Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra kybernetiky

Disertační práce
k získání akademického titulu doktor
v oboru *Kybernetika*

Ing. Martin Grůber

Syntéza expresivní řeči s využitím dialogových aktů k popisu expresivity

Školitel: doc. Ing. Jindřich Matoušek, Ph.D.

Plzeň 2012

University of West Bohemia
Faculty of Applied Sciences
Department of Cybernetics

Dissertation
submitted in conformity with the requirements
for the degree of Doctor of Philosophy
in the field of *Cybernetics*

Ing. Martin Grůber

Dialogue-Act Based Expressive Speech Synthesis

Supervisor: doc. Ing. Jindřich Matoušek, Ph.D.

Plzeň 2012

Prohlášení

Předkládám tímto k posouzení a obhajobě disertační práci zpracovanou na závěr doktorského studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že tuto práci jsem vypracoval samostatně s použitím odborné literatury a dostupných pramenů uvedených v seznamu, jenž je součástí této práce.

V Plzni dne Martin Grůber

Poděkování

Tato disertační práce vznikla za podpory mezinárodního projektu Companions (`http://www.companions-project.org`) sponzorovaného Evropskou komisí v programu IST-FP6-034434 a dále také projektu GAČR 102/09/0989.

Dále bych chtěl poděkovat:

- svému školiteli doc. Ing. Jindřichu Matouškovi, Ph.D., za poskytnutí cenných odborných rad a konzultací,
- kolegům Ing. Milanu Legátovi a Ing. Danielu Tihelkovi, Ph.D., za spolupráci při řešení některých problémů,
- všem ostatním kolegům oddělení umělé inteligence Katedry kybernetiky za vytvoření dobrých pracovních podmínek,
- rodičům za jejich všestrannou podporu, kterou mi v průběhu celého studia věnovali,
- přítelkyni Oldřišce za pochopení a trpělivost.

Obsah

Seznam obrázků

Seznam tabulek

Seznam použitých symbolů a zápisů

a	skalární proměnná
$\mathbf{a}$	vektor
$\mathbf{A}$	matice
$\mathbf{A}^T$	transponovaná matice
$\hat{a}$	odhad parametru a
$APOLOGY$	označení dialogového aktu
$F0$	základní hlasivková frekvence
$F1$, $F2$, $F3$	formantové frekvence
RMS	hodnota RMS signálu

Kapitola 1

Úvod

Komunikace mluvenou řečí je základní, nejpřirozenější a nejdůležitější forma přenosu informace mezi lidmi. Je proto samozřejmé, že s rostoucím využitím výpočetní techniky roste i snaha o uplatnění této formy komunikace při dialogu člověka s počítačem. Než však bude počítač schopen s lidmi komunikovat na stejné úrovni jako lidé mezi sebou, je potřeba vyřešit několik problémů s tím spojených. Jedním z nich je také syntéza řeči, což je proces, při kterém dochází k umělému vytváření řečového signálu.

Historie syntézy řeči začíná již koncem 18. století, kdy vznikala jednoduchá zařízení napodobující hlasový trakt člověka. Jako první syntetizér je označován stroj, který v roce 1791 sestrojil Wolfgang von Kempelen. Jeho zařízení, sestávající se z měchů, rákosu, kožené trubice nebo pryže, bylo schopno vytvářet primitivní slova i věty. Další badatelé pak v 19. století objevovali různé nástupce tohoto Kempelenova mechanického přístroje. Jejich cílem bylo samozřejmě vylepšení kvality produkovaných zvuků.

S rozvojem jednoduchých elektrických obvodů počátkem 20. století se začínají objevovat první elektronické syntetizéry. První takový přístroj, který dokázal vytvářet souvislou řeč, vytvořil v roce 1936 Homer Dudley. Jeho *Voder* (Voice Operating Demonstrator) se skládal z paralelně zapojených pásmových propustí, které byly buzeny buď periodickým signálem, nebo šumem. Výstupy jednotlivých propustí se nezávisle na sobě zesilovaly pomocí potenciometrů a sčítaly. Stroj byl ovládán pomocí několika tlačítek a nožního pedálu a jeho obsluhu zvládal pouze zkušený a trénovaný operatér, ale řeč vytvářená tímto zařízením již byla srozumitelná.

V 50. letech 20. století se s nástupem číslicových počítačů mění přístup k syntéze řeči. Místo spíše mechanických zařízení už vznikají digitální syntetizéry využívající výhod, které poskytují právě číslicové systémy. Techniky syntézy řeči se také začínají dělit na dvě různé metody: modelový přístup (modeluje celý proces vytváření řeči člověkem) a signálový přístup (mode-

luje pouze výsledný akustický signál). A tak se v následujících desetiletích objevují formantové a artikulační syntetizéry, a v 70. letech i první konkatenační syntetizéry.

Syntéza české řeči se objevuje v 60. letech, kdy je vytvořen první český syntetizér. Koncem 20. století vznikají převážně konkatenační syntetizéry a na katedře kybernetiky Západočeské univerzity v Plzni je vytvořen první český konkatenační syntetizér využívající rozsáhlý řečový korpus.

V polovině 20. století se také začíná rozvíjet problematika syntézy řeči z libovolného textu (*Text To Speech, TTS*), která se v české literatuře označuje jako konverze textu na řeč, což je nejobecnější úloha syntézy řeči. Než dojde k samotnému vytvoření řečového signálu, je totiž potřeba provést předzpracování a analýzu textu. Zpracování přirozeného jazyka (*Natural Language Processing, NLP*) je prvním modulem, kterým musí text projít. V rámci tohoto modulu se provádí předzpracování textu (detekce slov, detekce konců vět, normalizace textu), dále pak morfologická analýza (navrhování možných mluvnických kategorií jednotlivých slov), kontextová analýza (redukování možných mluvnických kategorií z morfologické analýzy na základě kontextu okolních slov), syntakticko-prozodický rozbor (rozdělení vět na větné úseky), fonetická transkripce (převod ortografické podoby do podoby fonetické) a generování prozodie (prozodické charakteristiky popisují intonaci, rychlost, hlasitost, rytmus a členění řeči). Dalším modul pro digitální zpracování signálu (*Digital Signal Processing, DSP*) je potom určen k samotné syntéze řeči.

V současné době je syntéza řeči na velmi kvalitní úrovni, je srozumitelná a pro posluchače příjemná. Nicméně v ní stále chybí prvek přirozenosti, který může být dosažen pouze použitím metod, jež syntetické řeči dodají „lidštější“ podobu. Tyto metody se snaží přenést pomocí syntetické řeči na posluchače i postoj a náladu mluvčího, tj. toho, koho má syntetizér představovat. Zde je samozřejmě možné přemýšlet nad možnou aplikací takových systémů syntézy řeči. Jejich využití je obecně široké, nicméně uveďme zde alespoň základní příklady.

Na jednu stranu může být systém syntézy řeči použit jako výstup nebo jeden z mnoha výstupů nějakého stroje nebo přístroje, např. počítače, mobilního telefonu nebo GPS navigace, kde uživatele informuje např. o svém stavu nebo ho navádí do cíle jeho cesty. V tomto případě stojí jistě za úvahu zamyslet se nad tím, zda takový stroj vůbec může (popř. má zapotřebí) navenek prezentovat nějaké své postoje, nálady nebo emoce. Je spíše pravděpodobnější, že bude používat zdůraznění některých zpráv, jejich částí nebo tak podobně.

Na straně druhé můžeme uvažovat o případu, kdy syntetizér nahrazuje mluvenou komunikaci mezi lidmi, tedy např. v dialogových systémech, při čtení emailů a SMS zpráv nebo kompenzuje handicap, který člověku jakým-

koliv způsobem znemožňuje produkovat řeč vlastními silami (např. kvůli nějakému úrazu či nemoci) a nahrazuje tak tedy jeho vlastní hlas a usnadňuje mu tak sociální začlenění do společnosti. V tomto případě se může jednat skutečně o původní hlas takto postiženého člověka, jenž byl v minulosti nějakým způsobem zakonzervován, nebo jde o hlas patřící původně někomu jinému, popř. je vytvořený zcela uměle. Zde je již jistě na místě uvažovat o tom, že syntetizér by měl nejen úspěšně produkovat srozumitelnou řeč, ale také jejím prostřednictvím prezentovat pocity a nálady toho, kdo např. psal daný email či SMS zprávu, nebo využívá syntetizér přímo pro komunikaci při rozhovorech s ostatními lidmi.

Pravděpodobně teprve až potom, co bude vyřešena úloha začlenění postoje mluvčího do syntetické řeči, bude dialog člověka a počítače moci probíhat přirozeně, včetně expresivního vyjádření obou účastníků. Toto je samozřejmě myšleno pouze z hlediska syntézy řeči jako takové. K přirozenému dialogu mezi člověkem a počítačem (strojem) je totiž také třeba zdárně vyřešit problém jak rozpoznávání řeči (včetně rozpoznání postoje mluvčího), tak i řízení dialogu. To je mechanismus, který počítači (obecně jakémukoliv systému) nějakým způsobem interpretuje co člověk říká a jakým způsobem je potřeba v dané fázi dialogu reagovat či odpovědět.

Je samozřejmě jasné, že vytvoření vysoce kvalitní, srozumitelné a hlavně přirozené řeči, která by nebyla odlišitelná od promluvy člověka, je velmi náročný úkol, který zahrnuje problematiku nejen z oblasti umělé inteligence, ale také z oblastí akustiky, fonetiky, fonologie, lingvistiky, psychologie a dalších jim podobných.

V této práci se v kapitolách 2 – 4 věnujeme popisu současného stavu v oblasti počítačové syntézy (expresivní) řeči a používaných přístupů pro popis expresivity. V kapitole 5 uvedeme cíle této práce, kterých se snažíme dosáhnout. Postup vývoje systému pro syntézu expresivní řeči v dialogu pak popisují kapitoly 6 – 8 a v kapitole 9 prezentujeme dosažené výsledky. Závěrečné shrnutí je uvedeno v kapitole 10.

Kapitola 2

Syntéza řeči obecně

V úloze syntézy řeči existuje v současnosti několik základních přístupů. Jsou jimi: formantová syntéza, artikulační syntéza, HMM syntéza a konkatenační syntéza (pod kterou se řadí známá difonová syntéza – obecně metoda s jedním reprezentantem a dále také metoda dynamického výběru jednotek). Zajímavých výsledků bylo v minulosti dosaženo použitím formantové syntézy, nicméně její vývoj není v současnosti příliš rozšířen a kvalita takto produkované řeči je většinou překonána novějšími přístupy. Nejvíce používanou metodou je nyní především konkatenační syntéza a hlavně metoda výběru jednotek, proto bude tato práce věnována v první řadě jí. Ovšem do popředí zájmu se dostává také HMM syntéza a to hlavně pro její flexibilitu, kapacitní nenáročnost a možností rychlé tvorby nových hlasů. Artikulační syntéza se zatím jeví jako příliš komplikovaná a pro její realizaci je i nedostatek reálných dat. Předpokládá se, že tento přístup má velký potenciál a mohl by být využit spíše v budoucnosti.

