy anotátorů pro konečný výstup	0.34 : 0.16 : 0.5	0.5 : 0.5	0.375 : 0.625	0.625 : 0.375	0.4 : 0.4 : 0.2	průměr

Tabulka 5.1: Hodnoty, parametry a nastavení jednotlivých metod předzpracování, trénování a po–zpracování samostatných systémů, jenž vedly k dosažení nejlepší predikce na vývojové sadě. *Bez redukce pomocí analýzy hlavních komponent. **GS = Gold standard.

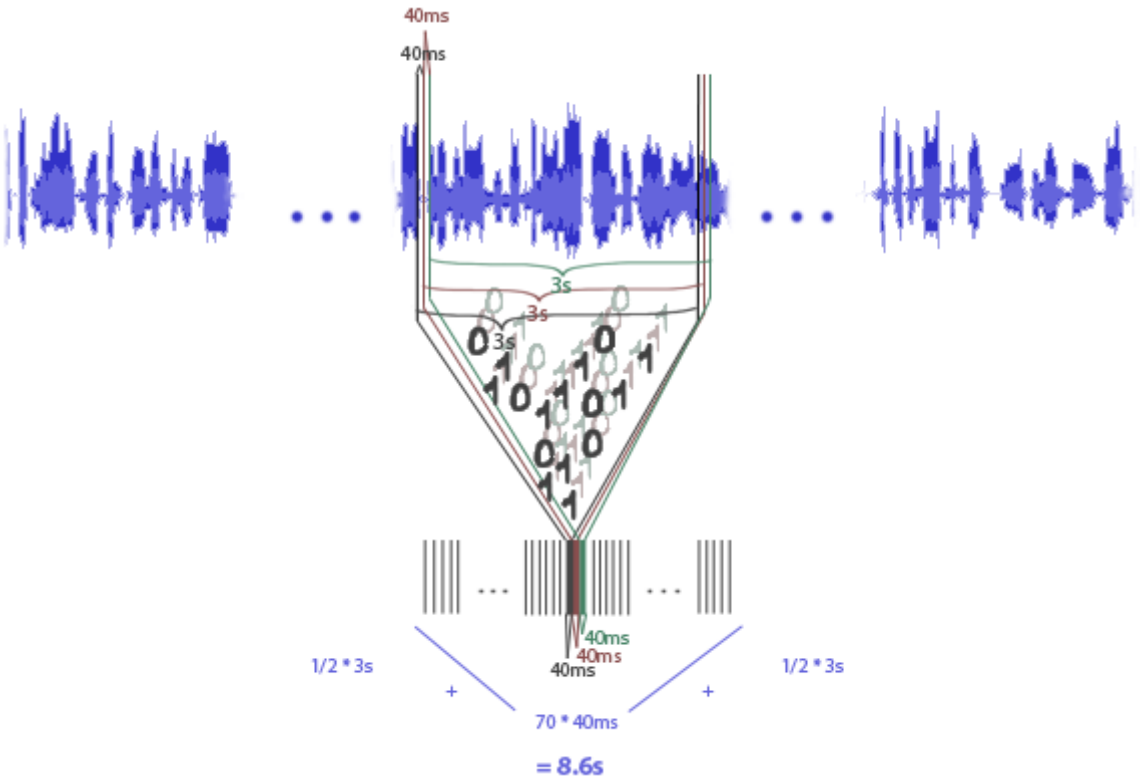
novaný pouze na úseky s řečí horší v predikci valence o 13 % relativně a v predikci arousalu o 11 % relativně. Tento výsledek lze připočítávat tomu, že pozornost je cílena na jednoho člena dialogu, jehož emoce jsou třeba rozpoznávat i během toho, co hovoří člen druhý, tudíž i tehdy, kdy je v nahrávce ticho. Detekcí řečové aktivity jsou ale tyto úseky ticha odstraněny — do ticha lze zahrnout i jakékoli jiné zvukové projevy mluvčího, než je samotná řeč (smích, povzdechy, ...). Z tohoto důvodu zmíněná metoda ve finálním systému použita nebyla.

5.1.2 Aplikace časového kontextu

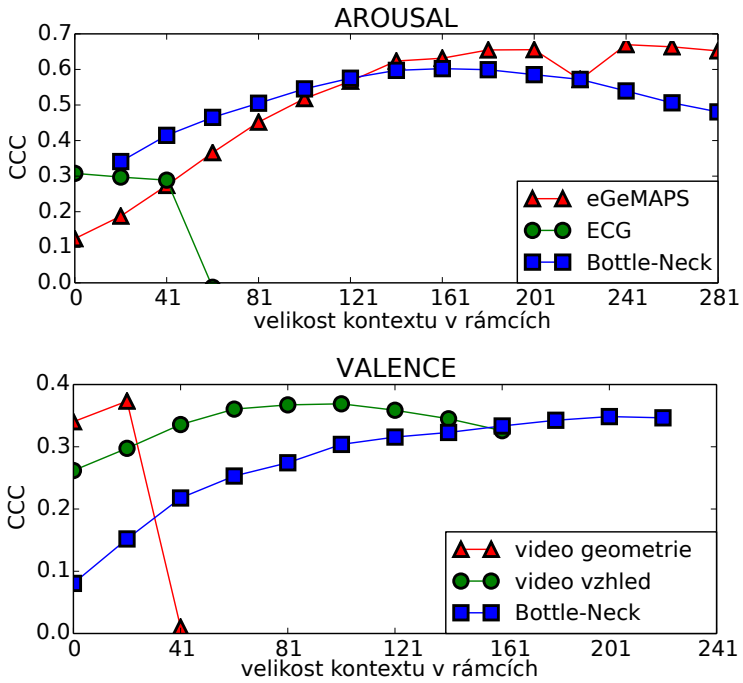
Baseline příznaky jsou extrahovány z 3-4 s dlouhých kusů nahrávek, tudíž v sobě již nesoucí jistou kontextovou informaci. Tato informace je ale zprůměrovaná, tudíž v sobě neuchovává informaci o vývoji v čase neboli trendy. Experimenty ukazují, že tento typ chybějící kontextové informaci je třeba dodat. Obecně je míra této kontextové informace nejvíce významná u audio příznaků, což dává smysl v souvislosti s tím, že tón řeči se rychle mění, zatímco emoce ještě může doznívat. U videa je tento jev slabší a u fyziologických záznamů bychom spíš chtěli tento kontext ubírat — například rychlost bušení srdce se nám neuklidní hned po tom, co už se cítíme lépe.

Příklad tvorby časového kontextu o velikosti 141 rámců (70 + 1 + 70) je ukázán na obrázku 5.1. Je zde zobrazena i okrajová kontextová informace, obsažená již ve vyextrahovaném příznaku. Proto skutečná délka kontextové informace je *počet sloučených rámců * velikost rámce + velikost úseku pro extrakci/2*.

Aplikace časového kontextu je tedy první metoda, která posunula samostatné systémy používající audio příznaky silně dopředu. Obecně platí nutnost uchovávání větší kontextové



Obrázek 5.1: Ukázka tvorby kontextové informace, která je použita při rozpoznávání.