V oblasti syntézy řeči se v minulosti zatím řešila především srozumitelnost a kvalita syntetizované řeči. Nekladl se tak velký důraz na přirozenost, proto byla výstupem především neutrální řeč. V posledních letech dochází v tomto ohledu ke změně. Srozumitelnost a kvalita jsou již na celkem vysoké úrovni, začíná se tedy řešit i otázka přirozenosti řeči. Snahou je tedy vytvořit syntetickou řeč, která bude také přirozená, pokud možno nerozeznatelná od té, kterou produkuje člověk. V souvislosti s „odbouráváním“ neutrality řeči se v této oblasti objevuje několik pojmů, které se budou vyskytovat i v následujících kapitolách. Tyto pojmy je nutné upřesnit a vysvětlit. V dostupné literatuře se totiž vyskytuje několik různých označení: expresivní syntéza, afektivní syntéza, syntéza emotivní řeči, syntéza určitého stylu, apod. Všechny tyto pojmy většinou označují velmi podobné postupy, ale jisté rozdíly můžeme pozorovat.

Výrazy *syntéza určitého stylu* a *syntéza emotivní řeči* shrneme pod pojem

emotivní syntéza. Tímto označením bude myšlena syntéza s různým emotivním zabarvením, tedy např. smutně, vesele, znuděně, unaveně, atd., jak je dále uvedeno v části 3.1.

Výrazy *expresivní* a *afektivní syntéza* označíme pouze jako *expresivní syntéza*. Tento pojem bude v následujícím textu vždy znamenat syntézu řeči, která se jakkoliv odlišuje od řeči neutrální a vyjadřuje jakýkoliv postoj, náladu nebo tzv. *afektivní stav* řečníka. Může to tedy být syntéza obsahující neřečové události (zakašlání, povzdech, hlasitý nádech, úsměv, smích, souhlasné přitakání, další různé citoslovce, apod.), syntéza se zdůrazněním (určité slovo v syntetizované promluvě nebo její část je nějakým způsobem zdůrazněna), syntéza textu, který má v posluchači vyvolat určitý pocit, a z předchozího je patrné, že i emotivní syntéza bude patřit do množiny expresivní syntézy.

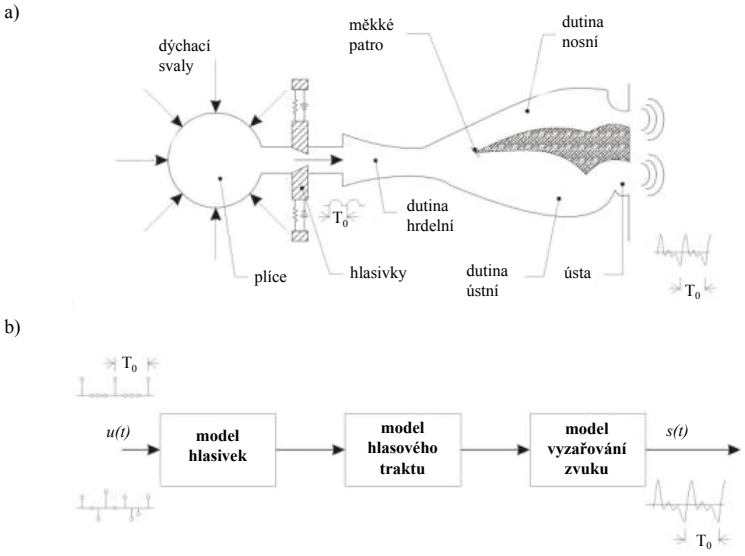
Na následujících stránkách jsou ve stručnosti uvedeny hlavní principy všech výše uvedených metod syntézy neutrální řeči a v kapitole 4 pak jejich možné použití při emotivní, resp. expresivní syntéze. Nejvíce prostoru je pak věnováno konkatenační syntéze a podrobnějšímu postupu, jakým se syntetická řeč touto metodou produkuje. V části 4.1.2 je pak tato metoda detailněji popsána s ohledem na využití pro syntézu expresivní řeči, v kapitole 6 s ohledem spíše na konkrétní použití v dialogovém systému.

2.1 Formantová syntéza

Formantová syntéza, také známá jako syntéza podle pravidel, vytváří řečový signál pouze na základě znalostí o jednotlivých řečových zvucích. Tyto znalosti jsou pak formulovány ve formě pravidel. Jak již z názvu vyplývá, půjde zejména o pravidla pro určování formantových frekvencí. V základním přístupu nejsou během procesu syntézy využity žádné části přirozené řeči.

Formantové syntetizéry se nesnaží modelovat do detailů lidské hlasové ústrojí, ale vytvářejí výsledný signál pomocí výše zmíněných pravidel. Ta jsou založena na teorii produkce lidské řeči, viz obrázek 2.1. Schéma takového formantového syntetizéru je pak znázorněno na obrázku 2.2. Chování zdroje popisuje ve frekvenční oblasti jeho spektrum $H_g(z)$. Zdroj budí hlasový trakt s přenosovou funkcí $H(z)$, zjednodušeně popsatelný jako skupina rezonátorů. Výslednou podobu pak zvuku dá model vyzařování řečového signálu $H_L(z)$.

Nejjednodušším použitelným modelem buzení pro znělou řeč je posloupnost impulsů s časovým odstupem odpovídajícím jedné periodě základního hlasivkového tónu (tzv. *pitch-periodě*) a simulující tak funkci hlasivek. Použití takového zdroje dává srozumitelnou řeč, ale výsledek zní „strojově“. Aby se buzení více blížilo chování skutečných hlasivek, je třeba impuls vhodně natvarovat. To může být provedeno použitím jednoduchého lineárního fil-

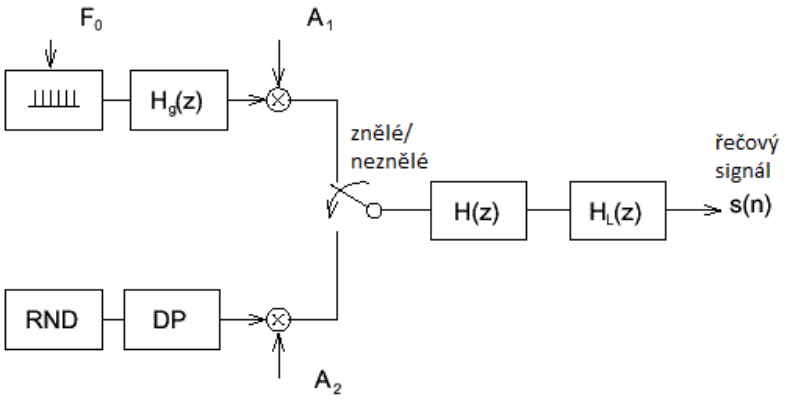


Obrázek 2.1: Model produkce řeči

tru. Pro generování neznělých úseků řeči je potřeba vytvořit model buzení, který odpovídá otevřeným hlasivkám. Tento stav lze simulovat generátorem bílého šumu, který je na obrázku 2.2 znázorněn jako blok RND, procházející filtrem DP. Nicméně toto nahrazení není úplně přesné, neboť tento model předpokládá, že náhodný šum vychází již z hrtanu. Šum se však ve skutečnosti vytváří až v hlasovém traktu v místě největšího zúžení, kde dochází ke vzduchovým turbulencím.

Základním prvkem modelu hlasového traktu u formantového syntetizéru jsou tedy rezonátory. Ty slouží k modelování zvýrazněných (rezonančních) frekvencí – vrcholů ve spektrální obálce – formantů. Rezonátory lze simulovat jako zpětnovazební filtry druhého řádu typu pásmová propust. Pro modelování utlumených frekvencí – propadů ve spektrální obálce – antiformantů, vznikajících vlivem vedlejších rezonátorů v hlasovém traktu, se používá antirezonátorů. Ty především simulují nazály. Dále také upravují spektrum budícího zdroje, přesněji řečeno modelují propojení zdroje buzení a hlasového traktu. I ty lze simulovat filtrem, tentokrát typu pásmová propust, který je bez zpětných vazeb. Různé nastavení filtrů zastupujících rezonátory a antirezonátory vlastně odpovídá různému nastavení artikulačních orgánů (čelist, jazyk, měkké patro, atd.) při vytváření řeči člověkem.

Po průchodu budícího signálu modelem hlasového traktu ještě zbývá simulovat chování signálu při vyzařování do prostoru, což odpovídá šíření zvu-



Obrázek 2.2: Formantová syntéza.

kového signálu rty do okolí. S dostatečnou přesností lze i tady vystačit s jednoduchým filtrem typu horní propust. Podrobnější informace o formantové syntéze lze nalézt např. v [105].

Výsledná syntetická řeč však zní i přes všechny snahy nepřirozeně a roboticky v porovnání se současnou konkatenační syntézou. Na druhou stranu, tento přístup poskytuje velkou volnost a flexibilitu v nastavení různých parametrů popisujících proces produkce řeči, což by mohlo být zajímavé pro modelování expresivity a emocí v řeči, viz [13]. Dále se také jako zajímavá jeví možnost kombinace základní formantové syntézy s využitím nahraného řečového korpusu pro extrakci parametrů [20]. Tento postup by se dal u syntézy expresivní řeči taktéž využít.

2.2 Artikulační syntéza

Artikulační syntéza se snaží komplexně modelovat lidský hlasový trakt. Matematicky popisuje celé hlasové ústrojí se všemi orgány, které se na produkci řeči podílejí (hrtan, hlasivky, měkké patro, jazyk, rty, čelisti, atd.). Modelování je možno provádět nejen matematicky, ale lze vytvářet i 3D obrazy hlasového traktu a tím umožnit i audio-vizuální syntézu [32]. Nastavitelné parametry, které ovlivňují artikulační syntézu, je potřeba určit co nejpřesněji, aby výsledná řeč zněla co nejpřirozeněji. Tyto parametry se dají získat z přirozeného řečového signálu [5], nebo pomocí magnetické rezonance (MRI), viz např. [6, 78, 87].

Tato metoda se zdá být pro syntézu expresivní řeči v současné době nepoužitelná, protože ještě není zdaleka vyřešena ani na úrovni neutrální řeči, ač její historie sahá do dnes již poměrně dávné minulosti (např. [102]). Zís-

kávání potřebných parametrů a samotné modelování je stále příliš složité. Nicméně právě možnost ovlivňovat výslednou řeč pomocí parametrů hlasového traktu je velmi důležitá z hlediska syntézy expresivní řeči. Teoreticky by tedy tento přístup mohl být v budoucnosti využíván, avšak v současné době je to prakticky nemožné.

2.3 Konkatenační syntéza

Základním principem konkatenační syntézy je použití částí přirozeného řečového signálu, který je ve formě promluv uložen v tzv. *řečovém korpusu*. Předpokládá se, že řeč se skládá z *řečových (akustických) jednotek*. Reálný řečový signál v řečovém korpusu je tedy pomocí metod automatické [72, 73] nebo ruční segmentace rozdělen na jednotlivé *segmenty*, které odpovídají těmto řečovým jednotkám.

V současné době se využívá především metod automatické segmentace, neboť používané řečové korpusy jsou většinou značně rozsáhlé a jejich ruční segmentace by tak byla značně pracná a také samozřejmě časově náročná. Kromě časové náročnosti hovoří proti ruční segmentaci také fakt, že ji musí provádět trénovaní experti, kteří se ovšem mohou díky velkému množství dat dopustit chyb v segmentaci (navíc se někdy ani dva experti neshodnou na správném určení časových hranic jednotlivých jednotek). Je pravdou, že chyby se objevují také při segmentaci automatické, zde však dochází k výskytu spíše systematických chyb, zatímco u ruční segmentace se mohou objevit chyby především náhodné. Automaticky segmentovaný řečový korpus by tedy měl být v tomto ohledu konzistentní.