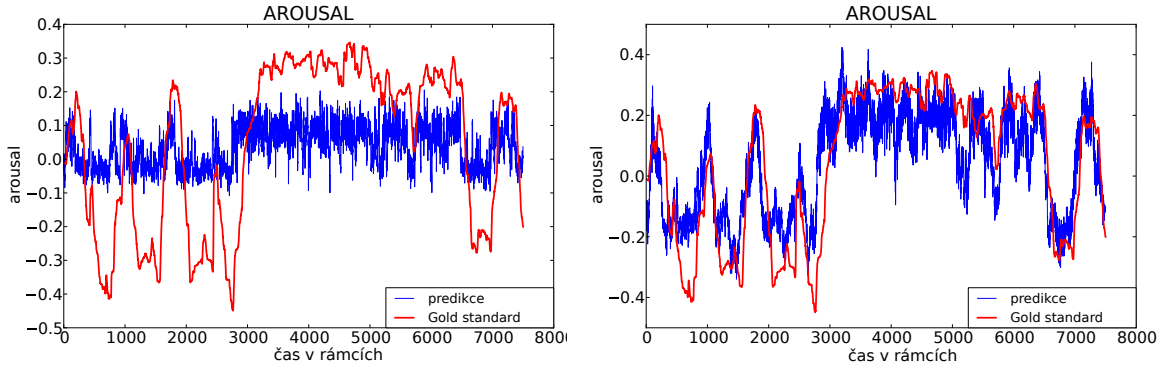
Obrázek 5.2: Závislost CCC na velikosti časového kontextu v rámcích. Do časového kontextu je započítán současný rámec a stejný počet rámců dopředu i dozadu.

informace u predikce valence, než u arousalu. Největší kontext byl použit u predikce valence z obou audio příznaků — eGeMAPS a Bottle–Neck — a to skrze 181 rámců (skutečná délka je potom asi 7 s + okolní kontextová informace, obsažená již ve vyextrahovaném příznaku). Ostatní hodnoty je možné vidět v shrnující tabulce 5.1. Délka kontextu je zde udávána v počtu sdružených sousedních rámců do jednoho, kde velikost jednoho rámce se rovná 40 ms.

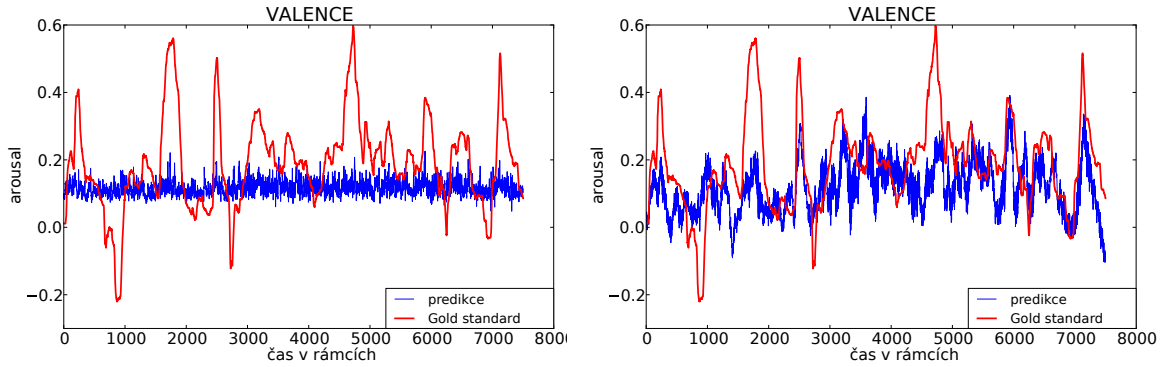
Díky aplikaci časového kontextu došlo ke zlepšení predikce arousalu z původní hodnoty CCC = 0.052 na hodnotu CCC = 0.370 a valence z CCC = 0.071 na CCC = 0.221 při použití audio eGeMAPS příznaků. Systémy používající audio Bottle–Neck příznaky na tom byly ještě o něco lépe a u nich byla predikce valence CCC = 0.304 a predikce arousalu CCC = 0.539. Systémy již v této fázi předčily baseline výsledky, prezentované v článku zveřejněném k evaluaci AV+EC 2015 jako vodítko [14]. Systém používající eGeMAPS příznaky předčil baseline o 70 % a systém používající Bottle–Neck příznaky o 139 % relativně k hodnotě průměru výsledku baseline systému predikce valence (CCC = 0.069) a arousalu (CCC = 0.287). Ukázku závislosti velikosti kontextu na úspěšnosti predikce systémů je možné vidět v grafu 5.2 a ukázku zlepšení predikce na jedné konkrétní nahrávce v grafu 5.3 a 5.4.

5.1.3 Redukce dimensionality pomocí analýzy hlavních komponent

Všechny systémy byly testovány na redukci dimenzí pomocí analýzy hlavních komponent. Například u samostatného systému používající audio eGeMAPS příznaky byla určitá redukce nutná, z důvodu potřeby uchovávat velkou kontextovou informaci a s tím související zaplnění paměti. Experimentováním byl nalezen kompromis, kdy redukce pomocí analýzy hlavních komponent přinášela užitek ve formě dekorelace příznaků a zároveň ještě neubrala



Obrázek 5.3: Přesnosti predikce emoční dimenze arousal před (vpravo) a po (vlevo) aplikaci časového kontextu skrze 161 rámců.



Obrázek 5.4: Přesnosti predikce emoční dimenze valence před (vpravo) a po (vlevo) aplikaci časového kontextu skrze 181 rámců.

na vypovídající hodnotě.

Analýza hlavních komponent s využitím velkého časového kontextu potom zafungovala příznivě pro samostatné systémy používající audio eGeMAPS příznaky, u valence byla dimensionalita snížena z 102 na 24 a u arousalu z 102 na 13. V systémech používající příznaky z videa nebo fyziologických záznamů naopak byla provedena redukce dimensionality velice malá nebo vůbec žádná. Konkrétní číselné hodnoty těchto redukcí můžete vidět v tabulce 5.1.

5.1.4 Diskrétní kosinová transformace

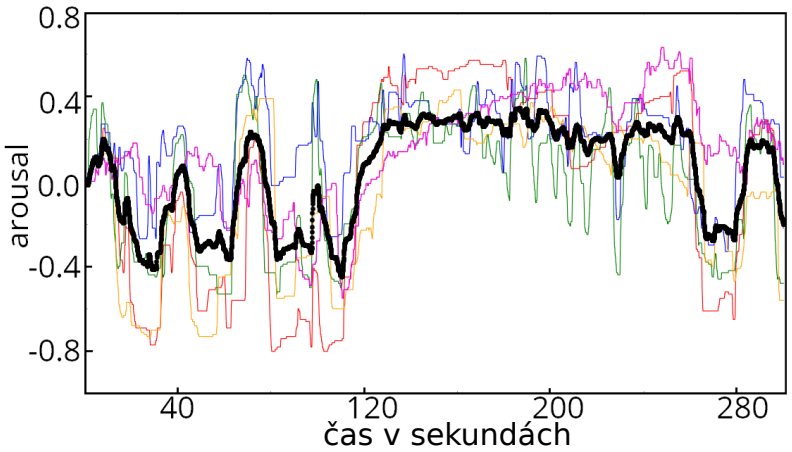
Spojování sousedních rámců za účelem přítomnosti kontextové informace vede k zvyšování dimensionality. Obecně je v této informaci obsaženo i určité množství redundance — již samotné příznaky jsou generovány tak, aby v nich byla jistá kontextová informace už obsažena. Pro redukci tohoto jevu spolu s využitím maximálního potenciálu kontextové informace je použita diskrétní kosinová transformace.