Zmíněné segmenty z řečového korpusu jsou uloženy v tzv. *inventáři řečových jednotek*, což je seznam všech jednotek, které lze použít pro potřeby umělého vytváření řeči. Syntetizovaná řeč se pak vytváří řetězením (konkatenací) vhodných řečových segmentů z inventáře řečových jednotek. Z výše uvedeného je zřejmé, že výsledná syntetická řeč napodobuje hlas řečníka, který namluvil promluvy řečového korpusu. Jak je patrné z předchozího, jeden řečový korpus může namluvit pouze jeden jediný řečník, jinak by poté při vlastní syntéze docházelo k řetězení částí přirozené řeči od několika různých řečníků, což je samozřejmě nepřijatelné.

Jak již bylo uvedeno, základním stavebním kamenem mluvené řeči je tzv. řečová jednotka, což je abstraktní termín pro pojmenování stejného typu řečových zvuků. Konkrétní akustická realizace konkrétní řečové jednotky se pak označuje jako *zástupce (reprezentant) řečové jednotky*. V úloze syntézy řeči je možné volit délku těchto jednotek různě, s ohledem na maximální pokrytí koartikulačních jevů a bezproblémové řetězení [45].

2.3.1 Inventář řečových jednotek

Při vytváření inventáře řečových jednotek z nasegmentovaného řečového korpusu musíme určit, které realizace jednotek (zástupce) do inventáře uložit. Těchto zástupců pak totiž bude použito při syntéze a na jejich vlastnostech bude záviset kvalita výsledné syntetizované řeči.

Pro výběr jednotek máme v zásadě dvě možnosti:

- **Systémy s jedním zástupcem** – Do inventáře je uložen pouze jeden zástupce každé řečové jednotky, podle nějakého kritéria ten nejlepší (výběr zástupce tak probíhá *off-line*). Nejlepším v tomto případě většinou rozumíme „nejprůměrnější“, protože při syntéze potom dochází použitím různých metod k prozodické modifikaci jednotky (viz dále část 2.3.2). Pokud by byl v inventáři uložen nevhodný zástupce, bylo by potřeba při syntéze provádět větší množství modifikací, a tím by se pochopitelně také zhoršovala kvalita dané jednotky a spolu s tím i kvalita výsledné řeči. Je tedy zřejmé, že návrh kritéria výběru nejlepšího kandidáta je velmi důležitý. Protože je v inventáři uložena pouze jedna realizace každé jednotky, nejsou nároky na prostor, ve kterém je inventář uložen, příliš vysoké. Proto byl největší rozmach tohoto přístupu v devadesátých letech 20. století, kdy nebyla výpočetní technika (a zvláště pak paměťová kapacita počítačů) ještě na takové úrovni jako dnes.

- **Systémy používající výběr jednotek** – Do inventáře jsou uloženy buď všichni zástupci každé jednotky, kteří se vyskytly v řečovém korpusu, nebo několik vybraných odlišných (a to jak prozodicky, tak i spektrálně) zástupců. Během syntézy se pak použije ten reprezentant, který je v dané situaci nejvhodnější. K výběru zástupce tak dochází *on-line*. Výhoda tohoto přístupu označovaného jako *dynamický výběr jednotek* (nebo-li *unit selection* [56]) je v tom, že pokud vybereme vhodnou jednotku z inventáře, odpadá nutnost provádět prozodické modifikace (či je lze provádět pouze v minimální míře), a tudíž nedojde ke zkreslení původního přirozeného řečového signálu jednotky. Kvalita syntetizované řeči je tak řádově vyšší v porovnání s předchozím přístupem. Je patrné, že tato metoda je mnohem náročnější na kapacitu úložného prostoru, což však v dnešních podmínkách již neznamená problém. Dále je velmi obtížné volit kritérium on-line výběru jednotek (viz část 2.3.3), protože vlastní výběr je výpočetně velmi náročný (více než 50 % času syntézy). V současné době však tato metoda určuje trend v přístupu k syntéze řeči řetězením.

2.3.2 Metoda s jedním zástupcem

Metoda s jedním zástupcem (s jednou realizací jednotky, *single instance*) využívá přirozený řečový signál, který je v době syntézy modifikován. V inventáři řečových jednotek je pro každou jednotku uložen pouze jeden zástupce, který byl vybraný z řečového korpusu, a je podle nějakého kritéria ten nejlepší. Pro požadovanou promluvu jsou při syntéze vygenerovány prozodické parametry, především kontura $F0$ a doby trvání jednotlivých řečových jednotek. Tyto parametry mohou být určeny např. na základě sady pravidel nebo mohou být získány z reálných řečových dat. Akustický signál jednotky uložené v databázi je v průběhu syntézy modifikován tak, aby odpovídal požadovaným parametrům. Je zřejmé, že tato modifikace má velký vliv na kvalitu příslušné jednotky, neboť každý zásah do přirozeného signálu kvalitu řeči snižuje. K modifikacím se používají různé metody, např. PSOLA (TD-PSOLA, FD-PSOLA, LP-PSOLA), HNM, RELP (podrobnosti lze nalézt např. v [71, 109]).

Pro syntézu expresivní řeči je výhoda tohoto přístupu jasná. Pokud můžeme měnit parametry, které mají vliv na expresivní zabarvení řeči (viz část 7.3), můžeme tak syntetizovat nejen řeč přirozenou, ale i expresivní. Přichází tak zřejmě v úvahu získání potřebných parametrů (hlavně tedy doby trvání a kontury $F0$) z reálného řečového korpusu, ve kterém se vyskytují expresivní věty, a použití takto získaných parametrů při vlastní syntéze expresivní řeči. Některé experimenty používající tuto metodu jsou popsány např. v [11] nebo [91]. Nevýhodou pak jsou větší modifikace přirozeného řečového signálu spojené s velkou variabilitou parametrů v expresivní řeči.

2.3.3 Metoda dynamického výběru jednotek

Metoda výběru jednotek (*unit selection* [56]), narozdíl od metody s jednou realizací, využívá obecně všechny realizace jednotlivých jednotek z řečového korpusu a snaží se v reálném čase běhu syntézy vybrat nejlepší možnou posloupnost po sobě jdoucích segmentů. Kritériem pro výběr je snaha minimalizovat jednak modifikace řečového signálu a jednak počet řetězení (snaha vybírat ty kandidáty, kteří se v původním namluveném korpusu nacházeli vedle sebe).

Pro určení vhodnosti jednotlivých kandidátů se používají dvě hodnotící funkce, a to *cena cíle* (*target cost*) a *cena řetězení* (*concatenation cost* nebo *join cost*).

Cena cíle

Cena cíle C^t hodnotí odlišnost prozodických charakteristik realizace řečové jednotky nalezené v inventáři od charakteristik požadovaných. Na obrázku 2.3 jsou zobrazeny některé prozodické charakteristiky, které tvoří vektor příznaků. Jsou zde vidět příznaky jak pro požadovanou jednotku, tak pro jednotky v řečovém inventáři (kandidáty). V tomto případě se jedná o příznaky symbolické, neboť v našem systému syntézy řeči ARTIC [74] se tento přístup osvědčil. Nicméně lze použít i jiný popis, např. přímo modelovat konturu F0, tedy každé cílové jednotce přiřadit její požadovanou hodnotu, buď konkrétním číslem nebo třeba slovním popisem (vysoká, nízká, klesající, stoupající, apod.).

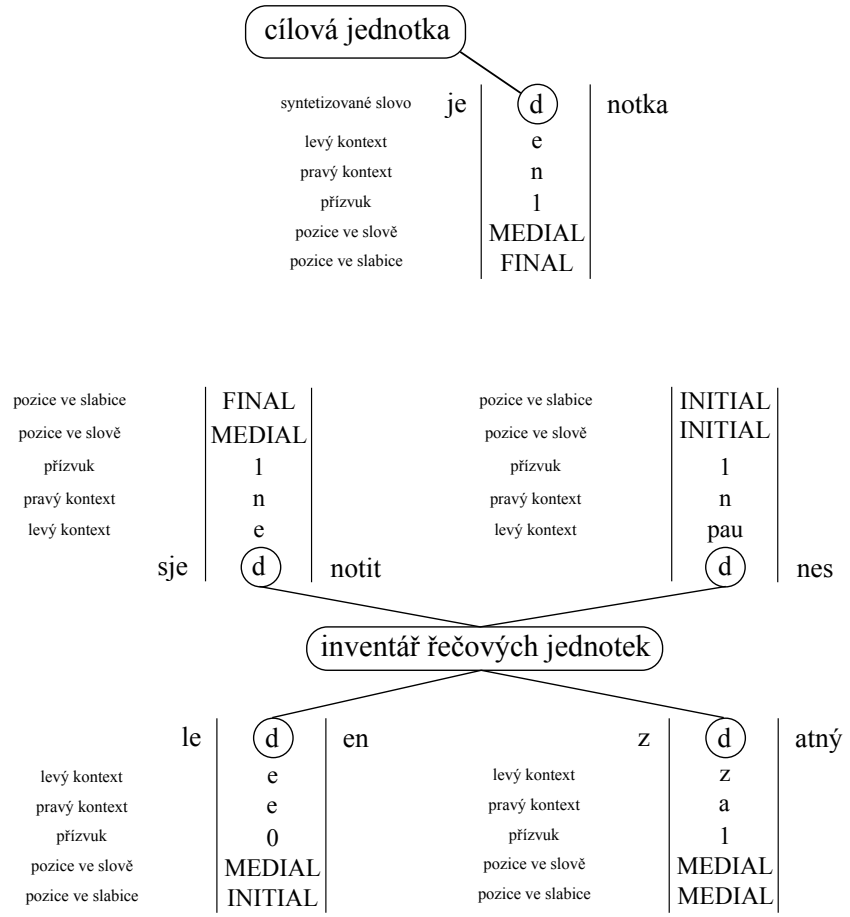
Celková cena cíle pro daného kandidáta je potom počítána jako vážený součet vzdáleností mezi jednotlivými elementy vektoru příznaků cílové jednotky a vektoru příznaků konkrétního kandidáta podle rovnice

$$C^t(t_i, u_i) = \sum_{k=1}^{p} w_k^t C_k^t(t_i, u_i), \tag{2.1}$$

kde t_i je cílová jednotka, u_i je jednotka v inventáři odpovídající cílové jednotce t_i, w^t je vektor vah, C_k^t je míra vzdálenosti k-tého příznaku a p je počet všech příznaků.

Cena řetězení

Cena řetězení C^c hodnotí, jak dobře nebo špatně se dvě sousedící jednotky při syntéze řetězí, nebo-li jak hladce se daný kandidát u_i bude spojovat s potencionálním kandidátem pro předchozí jednotku u_{i-1}. Příznaky se v tomto případě počítají přímo z řečového signálu v místě řetězení. Nejvíce se v současné době používají MFCC koeficienty, popř. další parametry jako $F0$ nebo intenzita signálu. Existuje však mnoho různých koeficientů, a snahou je získat co nejlepší popis nespojitosti — v nejlepším případě tak, jak ji vnímají sami posluchači. Stejně tak je na výběr velké množství měr, které lze použít k určení vzdálenosti dvou po sobě jdoucích vektorů příznaků. Některé metody, jak tyto příznaky vybrat (resp. jaké koeficienty a jaké míry použít), jsou uvedeny v [119], popř. [80]. Mimo již zmíněných MFCC koeficientů jmenujme například parametry LSF (*Line Spectral Frequencies*) či MCA (*Multiple Centroid Analysis*), z možných měr pak jednoduchou Euklidovskou vzdálenost nebo některé složitější jako divergence Kullback–Leibler či vzdálenost Itakura-Saito.