Diskrétní kosinová transformace neprospěla úplně všem systémům, ale některým naopak prospěla výrazně. Navíc její použití bylo výhodné až společně s aplikací mediánového filtru, o kterém bude řeč až později. Za podmínek ponechání pouze prvních 7–mi bází diskrétní

kosinové transformace se zvýšila přesnost samostatného systému rozpoznávající arousal a používající Bottle–Neck příznaky o 10 %. Systém na stejné bázi rozpoznávající valenci se posunul ponecháním těchto prvních 30-ti bází o 13 %. V systému používající video–vzhled příznaky pro rozpoznávání valence byly použity první 3 báze a zlepšení bylo o 10 %. Přesnost valence z video–vzhled příznaků v této fázi s CCC = 0.296 předčila AV+EC 2015 baseline systém o 8 %. V systémech používající audio eGeMAPS příznaky diskrétní kosinová transformace použita nebyla.

Obecně je možné říci, že diskrétní kosinová transformace více prospívá systémům rozpoznávající valenci.

Ještě před touto projekcí lze zmenšit datový prostor snížením počtu rámců uvažovaných v přidané kontextové informaci tak, aby délka této informace byla zachována. Lze tak učinit započítáním například pouze každého druhého rámce. Hodnoty těchto *podvzorkování* lze opět nalézt ve shrnující tabulce 5.1.

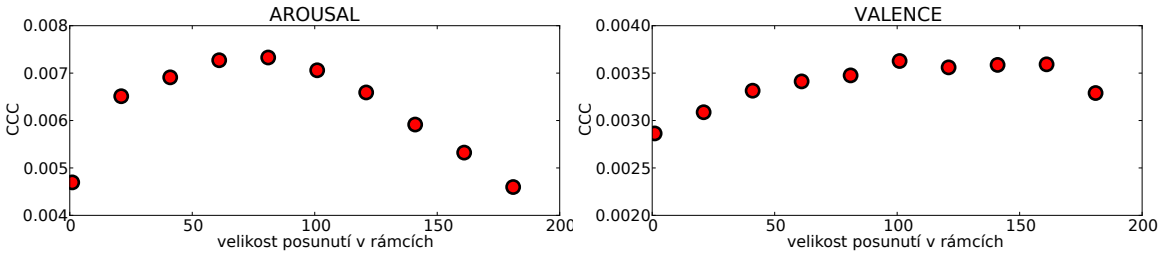


Obrázek 5.5: Ukázka hodnocení emoční dimenze arousal šesti různých anotátorů jedné promluvy, včetně Gold standardu — tučná černá linka.

5.2 Výsledky trénování

Ve všech samostatných systémech rozpoznávající arousal a ve všech samostatných systémech rozpoznávající valenci kromě systému používající příznaky geometrie obličeje — tam je použita neuronová síť — je použit lineární regresor. Neuronová síť má jednu skrytou vrstvou a topologii 948–474–3. Velikost skryté vrstvy se rovná polovině počtu vstupních dat, kde tato hodnota byla experimentálně vyhodnocena jako nejvhodnější. Zmíněný výběr trénovacího nástroje byl učiněn po několika prvotních experimentech. Mohlo by se zdát zvláštní, že pro většinou systémů lépe fungovala lineární regrese než v dnešní době velmi populární neuronová síť. Vzhledem k tomu, že hlavní síla neuronových sítí je ale učit se z velkého množství příkladů určitého jevu, je tento výsledek celkem logicky vysvětlitelný. V data–setu, který je k dispozici pro účely této práce není dostatek různých příkladů konkrétních hodnot emocí, spíše se dá říci, že hodnoty emocí se vyskytují jedinečně. Tím se dá vysvětlit, že lineární regrese podává ve většině případů lepší výsledky, než neuronová síť.

Možnosti pro trénování regresoru jsou rozmanité. K dispozici je šest různých anotací a k nim jeden referenční výstup — Gold standard (normalizovaný a průměrovaný), detailněji



Obrázek 5.6: Závislost CCC mezi Bottle–Neck příznakovým setem a posunutou referencí.

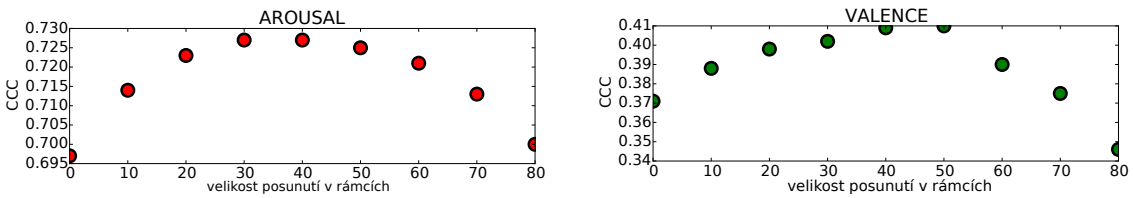
popsáno v [14]. Jednotlivé anotace mezi sebou mají jisté odchylky. Pro představu jsou anotace emoční dimenze arousal jedné z nahrávek ukázány na obrázku 5.5. Experimentálně bylo zjištěno, že pro emoční dimenzi arousal, zjišťovanou ze samostatného systému používající Bottle–Neck příznaky, je nejvýhodnější trénovat regresor na prvním až třetím anotátorovi. Tato změna přinesla zlepšení systému o dalších 7 %. Pro emoční dimenzi valence se ukázalo jako optimální použití prvních dvou anotátorů, systém se tím zlepšil o 4 %. Systém pro arousal, používající audio eGeMAPS příznaky se za použití prvního a třetího anotátora zlepšil o 35 % a systém pro valenci, používající stejné příznaky, se výběrem prvních třech anotátorů zlepšil relativně o 5 %.

Vzhledem k velmi malému množství dat k trénování byla snaha o rozšíření trénovací sady přidáním dat ze sady vývojové. Trénování systému tedy probíhalo na všech trénovacích datech a na všech vývojových kromě jednoho, které bylo zrovna testováno na úspěšnost (technika známá pod názvem *leave-one-out*). Kupodivu tato metoda nevedla k lepší úspěšnosti než ta, při které byla požita původní menší trénovací sada.

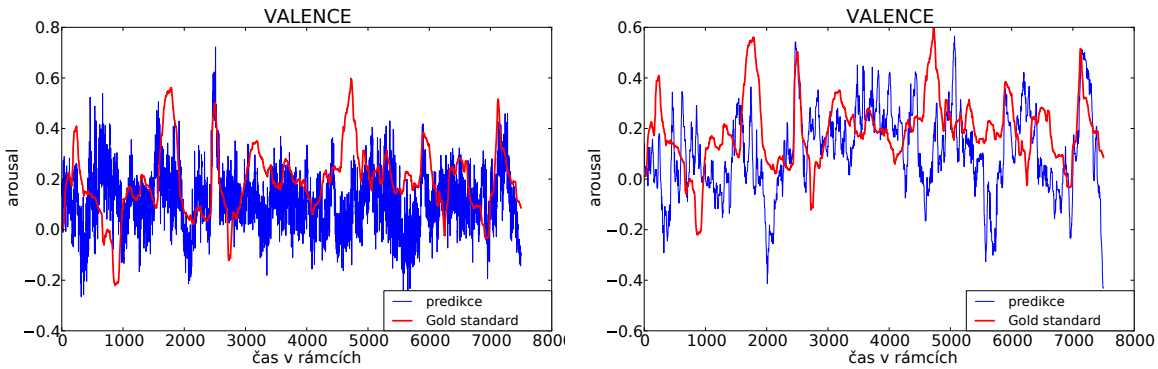
5.2.1 Problém reakční prodlevy

Regresor tak, jak je použitý v této práci, produkuje predikce emocí v řádu 40 ms. Tak krátký časový úsek je však pro lidi nerozlišitelný. Referenční hodnoty emocí tedy musí mít jistou časovou odchylku, které říkáme reakční prodleva, i když byly tvořeny profesionálními anotátory. Znamená to, že příznaky vyextrahované z konkrétního časového úseku patří k hodnotě emoce, která je ve skutečnosti o něco dříve. Tento jev může být také způsoben tím, že samotná extrakce již v sobě zahrnuje časový kontext — například u audio příznaků eGeMAPS jde o kontext 3 s. Přiřazeny jsou ale k rámci o velikosti 40 ms, otázkou tedy je, kam přesně v úseku 3 s tato hodnota emocí patří. Zmíněnou odchylku je samozřejmě vhodné řešit.

K odhadu délky reakční prodlevy byla použita metoda maximalizace CCC mezi přízna-



Obrázek 5.7: Závislost posunutí referenčních hodnot pro trénování na úspěšnosti systémů používajících Bottle–Neck příznaky.



Obrázek 5.8: Přesnosti predikce emoční dimenze valence před (vpravo) a po (vlevo) aplikaci mediánového filtru skrze 31 rámců.

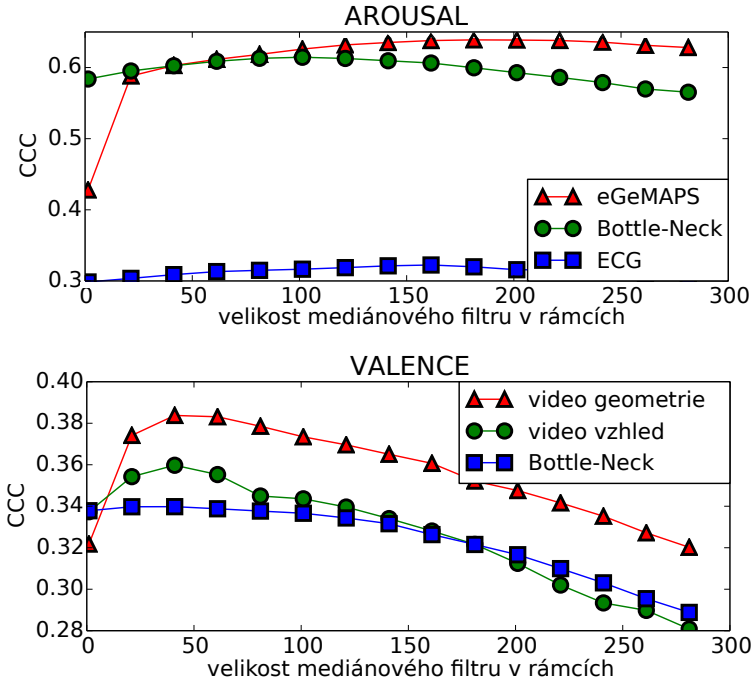
kovým setem a referenčními hodnotami (inspirováno [6]), která spočívala ve vyhodnocení korelačního koeficientu (CCC) mezi jednotlivými příznakovými sety a různě posunutými referenčními hodnotami. Takto provedená analýza vedla například u samostatného systému používající Bottle–Neck příznaky k výsledku zaznamenaném v grafu 5.6. S tímto odhadem velikosti posunutí byl systém trénován, a dále byla tato hodnota laděna k dosažení nejlepšího výsledku. Ukázku průběhu těchto pokusů můžete vidět na grafu 5.7. Z grafů je patrné, že konečná hodnota posunutí se liší od té původní vypočtené metodou maximalizace CCC mezi příznakovým setem a posunutou referencí. Pro systém využívající Bottle–Neck příznaky bylo tedy nakonec použito posunutí o délce 40/50 rámců (arousal/valence), což odpovídá asi 1.8 s. Pro samostatné systémy používající audio eGeMAPS příznaky se ukázalo jako nejvýhodnější posunutí o 70 rámců (necelé 3 s), u valence i u arousalu. Toto posunutí vedlo u valence k relativnímu zlepšení o 11 % a u arousalu o 5 %.

5.3 Výsledky po–zpracování

Výsledky produkované regresorem byly podle očekávání poněkud kostrbaté. Pro vylepšení těchto výstupů se nabízely dvě operace:

1. Byl–li systém trénován na více anotátorech, jeho výstup byl přirozeně také vícerozměrný. Jednotlivým výstupům lze přiřadit váhy a ladit je tak dlouho, dokud nebude nalezen optimální poměr těchto výstupů. Konkrétní hodnoty vah pro optimalizaci výsledků jsou v tabulce 5.1.

2. Pro konečné vyhlazení výstupů je aplikován mediánový filtr. Po jeho aplikaci došlo opět ke znatelnému zlepšení. Mediánový filtr byl aplikován na systémy ve fázi, kdy používaly pouze metodu aplikace časového kontextu, tudíž nyní uvedené zlepšení je bráno relativně k výsledkům dosažených aplikací časového kontextu, prezentované v kapitole 5.1.2. Zlepšení systémů používající audio eGeMAPS příznaky bylo potom pro emoční dimenzi arousal 7 % (filtr skrze 201 rámců ≙ 8 s) a pro valenci 26 % (filtr skrze 31 rámců ≙ 1s) — ukázka tohoto zlepšení na jedné nahrávce je na obrázcích 5.8. U systému používající Bottle–Neck příznaky bylo toto zlepšení v průměru pouze o 1 %. Jednotlivé velikosti použitého mediánového filtru je opět možné vidět v tabulce 5.1. Graf 5.9 ukazuje závislost různých modalit na velikosti tohoto kontextu. Zpravidla platí, že větší filtr je použit při rozpoznávání arousalu než při rozpoznávání valence.



Obrázek 5.9: Závislost CCC na délce mediánového filtru v rámcích, aplikovaného po výstupu z regresoru.

5.4 Závěry a komentáře k experimentům se samostatnými systémy

Výsledky všech zkoumaných samostatných systémů jsou zobrazeny v tabulce 5.2 spolu s výsledky zveřejněnými organizátory evaluace AV+EC 2015 — popis těchto systémů je zveřejněn v oficiálním článku k evaluaci, viz [14]. Je zřejmé, že ve většině případů produkují systémy popsané v této bakalářské práci výsledky lepší. A také potvrzují, že emoční dimenze arousal je lépe rozpoznatelná z audia, na rozdíl od emoční dimenze valence, které více svědčí video. Největší přínos pro tyto systémy spočíval v:

- Použití velkého časového kontextu, v některých systémech o velikosti až 6–7 s.
- Aplikaci mediánového filtru na výstupní hodnoty z regresoru.