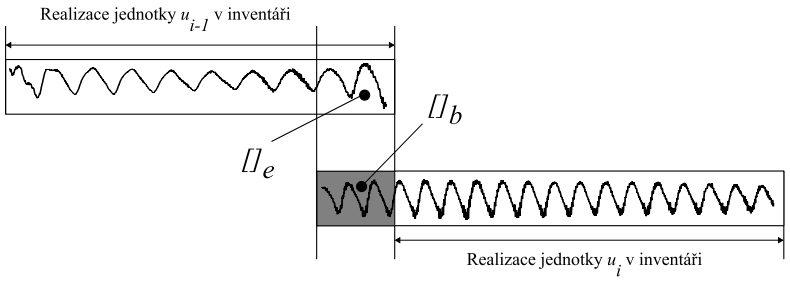
Obrázek 2.3: Některé prozodické charakteristiky fonémové jednotky [d], které se mohou využívat pro výpočet ceny cíle. Jak je vidět, nejlepší volbou z hlediska ceny cíle by byl kandidát ze slova „sjednotit“, neboť jeho prozodické charakteristiky jsou stejné jako u cílové jednotky

Stejně jako u ceny cíle je i celková cena řetězení ve většině případů počítána jako vážený součet konkatenačních subcen, a to podle rovnice

$$C^c(u_{i-1}, u_i) = \sum_{k=1}^{q} w_k^c C_k^c(u_{i-1}, u_i), \tag{2.2}$$

kde u_{i-1} je jednotka předcházející jednotce u_i, w^c je vektor vah, C_k^c je míra vzdálenosti k-tého příznaku a q je počet všech příznaků.

Jak je zobrazeno na obrázku 2.4, cena řetězení se obvykle počítá mezi koncovým úsekem realizace jednotky u_{i-1} (označeném jako $[]_e$) a úsekem



Obrázek 2.4: Určování ceny řetězení

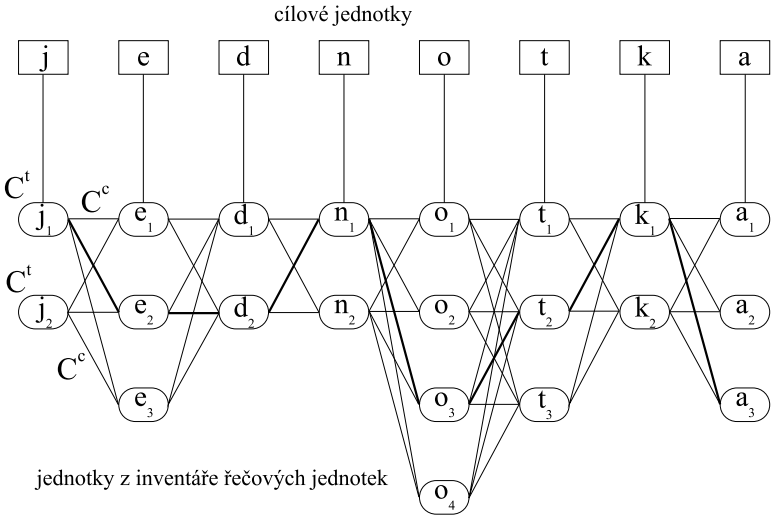
bezprostředně předcházejícím realizaci jednotky u_i (označeném jako $[]_b$). Tím je zaručeno, že cena bude nulová v případě, že tyto dvě uvažované realizace jednotek byly v korpusu uloženy za sebou.

Výběr optimální posloupnosti kandidátů

Z jednotlivých kandidátů je potřeba na základě určených cen vybrat jejich nejlepší posloupnost, a to tak, aby výsledná celková cena byla minimální. Různé kombinace těchto posloupností mohou být znázorněny jako graf, kde ohodnocení uzlu odpovídá cena cíle a ohodnocení hrany odpovídá cena řetězení, jak je vidět na obrázku 2.5. Problémem tedy potom je najít v grafu cestu s nejnižší cenou, což lze řešit aplikací různých optimalizačních algoritmů, například Viterbiho algoritmem. Detailnější informace o metodě dynamického výběru jednotek můžeme nalézt mimo jiné např. v [111].

2.4 HMM syntéza

HMM syntéza patří do skupiny metod syntézy řeči pomocí statistické parametrizace. U této skupiny metod je z přirozeného řečového signálu vytvořena (natrénována) množina statistických modelů, které reprezentují parametry jednotlivých řečových jednotek. Protože se téměř výhradně využívá skrytých třístavových nebo pětistavových Markovových modelů (HMM z anglického *Hidden Markov Model*), mluvíme většinou právě o HMM syntéze. Statistické modely pak slouží k modelování průběhu řečového signálu a původní přirozený řečový signál se tak již v době syntézy nepoužívá. Použití těchto modelů má dvě fáze – trénovací část a vlastní syntézu. Proces HMM syntézy je znázorněn na obrázku 2.6 a popsán např. v [130, 132] nebo pro češtinu pak v [48, 49].

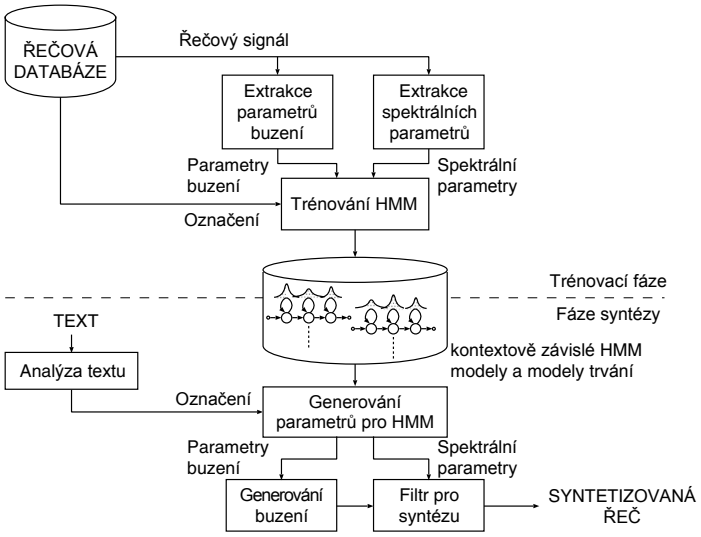


Obrázek 2.5: Graf požadovaných jednotek a jejich realizací z řečového korpusu použitých k syntéze slova „jednotka“. Jednotlivé uzly grafu představují kandidáty z inventáře řečových jednotek. Každému uzlu je přiřazena cena cíle C^t a každé hraně grafu je přiřazena cena řetězení C^c. Zvýrazněná cesta je cesta s minimálním ohodnocením.

V první fázi je potřeba HMM natrénovat, což se provádí pomocí reálných dat, podobně jako v úloze rozpoznávání řeči. Rozdíl je v tom, že z řečové databáze se v tomto případě extrahují jak spektrální parametry řeči (např. MFCC koeficienty a jejich dynamické změny – delta a delta-delta koeficienty), tak i parametry pro buzení ($\log F0$ a její dynamické změny – delta a delta-delta koeficienty). Tímto způsobem se trénují parametry kontextově závislých HMM (v úvahu se bere kontext fonetický, lingvistický a prozodický). Hodnoty $F0$ se modelují pomocí tzv. víceprostorového rozdělení pravděpodobnosti [113]. Každý stav HMM navíc obsahuje i model trvání stavu, čímž se modeluje doba trvání. Kvůli větší robustnosti systému se často využívá shlukovacích algoritmů, kdy jednotlivé kontextově závislé modely vytvářejí shluky na základě rozhodovacích stromů. Tím se zamezuje problémům, které mohou vzniknout v důsledku nedostatečného počtu dat.

Ve druhé fázi dochází k vlastní syntéze. Vstupní text je převeden do posloupnosti kontextově závislých HMM modelů[1], tato posloupnost tvoří

[1]Zde se využívají rozhodovací stromy, které vznikly při trénování.



Obrázek 2.6: Znázornění procesu HMM syntézy, a to jak trénovací části, tak i fáze vlastní syntézy.

tzv. HMM promluvu[2]. Dále jsou určeny doby trvání jednotlivých hlásek na základě natrénovaných hustot pravděpodobností setrvání ve stavu. Z natrénovaných HMM jsou generovány parametry řeči (MFCC koeficienty a hodnoty $F0$) metodou maximální věrohodnosti. Nakonec je ze spektrálních parametrů a parametrů pro buzení vytvořen výstupní řečový signál použitím syntetického filtru (nejčastěji se využívá MLSA filtr). Podrobnější informace o HMM syntéze lze najít například v [48, 114, 115, 130] nebo [8], kde je uvedeno i porovnání HMM syntézy a syntézy metodou dynamického výběru jednotek.

[2]V jistém smyslu lze tedy i zde mluvit o konkatenaci, přičemž se řetězí HMM modely, na rozdíl od dříve popsané konkatenační syntézy, kde dochází k řetězení přirozeného (popř. modifikovaného) řečového signálu v podobě řečových jednotek.

Kapitola 3

Popis expresivity

Obecný popis expresivity jako takové není sám o sobě jednoduchým úkolem a zabývají se jím různé studie, např. [25]. Pro různé vědní obory a jejich úlohy existují různé možnosti popisu expresivity. V zásadě se však objevuje několik určitých principů, na jejichž základě jsme schopni expresivitu vyskytující se v lidském chování popsat. Na následujících řádcích bude představeno několik nejdůležitějších a nejpoužívanějších možností, které se liší tím, jak je vhodné pro daný úkol expresivitu definovat.

Pro obecné studie zabývající se lidským chováním jsou zřejmě nejpoužívanější teorie popisující emotivní stavy člověka, tedy emoce. V části 3.1 tedy bude představen princip takového popisu společně s krátkým úvodem k samotnému vzniku emotivních stavů u člověka, proč k nim dochází, jak se projevují a jaké různé varianty takových stavů rozlišujeme. Nejdůležitější informací pro naši další práci pak bude nastínění toho, jaký popis emocí se využívá v syntéze expresivní (v tomto případě můžeme mluvit o emotivní) řeči.

Pro výzkum v oblasti syntézy řeči a použití syntetické řeči v dialogu pak je v části 3.2 představen souhrn nejpoužívanějších popisů různých expresivních kategorií nazývaných povětšinou dialogové akty[1]. V podstatě se tyto různé přístupy diskrétního popisu expresivních kategorií liší především oblastí, ve které se používají. Jak již bylo uvedeno výše, obecný popis expresivity (a nejenom v řeči) je velmi komplikovanou záležitostí. Z tohoto důvodu se vyvíjejí různá alternativní schémata popisu dialogových aktů většinou (avšak ne vždy) přímo na míru pro určité aplikace, v nichž se plánuje jejich využití.

[1]Někdy se lze setkat i s označením *komunikační/komunikativní* funkce, což lze chápat jako nejobecnější označení pro jakýkoli úkon, který výpověď v průběhu řečové interakce plní, přičemž půjde o úkony na různých rovinách komunikační události [53]. Tohoto označení je užíváno i v našich publikacích, nicméně pro tuto práci jsme zvolili označení dialogový akt.

Na základě uvedených informací jsme se tedy i my vydali cestou dialogových aktů, neboť náš výsledný expresivní systém TTS by měl najít využití především jako aplikace dialogového systému v předem vymezené oblasti, jak bude později popsáno v kapitole 6. Námi vybrané a poté použité konkrétní dialogové akty budou popsány v části 6.4.