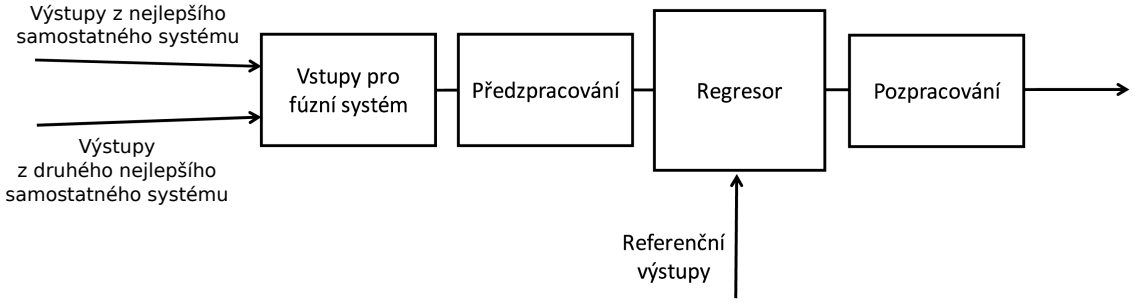
Tabulka 5.2: Srovnání samostatných systémů používající příznakové sady. Výsledky systémů prezentovaných organizátory AV+EC jsou v závorkách. *Vyhodnocení obsahovalo numerickou chybu a vzhledem k absenci referenční hodnot pro testovací sadu nebylo možné tento pokus opakovat.

	Vývojová sada		Testovací sada	
CCC	Arousal	Valence	Arousal	Valence
Audio	0.642 (0.287)	0.280 (0.069)	**0.595** (0.228)	0.160 (0.068)
Video geometrie obličeje	0.160 (0.231)	0.403 (0.325)	0.151 (0.162)	**0.302** (0.292)
Video vzhled	0.126 (0.103)	0.346 (0.273)	0.110 (0.114)	**0.334** (0.234)
ECG	0.305 (0.275)	0.231 (0.183)	–* (0.192)	–* (0.139)
EDA	0.117 (0.078)	0.235 (0.204)	0.118 (0.079)	0.226 (0.195)
Bottle–Neck	0.625	0.344	**0.525**	0.176

- Trénování systémů na konkrétních anotátorech a ne na Gold standardu, se kterým je pak predikce porovnávána.

5.5 Fúze

Pro fúzi byly vytvořeny dva systémy — odděleně pro arousal a pro valenci. Schéma fúzního systému můžete vidět na obrázku 5.10. Fúzní systém používá původní výstupy (nezprůměrované ani nevyhlazené mediánovým filterm) ze dvou nejlepších samostatných systémů. Tyto hodnoty jsou brány jako vstupní hodnoty do fúzního regresoru a je s nimi prováděná také jistá forma předzpracování a pozpracování. Nastavení parametrů systému použitých pro tvorbu konečného fúznho systému je vidět v tabulce 5.3. Výsledky z těchto fúzních systémů byly do evaluace AV+EC 2015 zaslány a tak bylo zjištěno, že úspěšnost těchto fúzních systémů na testovací sadě je CCC = 0.660 pro predikci arousalu a CCC = 0.504 pro predikci valence, což je relativně o 49% přesněší predikce arousalu a 31% přesnější predikce valence, než dávají fúzní baseline systémy, prezentované v článku [14]. Spolu s výsledky vyhodnocené na vývojové sadě jsou v tabulce 5.5.



Obrázek 5.10: Schéma fúzního systému.

Dle tabulky 5.2 byla fúze pro arousal tvořena výstupy ze samostatných systémů používajících audio baseline příznaky a Bottle–Neck příznaky. Pro trénování byla použita lineární regrese. Vstup do fúzního systému rozpoznávající valenci byl tvořen výstupy ze dvou systémů používjících dva druhy video příznaků. V tomto případě byla pro trénvání použita neuronová síť s jednou skrytou vrstou o topologii 486–243–1, kde o velikosti skryté vrstvy opět platí, že se rovná polovině vstupních dat a je určena experimentálně.

Tabulka 5.3: Nastavení parametrů pro předzpracování, trénování a pozpracování dat pro finální fúzní systém. *Použit lineární regresor.

	Arousal	Valence
Počet skrytých vrstev(velikost)	–*	1(243)
Velikost kontextové informace	121	101
Mediánový filtr	51	71

Tabulka 5.4: Zlepšení predikce díky použití fúzního systému.

Arousal — zlepšení o 22 %			Valence — zlepšení o 38 %		
Audio eGeMAPS	0.642	—> 0.772	Video geometrie	0.403	—> 0.518
Audio Bottle–Neck	0.625		Video vzhled	0.346	

Tabulka 5.5: Výsledky fúzního systému. V závorkách jsou výsledky fúzních systémů, prezentovaných organizátory evaluace AV+EC 2015.

CCC	Arousal	Valence
Vývojová sada	0.772 (0.476)	0.518 (0.461)
Testovací sada	0.660 (0.444)	0.504 (0.382)

5.6 Nerozpoznaný obličej ve video příznacích

Problémem objeveným při experimentování byly některé chybějící hodnoty u video příznaků. Tento jev nastával v situacích, kdy obličej byl z větší části mimo kameru, nebo také když snímaná osoba měla hlavu skloněnou — obličej tudíž nebyl rozpoznán. Vzhledem k tomu, že ve finální fúzi rozpoznávající valenci byly použity právě pouze ony zmíněné video příznaky, bylo zapotřebí tento problém řešit.

Úspěšnostní statistika Bottle–Neck příznaků při rozpoznávání valence je v pořadí hned za video příznaky, Bottle–Neck příznaky jsou tudíž vhodným kandidátem pro nahrazení rámců s nerozpoznaným obličejem. Samotná náhrada příznaků vstupujících do regresoru však ke zlepšení nevedla.

Následně byl zvolen přístup nahrazení až samotných výsledků v místech s nerozpoznaným obličejem. Byly vyzkoušeny dvě příznakové sady pro doplnění výsledků na místa s nulovými hodnotami — Bottle–Neck příznaky a EDA příznaky.

Jak už bylo zmíněno při popisu použitých dat, video příznaky byly dvojího charakteru — geometrie obličeje a vzhled. Prázdné hodnoty v jednotlivých rámcích se však lišily, proto bylo potřeba s nimi pracovat zvlášť. Následující postup byl tedy prováděn pro oba tyto typy příznaků:

Nejprve byly vytvořeny matice jedniček a nul, kde 1 znamenala vyplněný rámec a 0 chybějící hodnotu — označmě je *A1*. A k ní matice opačné, kde 1 znamenala nulovou hodnotu rámce a 0, že hodnota v daném rámci existuje — označme je *A0*. Dále byly použity dva výstupy z regresoru používající video příznaky a tyto výstupy byly vynásobeny maticí *A1*. Tím došlo k vynulování predikce v místech, kde byly hodnoty příznaků nulové. Poté byly výstupy z regresorů používající Bottle–Neck a EDA příznaky vynásobeny maticí *A0*. Tím zůstaly hodnoty predikce pouze pro ty rámce, které bylo potřeba nahradit. Nakonec byly tyto dvě výsledné matice sečteny a použity jako vstup do fúzního systému. Na celé vývojové sadě došlo ke zlepšení o 8 % relativně při nahrazení Bottle–Neck příznaky. Další statistické výsledky můžete vidět v tabulce 5.6.

Tabulka 5.6: Výsledky korekce predikce z nerozpoznaných rámců. Porovnání úspěšnosti na celé vývojové řadě a na nahrávce s více než 40 % nerozpoznanými rámci. *Relativně k původním hodnotám.