3.1 Emoce

Informace o tom, co jsou to emoce, jak u člověka vznikají, čím se projevují, a různé pohledy na ně uvádí ve stručnosti např. [54] nebo [100]. Podle [99] je problémem už samotné rozlišení emocí od ostatních typů afektivních stavů řečníka, jako např. jeho nálada, mezilidské vztahy, osobní postoj nebo osobitý charakter. V tabulce 3.1 jsou tyto jednotlivé rozdílné stavy řečníka popsány různými charakteristikami:

a) intenzitou;
b) trváním;
c) mírou časového sjednocení (synchronizací) reakcí jednotlivých částí organismu;
d) tím, do jaké míry je reakce zaměřena na událost, která ji vyvolala;
e) tím, jakou roli hrálo ve vyvolání reakce vyhodnocení stávající situace;
f) rychlostí změny stavu;
g) rozsahem, jakým tento stav ovlivňuje chování řečníka.

Tabulka 3.1: Rozdílné afektivní stavy podle [99]. Hodnoty jednotlivých charakteristik mají následující význam: 0 - nízký, 1 - střední, 2 - vysoký, 3 - velmi vysoký.

Typ afektivního stavu	a	b	c	d	e	f	g
emoce	2-3	1	3	3	3	3	3
nálada	1-2	2	1	1	1	2	1
mezilidské vztahy	1-2	1-2	1	2	1	3	2
osobní postoj	0-2	2-3	0	0	1	0-1	1
osobitý charakter	0-1	3	0	0	0	0	1

Z uvedené tabulky vyplývá následující:

Emoce je relativně krátká synchronizovaná reakce většiny (nebo všech) částí organismu na výsledek vyhodnocení vnějšího nebo vnitřního podnětu velkého významu (zlost, smutek, radost, strach, hanba, hrdost, ...).

Nálada je afektivní stav nízké intenzity a dlouhého trvání, který odpovídá změně subjektivního pocitu, často bez zjevné příčiny (srdečnost, sklíčenost, podrážděnost, netečnost, deprese, optimismus, ...).

Mezilidské vztahy vyjadřují afektivní přístup k druhé osobě v určité situaci, kdy dochází ke vzájemné interakci (zdrženlivost, bezcitnost, vroucnost, pohrdavost, ...).

Osobní postoj je relativně trvalé a afektivně zabarvené smýšlení o ostatních objektech nebo lidech (náklonnost, láska, nenávist, důležitost, touha, ...).

Osobitý charakter je vlastně stálá povaha a způsob chování, typický pro danou osobu (nervozita, starostlivost, bezstarostnost, mrzutost, zaujatost, závistivost, podezíravost, ...).

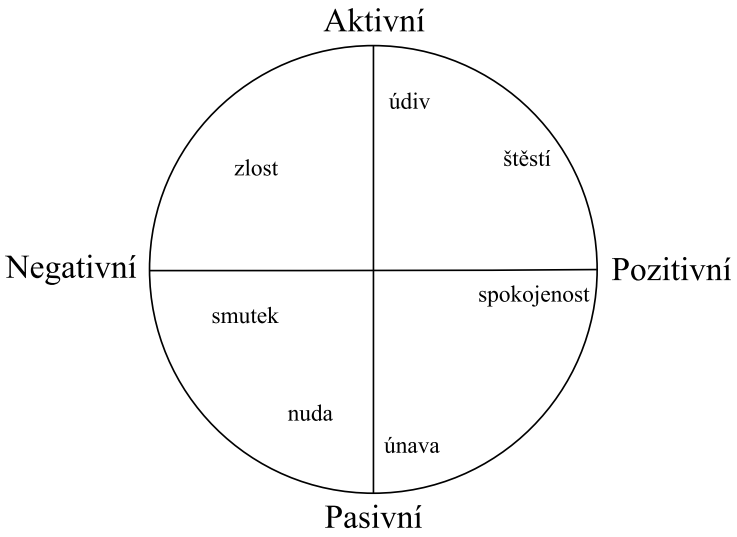
Emoce tedy tvoří zvláštní skupinu afektivních stavů, snadno odlišitelnou od všech ostatních. Vyznačují se vysokou intenzitou, ale velmi krátkou dobou trvání. Jsou hodně zaměřené na událost, která je vyvolala, velmi ovlivňují následné chování člověka a mohou se velmi rychle měnit. Je zde ještě zapotřebí zmínit existenci silného regulačního mechanismu, který určuje jak bude daná emoce prezentována navenek. Tato regulace je obzvláště nápadná u velmi intenzivních emocí, jako je např. vztek, zoufalost, velký strach, atd., které jsou často za normálních okolností maskovány.

Dále se budeme věnovat spíše „technickému“ popisu emocí. Jedním z prvních a zároveň základních problémů týkajících se syntézy emotivní řeči je rozdělení emocí. Otázkou je, jakým způsobem emoce klasifikovat. K dispozici máme dva různé teoretické základy.

Jedním z nich je rozdělení diskrétní, kde se předpokládá existence malého počtu základních (primárních, čistých) a většího počtu odvozených (sekundárních – nikoliv však menšího významu) emocí, které jsou charakteristické určitou fyziologickou odezvou a také obličejovým a hlasovým vyjádřením. Jako primární emoce jsou většinou označovány: strach, zlost, radost, smutek, překvapení a znechucení (tzv. „velká šestka“ [24]).

Druhý přístup vychází z myšlenky, že emoce jsou rozprostřeny spojitě ve dvourozměrném nebo vícerozměrném prostoru. Popisem emocí ve dvourozměrném prostoru se zabývá např. [98] (zobrazeno na obrázku 3.1), v tří-

rozměrném pak [76] – tzv. PAD reprezentace[2], tj. emoce jsou spojitě rozprostřené v prostoru na třech osách reprezentující pozitivní/negativní, vzrušené/nevzrušené, dominantní/submisivní charakteristiku té které emoce. Další teorií zabývající se spojitým rozdělením emocí je pak Plutchikovo těleso emocí [83, 81] (zobrazeno na obrázku 3.2).

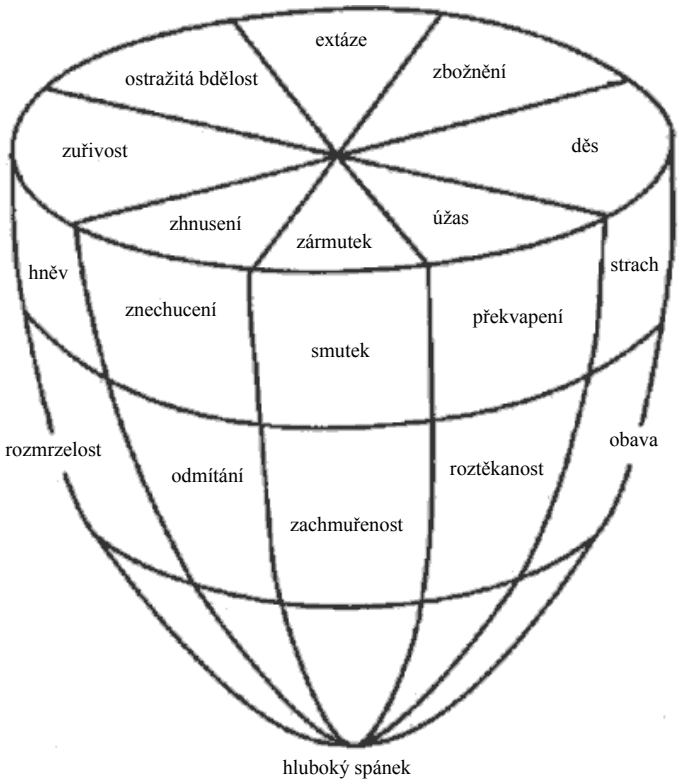


Obrázek 3.1: Prostor obsahující emoce je uzavřen kružnicí. Vodorovná osa naznačuje, zda se jedná o emoce negativní či pozitivní, svislá osa pak určuje, zda jde o emoce pasivní nebo aktivní.

Jaké z toho plynou důsledky pro výzkum expresivní/emotivní syntézy řeči? Jakým směrem by se měl popis emocí ubírat, pokud bychom ho plánovali využít v syntéze emotivní řeči?

Diskrétní rozdělení by bylo pro syntézu řeči samozřejmě mnohem jednodušší, spočívalo by v pouhém nadefinování jednotlivých kategorií, např. radost, štěstí, smutek, únava, nuda, zlost, odpor, strach, vzrušení, uvolněnost, atd. Jak ale tyto kategorie vybrat, aby bylo zastoupení možných stavů mluvčího nějak rovnoměrně pokryto alespoň v daném kontextu? A kolik jich vybrat? Jak moc jsou si jednotlivé kategorie podobné? Co třeba odlišuje radost od štěstí a jak moc jsou od sebe tyto dvě emoce „vzdálené“? Jakým způsobem by byly vyjádřeny různé stupně jednotlivých emocí? Na tyto otázky zatím nemáme uspokojivé odpovědi. Ale většina pokusů, které se týkaly emotivní syntézy a které nám jsou známy, používala právě diskrétní rozdělení emocí.

[2]Označení pochází z anglického pojmenování tří dimenzí, **p**leasure-displeasure, **a**rousal-nonarousal a **d**ominance-submissiveness.



Obrázek 3.2: Plutchikovo těleso emocí (převzato z [81]).

Naproti tomu by byl kontinuální popis emotivních stavů např. podle [98] pro potřeby syntézy emotivní řeči mnohem komplikovanější. Emoce lze tímto způsobem matematicky popsat ve dvourozměrném prostoru, nesyntetizovali bychom tedy řeč v emoci „smutek“, ale v emoci o souřadnicích např. $[-0{,}43; -0{,}2]$, pokud budeme uvažovat, že kruh zahrnující celý prostor emocí má poloměr $r = 1$. Pokud ovšem předpokládáme, že na vstupu syntetizéru máme čistý text, musíme jeho jednotlivým částem, popř. textu jako celku, také přesně přiřadit odpovídající souřadnice, abychom věděli, jak danou promluvu syntetizovat. V kontinuálním přístupu máme v tomto okamžiku nekonečně mnoho možností, jak tyto souřadnice vybrat, na rozdíl od diskrétního přístupu, kde lze vstupní text snáze zařadit do nějaké z několika málo daných kategorií. Vstupní text by tedy musel projít rozsáhlejší analýzou, při které bychom souřadnice museli určit, např. s využitím slovníku podobnému [121], kde jsou každému slovu přiřazeny číselné hodnoty pro 3 různé dimenze: příjemnost, aktivaci a představivost. To by zahrnovalo tvorbu takového slovníku pro příslušný jazyk, v našem případě češtinu, což jistě není triviální úkol.

Další otázkou ještě je, zda tyto souřadnice určovat pro celý vstupní text najednou, nebo zda je určovat po částech, např. po větách či slovech. V kontinuálním rozdělení máme tu výhodu, že souřadnice by se mohly plynule měnit během celé promluvy a výsledná syntetická řeč by zněla velmi přirozeně. Problémem pak bude syntéza jako taková, která zřejmě tyto malé odlišnosti nebude schopna vyprodukovat.

Na druhou stranu bychom se v případě kontinuálního rozdělení nemuseli omezovat pouze na vybrané emoce, měli bychom k dispozici celý prostor emocí i s jejich nejjemnějším rozdělením. Lehce bychom také mohli určit měřitelné rozdíly mezi jednotlivými emocemi, protože již jejich samotné rozprostření v prostoru mezi nimi určuje jisté vzdálenosti.

3.2 Dialogové akty

Jak již bylo uvedeno dříve, v námi vyvíjeném dialogovém systému se z hlediska syntézy řeči nebudeme snažit o vyjádření emocí, ale pouze jistých expresivních stavů, ve kterých ovšem mohou být emoce také obsaženy. Jedná se tedy o diskrétní popis expresivity, který byl již dříve použit u konkrétních dialogových systémů, jako např. [106, 129] nebo také jako pomocný mechanismus pro rozpoznávání řeči v dialogu [104].

Zjednodušeně řečeno, dialogový akt je určitá část rozhovoru řečená určitým způsobem nebo s určitým cílem, což může být z pohledu expresivity v řeči jak neutrálně, tak také expresivně. V tomto případě nás nezajímá konkrétní vyjádření určité emoce, ale předpokládáme její relevantní vyjádření v rámci dialogového aktu. Návrhem množiny dialogových aktů se zabývá mnoho studií, shrnutí a porovnání nejznámějších doposud navržených schémat nabízí např. [65]. Pro ukázku jsme vybrali čtyři možná schémata popisu dialogových aktů, z nichž jsme také vyšli při návrhu našeho vlastního. Důvodem pro návrh nového schématu podrobněji popsaného v části 6.4 byla nutnost popisu dialogových aktů v rámci konkrétní aplikace dialogového systému popsaného v kapitole 6.

3.2.1 DAMSL

Schéma popisu dialogových aktů s názvem DAMSL (Discourse Annotation and Markup System of Labeling, [23, 4]) bylo vytvořeno Discourse Research Initiative s cílem popsat obecně dialogové akty objevující se v různých typech dialogů. Využívá čtyři základní kategorie:

- komunikativní status – zaznamenává, zda je promluva srozumitelná a zda byla úspěšně dokončena;

- informační úroveň – charakterizuje významový obsah promluvy;
- dopředné funkce – reprezentují, jak současná promluva ovlivňuje budoucí akce;
- zpětné funkce – reprezentují, jaký vztah má současná promluva k minulé části dialogu.

Celé schéma DAMSL je znázorněno na obrázku 3.3, názvy dialogových aktů jsou v původním anglickém znění. Dodejme, že cílem bylo popsat spíše význam jednotlivých promluv v dialogu, nikoliv jejich expresivní vyjádření směrem k posluchači.

Communicative-status

- Uninterpretable
- Abandoned
- Self-talk

Information-Level

- Task
- Task-management
- Communication-management
- Other-level

Forward Looking Function

- Statement
  - Assert
  - Reassert
  - Other-statement
- Influencing-addressee-future-action
  - Open-option
  - Action-directive
- Info-request
- Committing-speaker-future-action
  - Offer
  - Commit
- Conventional
  - Opening
  - Closing
- Explicit-performative
- Exclamation
- Other-forward-function

Backward Looking Function

- Agreement
  - Accept
  - Accept-part
  - Maybe
  - Reject-part
  - Reject
  - Hold
- Understanding
  - Signal-non-understanding
  - Signal-understanding
    - Acknowledge
    - Repeat-rephrase
    - Completion
  - Correct-misspeaking
- Answer
- Information-relation

Obrázek 3.3: Schéma DAMSL pro popis dialogových aktů.

3.2.2 SWBD–DAMSL

Schéma SWBD-DAMSL je schéma dialogových aktů odvozené od předchozího DAMSL a je cíleno na popis korpusu SWITCHBOARD. To je korpus obsahující spontánní rozhovory mezi dvěma účastníky telefonního hovoru spojenými automatickým systémem – spojovatelkou – vytvořený firmou Texas Instruments. Sestává se z 2430 konverzací o průměrné délce 6 minut s celkovým počtem řečníků přesahujícím 500, mluvících americkou angličtinou na počítačem určené téma. Podrobnější popis systému je uveden například v [61]. Schéma dialogových aktů vyvinuté pro tento úkol je znázorněno na obrázku 3.4, názvy dialogových aktů jsou opět ponechány v původním anglickém znění.

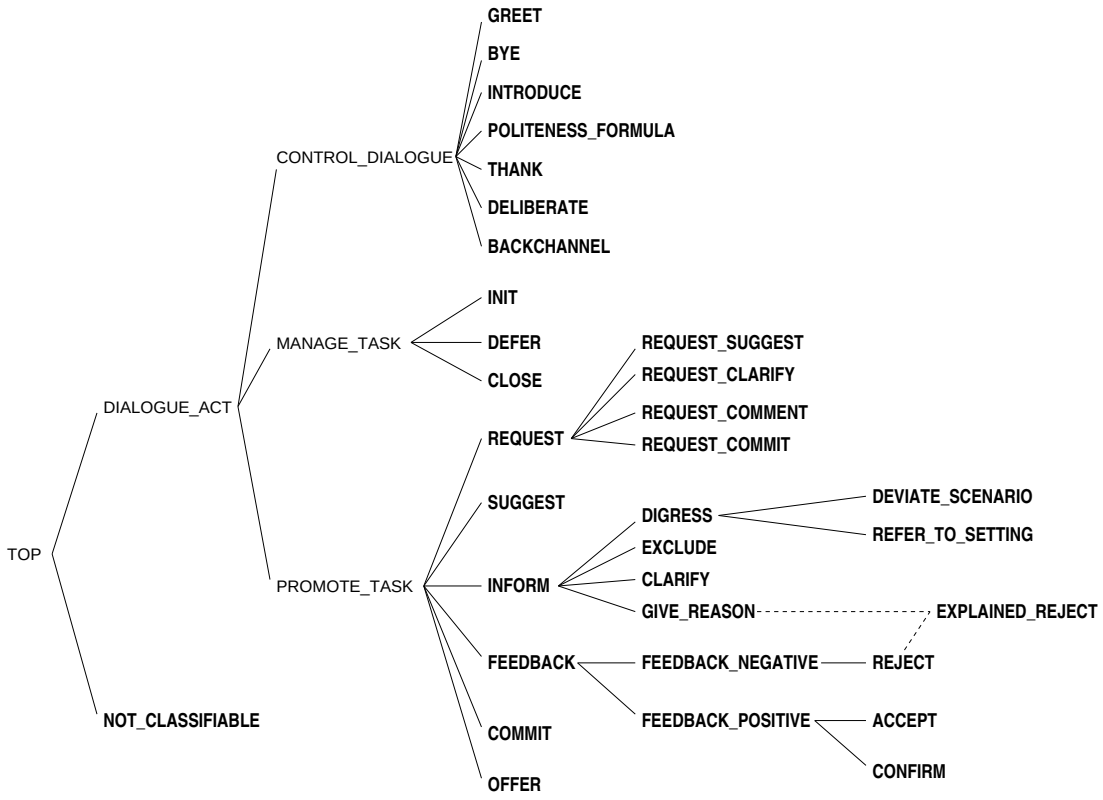
- Communicative-Status
  - Uninterpretable (%): *But, uh, yeah.*
  - Non-verbal (x): *[Laughter]*
  - Abandoned or Turn-Exit (% -): *So,-*
  - Self-talk (t1): *What's the world I'm looking for...*
  - 3rd-party-talk (t3): *My goodness, Diane, get down from there.*
- Forward-Communicative-Function
  - Statement
    - Statement-non-opinion (sd): *Me, I'm in the legal department.*
    - Statement-opinion (sv): *I think it's great.*
  - Influencing-addressee-future-action
    - Yes-No-Question (qy): *Do you have to have any special training?*
    - Wh-Question (qw): *Well, how old are you?*
    - Open-Question (qo): *How about you?*
    - Or-Clause (qrr): *Or is it more of a company?*
    - Declarative Yes-No-Question (qy^d): *So you can afford to get a house?*
    - Declarative Wh-Question (qw^d): *You are what kind of buff?*
    - Tag-Question (^g): *Right?*
    - Action-directive (ad): *Why don't you go first?*
    - Backchannel in question form (bh): *Is that right?*
    - Rhetorical-Questions (qh): *Who would steal a newspaper?*
  - Committing-speaker-future-action
    - Offers, Options Commits (oo,cc,co): *I'll have to check that out.*
  - Other-forward-function
    - Conventional-opening (fp): *How are you?*
    - Conventional-closing (fc): *Well, it's been nice talking to you.*
    - Thanking (ft): *Hey thanks a lot.*
    - Apology (fa): *I'm sorry.*
- Backwards-Communicative-Function
  - Agreement
    - Agree/Accept (aa): *That's exactly it.*
    - Maybe/Accept-part (aap/am): *Something like that.*
    - Reject (ar): *Well, no.*
    - Hold before answer/agreement (^h): *I'm drawing a blank.*
  - Understanding
    - Signal-non-understanding (br): *Excuse me?*
    - Response Acknowledgement (bk): *Oh, okay.*
    - Repeat-phase (b^m): *Oh, fajitas.*
    - Collaborative Completion (^2): *Who aren't contributing?*
    - Acknowledge (b): *Uh-huh.*
    - Summarize/reformulate (bf): *Oh, you mean you switched schools for the kids.*
    - Appreciation (ba): *I can imagine.*
    - Downplayer (bd): *That's all right.*
  - Answer
    - Yes answers (ny): *Yes.*
    - No answers (nn): *No.*
    - Affirmative non-yes answers (na,ny^e): *It is.*
    - Negative non-no answers (ng,nn^e): *Uh, not a whole lot.*
    - Other answers (no): *I don't know.*
    - Dispreferred answers (arp,nd): *Well, not so much that.*
- Other
  - Quotation (^q): *You can't be pregnant and have cats.*
  - Hedge (h): *I don't know if I'm making any sense or not.*

Obrázek 3.4: Schéma SWBD-DAMSL pro popis dialogových aktů.

3.2.3 VERBMOBIL

Projekt VERBMOBIL [59] kombinuje dvě výzkumné oblasti a to zpracování řeči a automatický překlad. Jeho cílem je vytvořit systém automatického překladu z jednoho jazyka do jiného (angličtina – němčina) pro doménu vymezenou rozhovorem dvou lidí, kteří se snaží si sjednat obchodní schůzku. Celý systém byl později rozšířen také na plánování cest a označen jako VERBMOBIL-2 [3], rozšířeno bylo i portfolio jazyků, se kterými je systém schopen pracovat (angličtina – němčina – japonština).

Pro zvýšení úspěšnosti takového systému se využívá predikce dialogových aktů [93]. Použité dialogové akty jsou znázorněné na obrázku 3.5. Schéma je převzato z [3] a názvy dialogových aktů jsou znázorněny v původním znění.



Obrázek 3.5: Schéma VERBMOBIL pro popis dialogových aktů.

3.2.4 AT&T

Jako poslední příklad uvedeme schéma dialogových aktů AT&T Labs popsané v [107]. Jejich práce a navržené dialogové akty se pro nás staly inspirací pro vytvoření vlastní množiny dialogových aktů uvedené v části 6.4. Další postup AT&T v oblasti syntézy řeči pro dialogový systém je popsán v [106, 103]. Dialogové akty AT&T navržené pro obecný dialogový systém jsou zobrazeny v tabulce 3.2.

Tabulka 3.2: Schéma AT&T pro popis dialogových aktů.