Valence	Původní video	Nahrazeno Bottle–Neck	Nahrazeno EDA
Celá vývojová sada	0.518	0.559 (zlepšení 8 %*)	0.539 (zlepšení 4 %*)
Nahrávka č.4	0.103	0.272 (zlepšení 164 %*)	0.109 (zlepšení 6 %*)

5.7 Bottle–Neck multi–task[1] systém

Výsledky systémů používající Bottle–Neck příznaky se ukázaly jako slibné. Vzhledem k tomu, že jsou extrahovány pouze z audia a tato modalita je objektem největšího zájmu, byl vytvořen další systém používající ony Bottle–Neck příznaky. Tento systém byl vyzkoušen pro trénování obou emočních dimenzí dohromady a výsledky v tabulce 5.7 ukazují, že tento přístup je pro Bottle–Neck systém výhodnější.

Tabulka 5.7: Porovnání multi–task a single–task přístupů k tvorbě systému používající Bottle–Neck příznaky. *Parametry byly nastaveny zvlášť pro každou emoční dimenzi. **Byly použity stejné hodnoty parametrů jako v multi–task systému.

	Vývojová sada		Testovací sada	
CCC	Arousal	Valence	Arousal	Valence
Single–task*	0.625	0.361	0.525	0.176
Single–task**	0.390	0.343	0.296	0.174
Multi–task	**0.699**	**0.386**	**0.596**	**0.293**

Tomuto systému po řadě prospělo:

1. Použití časového kontextu o celkové délce 181 rámců (asi 7 s), vedoucí ke zlepšení predikce arousalu o 240 % a valence o 510 % relativně (původní hodnoty CCCarousal = 0.159, CCCvalence = 0.051; hodnoty po aplikaci časového kontextu CCCarousal = 0.540, CCCvalence = 0.311).

2. Trénování na konkrétních anotátorech — u valence to byl první a druhý (zlepšení o 4 %), u arousalu první a třetí (zlepšení o 18 %).

3. Aplikace mediánového filtru skrze 183 rámců u arousalu a 145 rámců u valence. Odhad valence se tak zlepšil o 14 %, arousalu o 9 %.

4. Posunutí referenčních hodnot pro trénování systému tak, jak je ukázáno v grafu 5.7. U arousalu posunutí o 40 rámců vedlo ke zlepšení o 3 %, u valence posunutí o 50 rámců o 5 %.

5. Ponechání prvních 30–ti bází diskrétní kosinovy transformace vedoucí k zlepšení v průměru o 1 % a vedoucí ke konečné přesnosti predikce CCCarousal = 0.721, CCCvalence = 0.400.

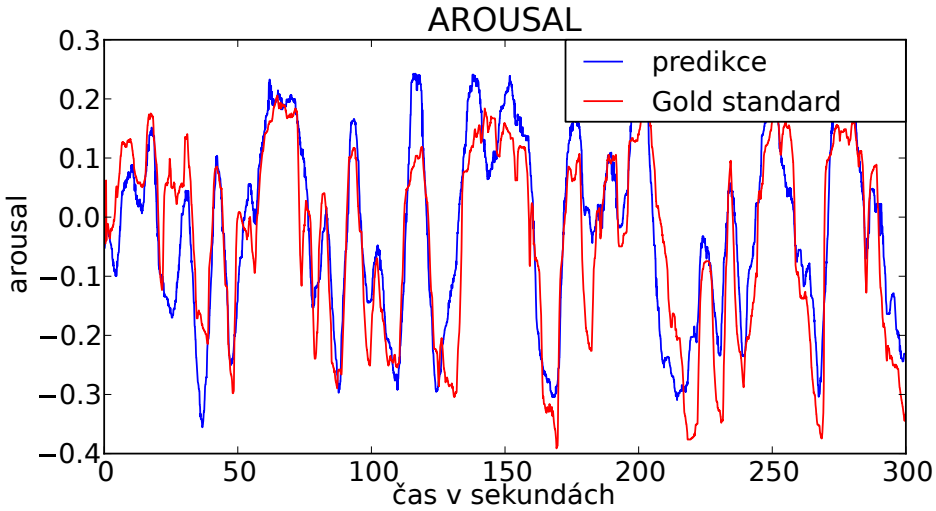
[1] multi–task je v tomto smyslu myšleno jako rozpoznávající obě emoční dimenze najednou, v jednom systému. Single–task je potom chápán jako systém rozpoznávající pouze jednu emoční dimenzi.

Kapitola 6

Závěr

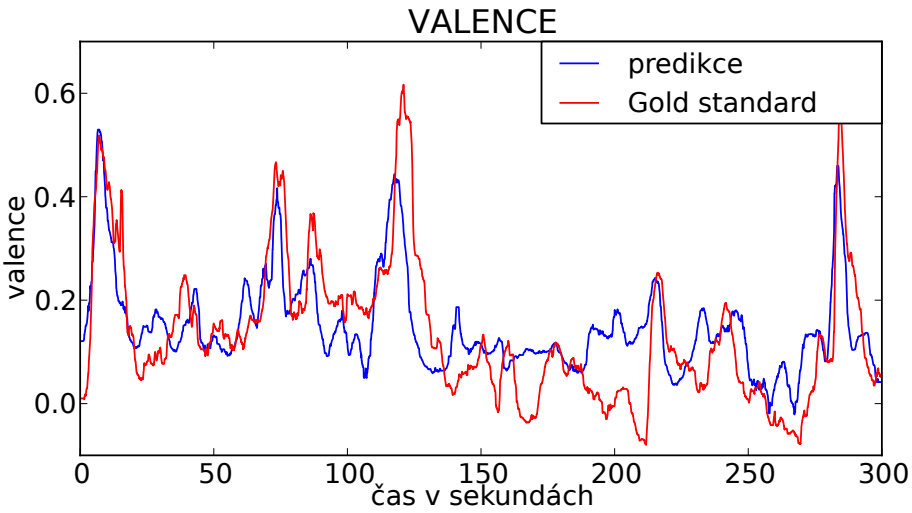
V této bakalářské práci je shrnut přístup k tvorbě systémů využívající různé modality (audio, video, fyziologické záznamy) k rozpoznávání emocí. Největší důraz je kladen na rozpoznávání z audia, kde byly nově vyzkoušeny Bottle–Neck příznaky, které jsou úspěšně používány i v dalších oblastech zpracovávání řeči.

Pro valenci i arousal je dosaženo nejlepších výsledků použitím fúze. Každá fúze je složena ze dvou podsystémů. První rozpoznávající arousal využívá audio baseline příznakový set (eGeMAPS) příznivě kombinované s Bottle–Neck příznakovým systémem. Druhá čerpá ze systémů používající dva druhy video příznaků — geommetrii obličeje a vzhled — a vyhodnocuje valenci. Skóre tohoto řešení dosahuje CCC = 0.64 mezi predikcí a Gold Standardem na vývojové sadě. Konkrétně CCCarousal = 0.772 a CCCvalence = 0.518. Ukázka predikce dvou různých nahrávek, dosahujících nejlepší predikce, je na obrázku 6.1 a 6.2. Pro porovnání je v příloze B připojena ukázka nejméně přesné predikce, opět pro dvě různé nahrávky.



Obrázek 6.1: Srovnání predikce emoční dimenze arousal finálního fúzního systému a Gold Standardu pro nahrávku z vývojové sady. CCC této nahrávky je rovno 0.851.

Vyhodnocení testovacích dat těchto fúzních systémů bylo zasláno do evaluace AV+EC 2015. Konečné skóre pro testovací sadu bylo pro arousal CCC = 0.660 a pro valenci CCC = 0.504. Tento výsledek přinesl páté místo, graf s ostatními účastníky je na ob-



Obrázek 6.2: Srovnání predikce emoční dimenze valence finálního fúzního systému a Gold Standardu pro nahrávku z vývojové sady. CCC této nahrávky je rovno 0.676.

rázku 6.3. Odkaz na článek výherců této evaluace viz [6].