Imperative: directs actions of others			
Speech Act	Abbr.	Num.	Examples
Request	Req	319	*Please enter your PIN.*
Directive	Dir	459	*Turn left onto Main Street.*
Warning	Warn	7	*Be prepared to stop.*
Repeat	Rept	62	*Pardon me?*
Wait	Wait	121	*Just a second please.*
Interrogative: solicits information from others			
Speech Act	Abbr.	Num.	Examples
Question-wh	Qwh	641	*Who should I call?*
Quest.-yes/no	Qyn	2394	*Are you flying to Cleveland?*
Quest.-mult.choice	Qmc	100	*Downtown or near the airport?*
Assertive: conveys factual information to others			
Speech Act	Abbr.	Num.	Examples
Inform.-detail	Idet	464	*VTL dash help at VT dot net.*
Inform.-general	Igen	4713	*You have four new messages.*
Affective: expresses the speaker's attitude			
Speech Act	Abbr.	Num.	Examples
Greeting	Grt	205	*Hi! Welcome to Call ATT.*
Apology	Apol	355	*I'm sorry.*
Exclam.-negative	Eneg	17	*Oops! Oh dear!*
Exclam.-positive	Epos	16	*Great!*
Thanks	Thks	129	*Thanks for calling.*
Goodbye	Gbye	39	*Bye bye.*
Cue phrase	Cue	349	*Meanwhile, ... Well, ...*
Back-channel	Fill	32	*Hmmm. Uh-huh.*
Other?			
Speech Act	Abbr.	Num.	Examples
Confirmation	Conf	1728	*All right.*
Disconfirmation	Dis	1670	*No, you must change terminals.*

Kapitola 4

Syntéza expresivní řeči

V kapitole 2 jsme popsali, jak můžeme v současné době různými technikami syntézy řeči vytvářet srozumitelnou řeč, která bude na velmi vysoké úrovni. Takto produkovaná syntetická řeč však nebude znít úplně přirozeně, dokud nebude vyjadřovat postoj mluvčího, jeho emotivní nebo afektivní stav, nebo jeho náladu, postoj, pocity. Proto se syntéza expresivní řeči stala v poslední době velmi diskutovaným tématem. Ačkoliv už bylo publikováno několik výsledků (a to zejména pro angličtinu [133, 47, 10, 14], japonštinu [57], němčinu [13, 11] či španělštinu [77]), problém syntézy expresivní řeči ještě nebyl uspokojivě vyřešen.

Jen pro úplnost uveďme, že v současné době existuje několik málo známých systémů TTS, které jsou schopné expresivní řeč produkovat. Z komerčních jmenujme *Nuance Vocalizer* od Nuance Communications (`http://www.nuance.com`), z volně dostupných pak *MARY Open-Source Emotional Text-to-Speech Synthesis System* od Deutches Forschungszentrum für Künstliche Intelligenz GmbH (`http://www.dfki.de`) nebo *EmoFilt* (`http://emofilt.syntheticspeech.de`), což však není zcela samostatný systém TTS (je to pouze modul, který se využívá ve spojení se systémem MBROLA [2]). Nesmíme zapomenout ani na expresivní syntetizér firmy IBM [82], japonské pracovní skupiny HTS [128] nebo nově vyvíjený *VivoText* (`http://www.vivotext.com`).

4.1 Používané přístupy

Základní rozdělení metod používaných pro syntézu expresivní řeči je stejné, jako bylo uvedeno v kapitole 2 pro syntézu neutrální řeči. Tyto metody, jejich parametry a algoritmy jsou pro simulaci expresivity v syntetické řeči různě modifikovány. Již dříve jsme uvedli, že nejpoužívanější metodou pro syntézu

řeči je v současnosti konkatenační syntéza. Je tedy celkem pochopitelné, že je snaha produkovat tímto způsobem i expresivní řeč. Také proto je cílem dalšího výzkumu zjistit především možnosti konkatenační syntézy v této oblasti, přestože podle [29] zde má tento přístup omezené použití. Navzdory tomu, že vývoj HMM syntézy pro češtinu je teprve v začátcích, pokusíme se alespoň nahlédnout do produkce expresivní řeči touto metodou pro jiné jazyky a také budeme prezentovat i naše prvotní experimenty s českou expresivní HMM syntézou. Jen ve stručnosti také představíme expresivní formantovou syntézu a uvedeme některé poznatky z literatury. Artikulační syntézou se v této práci kvůli její složitosti již nebudeme zabývat vůbec, přestože právě tato metoda by mohla být jistě v budoucnu přínosem pro výzkum v oblasti jak syntézy expresivní řeči, tak syntézy řeči obecně. Uvedeme také metodu, která ve skutečnosti není metodou syntézy řeči, avšak umožňuje produkovat expresivní řeč především na základě řeči neutrální, a tou je konverze hlasu.

4.1.1 Formantová syntéza

Jak bylo zmíněno v části 2.1, formantová syntéza produkuje syntetickou řeč na základě definovaných pravidel a různých parametrů. Velké množství těchto parametrů skýtá velké možnosti v nastavení kvality syntetické řeči. Je tedy snadné si představit, že správným nastavením těch správných parametrů můžeme i do formantově syntetizované řeči vnést přinejmenším známku expresivity. Příklad změny některých parametrů oproti neutrální řeči pro různé jazyky je zobrazen v tabulce 4.1 (převzato z [101]). Ve formantové syntéze samozřejmě nedochází k přímým změnám uvedených akustických parametrů, ale spíše k nastavení odpovídajících rezonátorů a antirezonátorů, které dané akustické parametry ovlivňují.

4.1.2 Konkatenační syntéza

Stejně, jako byly v části 2.3.1 uvedeny dvě základní možnosti pro vytvoření inventáře řečových jednotek, jsou i zde dvě různé možnosti, jak pracovat s nahraným řečovým korpusem. V následujících částech bude popsáno, jaké modifikace je potřeba implementovat do stávajících metod syntézy neutrální řeči, abychom získali na výstupu řeč expresivní.

Metoda s jedním zástupcem

Syntéza expresivní řeči použitím metody jednoho reprezentanta znamená změnu prozodických charakteristik jednotlivých jednotek. Dochází tedy

Tabulka 4.1: Změny v nastavení parametrů pro formantovou expresivní syntézu a úspěšnost rozpoznání zamýšlené emoce v syntetické řeči. Tabulka převzata z [101].

Emoce Jazyk Úspěšnost rozpoznání	Nastavení parametrů
Radost němčina 81 %	**F0 mean:** +50 % **F0 range:** +100 % **Tempo:** +30 % **Voice Qu.:** modal or tense; „lip-spreading feature“: F1 / F2 + 10 % **Other:** „wave pitch contour model“: main stressed syllables are raised (+100 %), syllables in between are lowered (-20 %)
Smutek americká angličtina 91 %	**F0 mean:** „0“, reference line „-1“, less final lowering „-5“ **F0 range:** „-5“, steeper accent shape „+6“ **Tempo:** „-10“, more fluent pauses „+5“, hesitation pauses „+10“ **Loudness:** „-5“ **Voice Qu.:** breathiness „+10“, brilliance „-9“ **Other:** stress frequency „+1“, precision of articulation „-5“
Zlost britská angličtina	**F0 mean:** +10 Hz **F0 range:** +9 s.t. **Tempo:** +30 wpm **Voice Qu.:** laryngealisation +78 %; F4 frequency -175 Hz **Other:** increase pitch of stressed vowels (2ary: +10 % of pitch range; 1ary: +20 %; empathic: +40 %)
Strach němčina 52 %	**F0 mean:** +150 % **F0 range:** +20 % **Tempo:** +30 % **Voice Qu.:** falsetto
Překvapení americká angličtina 44 %	**F0 mean:** „0“, reference line „-8“ **F0 range:** „+8“, steeply rising contour slope „+10“, steeper accent shape „+5“ **Tempo:** „+4“, less fluent pauses „-5“, hesitation pauses „-10“ **Loudness:** „+5“ **Voice Qu.:** brilliance „-3“
Nuda holandština 94 %	**F0 mean:** end frequency 65 Hz (male speech) **F0 range:** excursion size 4 s.t. **Tempo:** duration rel. to neutrality: 150 % **Voice Qu.:** modal or tense; „lip-spreading feature“: F1 / F2 + 10 % **Other:** final intonation pattern 3C, avoid final patterns 5&A and 12

k úpravě akustického signálu a tím také ke zhoršení kvality výsledné syntetizované řeči.

Změnou prozodických charakteristik se rozumí změna základní hlasivkové frekvence $F0$, popř. formantových frekvencí, změna doby trvání jednotlivých

řečových jednotek, atd. Měřitelné (a ovlivnitelné) prozodické charakteristiky, které se v expresivní řeči liší od hodnot stejných charakteristik v řeči neutrální, mohou být získány například prostřednictvím akustické analýzy řeči, která je popsaná v části 7.3.

Kromě prozodických charakteristik jednotlivých řečových jednotek (*segmentální charakteristiky*) je také potřeba brát v úvahu i celkovou prozodii syntetizované promluvy (*suprasegmentální charakteristiky*). Ta je vyjádřena např. pauzami mezi větnými úseky, celkovým průběhem $F0$ (intonací, melodií) nebo třeba důrazem či přízvukem.

Podle [77] (španělština) však modelování prozodie pomocí segmentálních a suprasegmentálních charakteristik nestačí. Byl proveden pokus, kdy pro syntézu byly vybrány jednotky např. z neutrálního korpusu a výsledná prozodie byla modelována použitím prozodických modelů pro všechny dostupné emoce. Potom byly vybrány jednotky z radostného korpusu a prozodie se opět modelovala všemi možnými způsoby. Vzniklo tedy několik syntetizovaných vět (druhá mocnina počtu emocí) s různými kombinacemi použitých korpusů a prozodických modelů. Poté byl proveden percepční test, ze kterého vyplynulo, že u některých emocí je pro správné rozpoznání důležitější použitý korpus, pro některé použitá prozodie. Další pokusy s modelováním prozodie lze najít např. v [91].

Vnímání expresivity v syntetizované řeči modelované pomocí změn prozodických charakteristik se tedy může lišit v závislosti na daném typu expresivního vyjádření, stejně tak jako může být závislé na daném řečníkovi. Každý jednotlivý řečník může pro různá expresivní vyjádření používat různé způsoby, velkou roli ve vnímání expresivity může také hrát tzv. barva hlasu (zahrnující takové charakteristiky jako např. zvučnost nebo nazalita) [38].

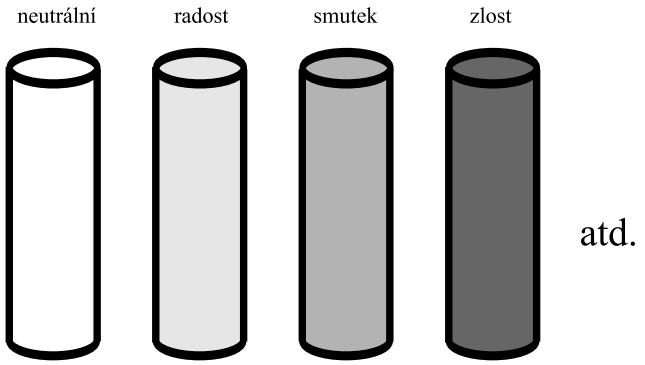
Metoda dynamického výběru jednotek

Pro syntézu expresivní řeči metodou dynamického výběru jednotek lze v zásadě použít dva postupy. Prvním zmíněným bude varianta nahrání korpusů pro každou kategorii popisující expresivitu zvlášť, druhý přístup používá jeden rozsáhlý korpus obsahující především neutrální řeč současně s menším množstvím nahraných expresivních dat. Z těchto korpusů jsou pak na základě zvoleného popisu expresivity vhodně vybírány řečové jednotky.