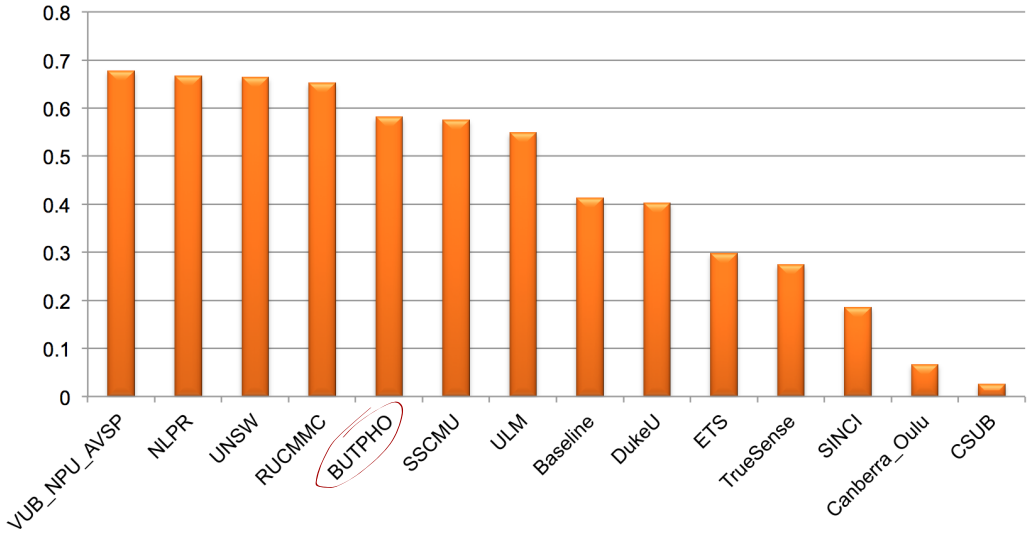
Výzkum prezentovaný v této práci potvrzuje důležitost správného výběru délky časového kontextu, anotací pro trénování a také délky mediánového filtru, aplikovaného na výstup z regresoru. V sekci 5 v tabulce 5.1 jsou uvedeny konkrétní hodnoty všech paremetrů ukazující závislost těchto parametrů jak na konkrétním příznakovém setu, popř. modalitě, tak na rozpoznávané emoční dimenzi.

Predikce valence je obecně známa jako hůře rozpoznatelná pouze z audia. V této práci byla tudíž soustředěna velká pozornost na vylepšení této skutečnosti experimentováním s Bottle–Neck příznaky. Součístí práce je i popis systému pracující pouze s Bottle–Neck příznaky a rozpoznávající obě emoční dimenze. Tento systém je potom pouze o 13 % relativně horší než výsledky produkující finální fúzní systémy. Tento multi–task systém rozpoznává valenci s CCC = 0.400 a arousal s až CCC = 0.721, čímž předčil výsledky systému používající audio eGeMAPS příznaky v průměru o 22 %.

6.1 Směry další práce

Letos bych se ráda znovu zapojila do evaluace AV+EC 2016, která se opět soustředí na rozpoznávání emocí založené na databázi RECOLA, a nově se navíc zabývá rozpoznáváním stupně deprese. Pro budoucí systém bych chtěla vyzkoušet další trénovací modely vhodné pro dataset tohoto typu, jako je Random Forest nebo Support Vector Regression (SVM).

Obecně pro budoucí práci má určitě smysl použití Bottle–Neck příznaků a jejich další výzkum v oblasti rozpoznávání emocí, popřípadě kombinace s příznaky dalšími. Dále by k zlepšení přesnosti predikce mohlo přispět rozšíření trénovacích dat — a to i skrze různé jazyky. Při větším množství dat pro trénování by totiž mohlo najít uplatnění použití neuronové sítě, která v sobě skrývá větší potenciál než obyčejná lineární regrese.



Obrázek 6.3: Výsledky evaluace AV+EC 2015. Systém popisovaný v této práci je označen jako BUTPHO.

Literatura

[1] Almaev, T. R.; Valstar, M. F.: Local Gabor Binary Patterns from Three Orthogonal Planes for Automatic Facial Expression Recognition. In *Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, ACII '13, Washington, DC, USA: IEEE Computer Society, 2013, ISBN 978-0-7695-5048-0, s. 356–361, doi:10.1109/ACII.2013.65.

[2] Douglas-Cowie, E.; Cowie, R.; Sneddon, I.; aj.: *Affective Computing and Intelligent Interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007 Proceedings*, kapitola The HUMAINE Database: Addressing the Collection and Annotation of Naturalistic and Induced Emotional Data. Berlin, Heidelberg: Springer Berlin Heidelberg, 2007, ISBN 978-3-540-74889-2, s. 488–500, doi:10.1007/978-3-540-74889-2_43.
URL http://dx.doi.org/10.1007/978-3-540-74889-2_43

[3] Ěerekoviæ, A.: An insight into multimodal databases for social signal processing: acquisition, efforts, and directions. *Artificial Intelligence Review*, ročník 42, č. 4, 2014: s. 663–692, ISSN 1573-7462, doi:10.1007/s10462-012-9334-2.
URL http://dx.doi.org/10.1007/s10462-012-9334-2

[4] Eyben, F.; Scherer, K.; Schuller, B.; aj.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Transactions on Affective Computing*, ročník PP, č. 99, 2015, ISSN 1949-3045, doi:10.1109/TAFFC.2015.2457417.

[5] Fér, R.; Matějka, P.; Grézl, F.; aj.: Multilingual Bottleneck Features for Language Recognition. In *Proceedings of Interspeech 2015*, 2015, s. 389–393.

[6] He, L.; Jiang, D.; Yang, L.; aj.: Multimodal Affective Dimension Prediction Using Deep Bidirectional Long Short-Term Memory Recurrent Neural Networks. In *Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge*, AVEC '15, New York, NY, USA: ACM, 2015, ISBN 978-1-4503-3743-4, s. 73–80.

[7] Iva Stuchlíková: *Základy psychologie emocí*. Portál, 2002, ISBN 978-80-7367-282-9.

[8] Jeong, D. H.; Ziemkiewicz, C.; Fisher, B.; aj.: iPCA: An Interactive System for PCA-based Visual Analytics. *Computer Graphics Forum*, ročník 28, č. 3, 2009: s. 767–774.

[9] Laukka, P.: *Vocal Expression of Emotion. Discrete ans Dimensional Accounts*. Dizertační práce, Uppsala University, Trädgårdsgatan 18, Uppsala, Sweden, 12 2004.

[10] Matějka, P.; Zhang, L.; Ng, T.; aj.: Neural Network Bottleneck Features for Language Identification. In *Proceedings of Odyssey 2014*, 2014, s. 299–304.

[11] McKeown, G.; Valstar, M.; Cowie, R.; aj.: The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations between a Person and a Limited Agent. *IEEE Transactions on Affective Computing*, ročník 3, č. 1, Jan 2012: s. 5–17, ISSN 1949-3045, doi:10.1109/T-AFFC.2011.20.

[12] Rainer Zelinski, P. N.: Adaptive Transform Coding of Speech Signals. *IEEE Transactions on Acoustic Speech and Signal Proccesing*, ročník 25, č. 4, 1977, doi:10.1109/TASSP.1977.1162974.

[13] Ringeval, F.; Eyben, F.; Kroupi, E.; aj.: Prediction of asynchronous dimensional emotion ratings from audiovisual and physiological data. *Pattern Recognition Letters*, ročník 66, 2015: s. 22 – 30, ISSN 0167-8655, doi:dx.doi.org/10.1016/j.patrec.2014.11.007, pattern Recognition in Human Computer Interaction.
URL http://www.sciencedirect.com/science/article/pii/S0167865514003572

[14] Ringeval, F.; Schuller, B.; Valstar, M.; aj.: AV+EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In *Proc. AVEC 2015, satellite workshop of ACM-Multimedia 2015*, Brisbane, Australia, Říjen 2015.

[15] Schlosberg, H.: Three dimensions of emotion. *Psychological Review*, ročník 61, č. 2, 1954: s. 81–88.

[16] Sonderegger, F. R. A.; Sauer, J.; Lalanne, D.: Introducing the RECOLA Multimodal Corpus of Remote Collaborative and Affective Interactions. In *Proc. Face and Gestures 2013, Workshop on Emotion Representation, Analysis and Synthesis in Continuous Time and Space (EmoSPACE)*, 2013.

[17] Sspnet.eu: AV+EC 2015 « SSPNET. Dostupné z: http://sspnet.eu/avec2015/, 2016 [cit. 2016-04-29].

[18] Ververidis, D.; Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods. *Speech Communication*, ročník 48, č. 9, 2006: s. 1162 – 1181, ISSN 0167-6393, doi:dx.doi.org/10.1016/j.specom.2006.04.003.
URL http://www.sciencedirect.com/science/article/pii/S0167639306000422

[19] Wikipedia, t. f. e.: Discrete cosine transform [online]. Dostupné z: https://en.wikipedia.org/wiki/Discrete_cosine_transform/, 2016-04-29 [cit. 2016-05-02].

[20] Wilhelm Max Wundt: *Outlines of psychology*. Leipzig, W. Engelmann; New York, G.E. Stechert, 1897.

Přílohy

Seznam příloh

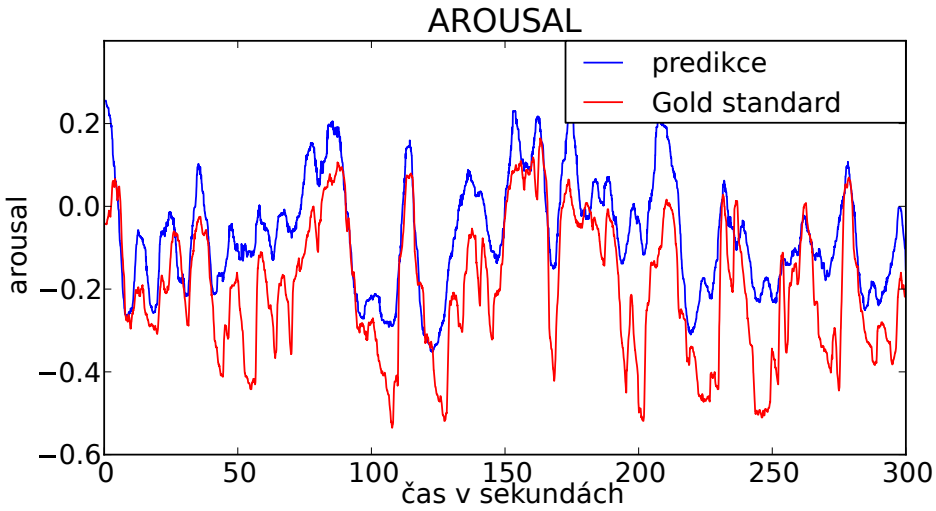
Příloha A

Obsah CD

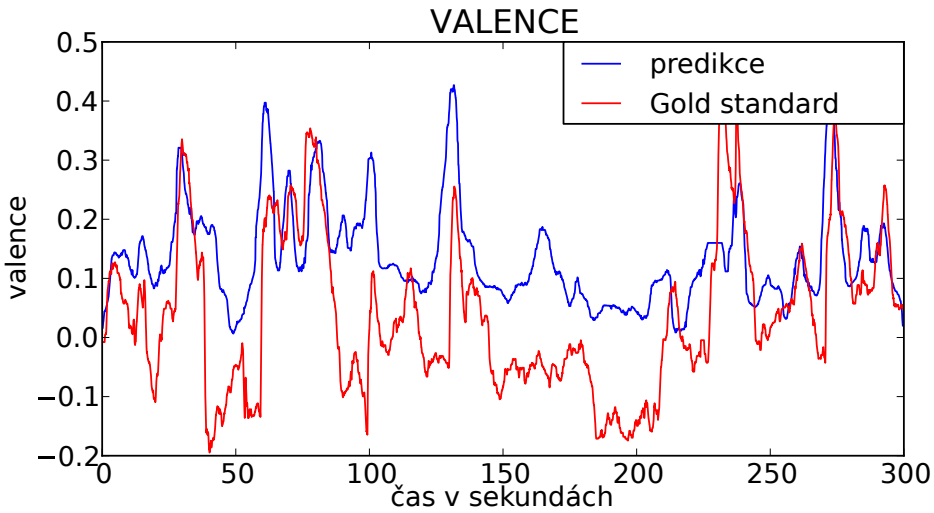
1. Skripty s regresory — spouštějí se s parametrem `train` (natrénuje systém) nebo `test` (vyhodnotí systém).
   - `r`egression.py — multi-task systém
   - `r`egression_single.py — single-task systém
   - `r`egression_arousal_fusion_audio+bnout.py — fúzní systém pro arousal
   - `r`egression_valence_fusion_video+geo+appear_filled.py — fúzní systém pro valenci
2. `f`eatures+labels/ — Složka s přichystanými daty, obsahuje:
   - `a`udio/, ecg/, eda/, video/ — normalizované příznaky roztříděné do složek dle modalit
   - `l`abels/arousal/, labels/valence/ — referenční výsledky roztříděné dle emoční dimenze (anotace + Gold standard)
   - `f`usion_inputs/ — výstupy z nejlepších samostatných systémů pro vstup do systémů fúzních
3. `s`rc/ — Obsahuje podpůrné skripty, během celého výzkumu jich bylo používáno více, ale z důvodu paměťové náročnosti jsou zde ponechány jen ty základní, umožňující spuštění trénování a testování již přichystaných dat.
4. `r`esults_BUTPHO_fusion/ — Obsahuje grafy predikce jednotlivých nahrávek z trénovací a vývojové sady společně s Gold standardem, rozdělené do složek podle emoční dimenze. Každá složka navíc obsahuje konkrétní hodnoty úspěšnosti jednotlivých nahrávek v evaluační metrice CCC (viz kapitola 3.4).
5. `l`ists/ — Obsahuje seznamy dat, které se mají použít pro trénování, validaci a testování.

Příloha B

Ukázka nejméně přesné predikce dvou různých nahrávek



Obrázek B.1: Srovnání predikce emoční dimenze arousal, produkované finálním fúzním systémem, a Gold Standardu pro nahrávku s nejmenší úspěšností, CCC = 0.582.



Obrázek B.2: Srovnání predikce emoční dimenze valence, produkované finálním fúzním systémem, a Gold Standardu pro nahrávku s nejmenší úspěšností, CCC = 0.355.