Kromě popisu expresivity, který jsme uvedli v kapitole 3, existuje ještě další možnost, jak vhodně popsat jednotky, které při dynamickém výběru z inventáře jednotek chceme preferovat. Tento postup, který uvádí např. [18], spočívá v popisu cílových jednotek ne dialogovými akty nebo jinými expresivními kategoriemi, jak bylo uvedeno dříve, ale přímo požadovanými akustickými parametry. Pro syntetizovanou promluvu namodelujeme průběh růz-

ných parametrů v rámci celé promluvy (např. F0, dobu trvání, apod.) a každou cílovou jednotku pomocí takového schématu tak popíšeme. Při samotném výběru pak upřednostňujeme ty jednotky, které nejlépe odpovídají požadovaným parametrům standardní cestou pomocí ceny cíle (viz část 2.3.3).

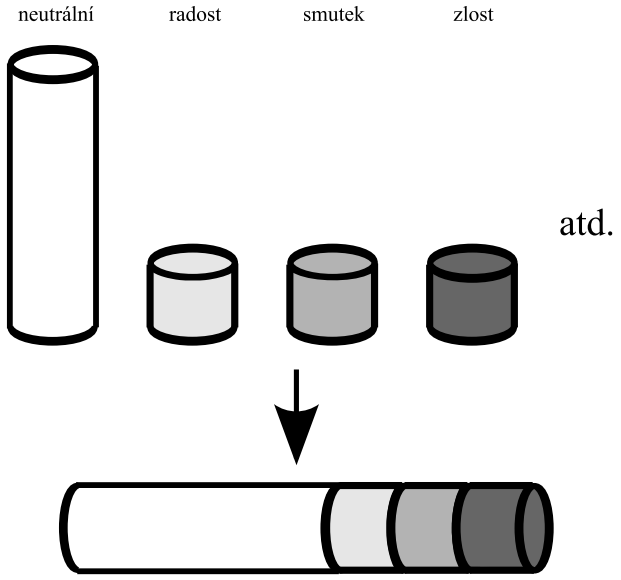
Více korpusů V tomto případě je potřeba obrovské množství dat, které musí řečník nahrát. Expresivní korpusy musí mít podobný rozsah, jako korpus neutrální, jak je znázorněno i na obrázku 4.1. Takový postup byl použit například v [58], kde se však pracuje s ne příliš velkými korpusy, požadavek na obrovské množství dat je tak velmi relativní. Velikost korpusů pak samozřejmě ovlivňuje kvalitu výsledné syntetické řeči. To pochopitelně platí jak pro syntézu řeči expresivní, tak i řeči neutrální.



Obrázek 4.1: Oddělené inventáře řečových jednotek: Pro každý korpus zvlášť jsou vytvořené jednotlivé inventáře řečových jednotek.

Během syntézy jedné promluvy jsou jednotky brány pouze z jednoho tohoto inventáře. A to z takového, který je nejblíže expresivnímu vyjádření, ve které chceme danou promluvu syntetizovat (ideálně máme nahraný přesně ten korpus, který potřebujeme). Jsme tak tedy závislí na jediném inventáři a to i v případě, že by se požadovaná jednotka vyskytovala v některém z ostatních korpusů např. v lepší kvalitě, vhodnějším kontextu, s lepšími prozodickými vlastnostmi, atd. Nemáme možnost, a to ani za cenu nějaké penalizace, použít jednotku z jiného korpusu. To se může, v porovnání s dále uvedenou metodou, jevit jako nevýhoda. Dále jsme také omezeni počtem expresivních kategorií podle toho, kolik máme expresivních korpusů. Výhodou naopak je, že teoreticky bychom měli pro expresivní řeč dosáhnout stejné kvality jako v případě syntézy neutrální řeči.

Jeden korpus Druhou variantou je nahrání rozsáhlého neutrálního korpusu a dále několika výrazně menších korpusů, opět pro každou kategorii popisující expresivitu zvlášť. Tyto korpusy jsou potom sloučeny dohromady, jak naznačuje obrázek 4.2, a jednotky získávají příznak, ze kterého korpusu pocházejí, tedy ke které expresivní kategorii patří. Nároky na množství nahraných dat jsou u tohoto přístupu značně menší. Úspora samozřejmě závisí na počtu kategorií, které mají být pro syntézu použity.



Obrázek 4.2: Společný inventář řečových jednotek: Jednotky ze všech nahraných korpusů (rozsáhlého neutrálního a menších expresivních korpusů) jsou uloženy v jediné databázi. U každé jednotky je zachován příznak, k jakému korpusu původně patřila.

Při syntéze promluvy, která je požadována v určité expresivní kategorii, se snažíme vybírat ty jednotky, které mají odpovídající příznak. Jednotky s jiným než požadovaným příznakem jsou penalizovány předem danou cenou podle penalizační matice, která tedy definuje podobnost různých expresivních kategorií. Příklad takové matice je vidět v tabulce 4.2 a lze si povšimnout, že matice nemusí být vždy nutně symetrická. Podrobný popis výpočtu konkrétních koeficientů pro námi použité dialogové akty je pak uveden v části 8.1. Aplikace penalizační matice tedy znamená, že při syntéze lze použít i jednotky zařazené k jiné expresivní kategorii než je ta požadovaná. Pokud totiž není nalezena prozodicky vhodná jednotka se správným příznakem, hledá se i mezi jednotkami s jinými příznaky, tedy v jiných korpusech. Snahou je v tomto případě minimalizovat celkovou cenu, viz část 2.3.3, přičemž penali-

zační matice je určitým způsobem začleněna do ceny cíle. Expresivní zabarvení syntetizované promluvy se tak může průběžně měnit (např. z expresivního na neutrální nebo naopak). Podle [54] lze řídit i intenzitu dané emoce, a to použitím většího (resp. menšího) počtu jednotek z expresivního korpusu doplněných jednotkami z neutrálního korpusu, a intenzitu expresí tak zesílit (resp. zeslabit). Avšak míchání jednotek z různých expresivních kategorií může s sebou také nést problémy spojené s kvalitou výsledné syntetické řeči.

Tabulka 4.2: Podle penalizační matice dochází k penalizaci v případě výběru jednotky s odlišným příznakem expresivní kategorie. Tato penalizace je součástí ceny cíle, viz část 2.3.3. Nastavení jednotlivých penalizačních hodnot je velmi důležité, zde uvedené hodnoty jsou uvedeny pouze jako příklad pro několik málo expresivních kategorií/emocí.

požadovaný příznak skutečný příznak	neutrální	radost	smutek	zlost
neutrální	0	0,25	0,25	0,25
radost	0,75	0	1	1
smutek	0,75	1	0	0,5
zlost	0,75	0,75	0,5	0

Další možností je kombinace obou výše uvedených postupů, a to smíchání stejně velkých, pokud možno co nejrozsáhlejších korpusů. Příklad je uveden v [55], kde byly takto smíchány tři méně rozsáhlé korpusy stejné velikosti (neutrální, radost, zlost; každý obsahoval 400 vět). Jednotky byly vybírány z výsledného korpusu, poměr expresivních a neutrálních jednotek pak vyjadřoval intenzitu dané expresivní kategorie.

4.1.3 HMM syntéza

V části 2.4 jsme uvedli, že základem pro HMM syntézu je správné natrénování HMM modelů z přirozeného řečového korpusu. I v tomto případě tak nejprve potřebujeme reálná řečová data, stejně jako u konkatenační syntézy. Pro účely expresivní syntézy řeči tak potřebujeme expresivní řečový korpus, ve kterém budou nějakým způsobem nadefinované expresivní kategorie.

Pro syntézu expresivní řeči (v práci [8] spíše označováno jako řečnický styl) byly např. v [127] navrženy dva postupy. První, označený jako *stylově závislé modelování*, předpokládá trénování HMM odděleně pro každý styl (expresivní kategorii). Vznikne tak tolik množin skrytých Markovových modelů,

kolik expresivních kategorií je v trénovací databázi. Druhá metoda, *stylově nezávislé modelování*, vytvoří pouze jednu množinu HMM. Expresivní kategorie se stává pouze jedním z příznaků kontextově závislého HMM. Oba dva zmíněné postupy jsou srovnatelné, co se kvality syntetizované řeči týká i s ohledem na napodobení požadovaného stylu. Syntetizovat pak tedy lze buď tu expresivní kategorii, která se vyskytla v původní trénovací databázi, nebo je možné využít např. metody interpolace stylů uvedenou v [108] pro syntézu jiné expresivní kategorie, která se v trénovacích datech nevyskytuje.

Naproti tomu [89] prezentuje další možnost HMM syntézy expresivní řeči a neomezuje se pouze na interpolaci mezi dvěma řečnickými styly. Je zde použit tzv. *model průměrné emoce*. HMM jsou natrénovány na databázi obsahující jak neutrální promluvy, tak i promluvy emotivně zabarvené. Před vlastní syntézou je pak ještě provedena fáze adaptace, kdy se natrénovaný model průměrné emoce adaptuje právě na tu emoci, která je vyžadována v syntetizované promluvě. K adaptaci je potřeba pouze malé množství řečových dat v dané emoci, ale není nutné, aby tato data byla využita pro trénování HMM.

Drobná úprava metody stylově nezávislého modelování je uvedena v [117]. Jako příznak kontextově závislého HMM je brána jak emoce, ve které řečník zamýšlel řeč pronést, tak i ta, která byla označena posluchači, resp. rozložení posluchači rozpoznaných emocí (v této práci se zabývají třemi diskrétními emocemi).

Další možnosti, jak uplatnit HMM syntézu pro umělé vytváření expresivní řeči je popsáno např. v [66], kde je využito omezeného rozsahu použití pouze na fotbalová oznámení – stačí menší trénovací množina a výsledky jsou pochopitelně také lepší. Prvotní výsledky české expresivní syntézy založené na metodě HMM jsou uvedeny v [43].

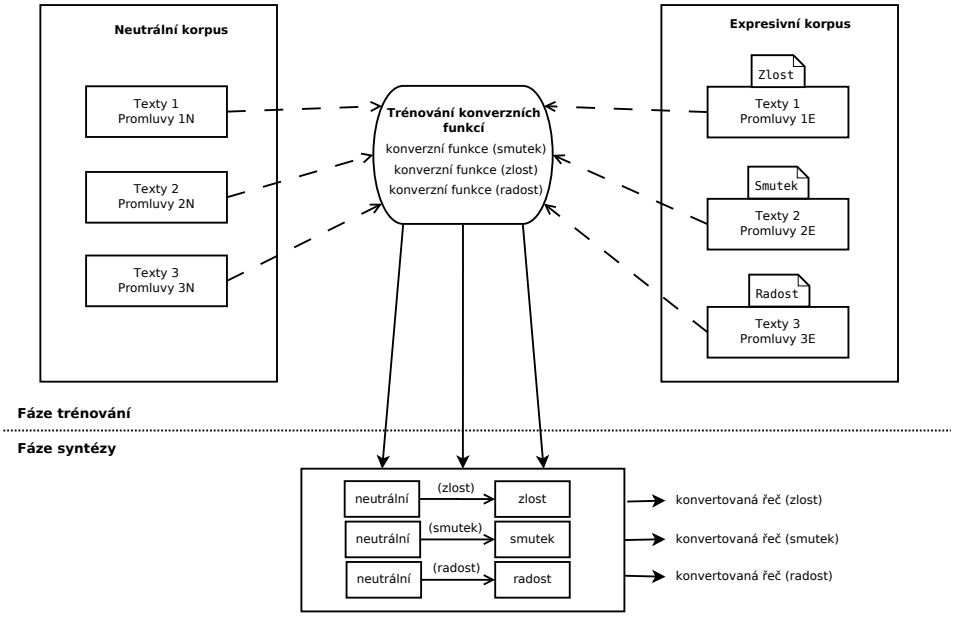
4.1.4 Konverze hlasu

Mimo výše uvedených metod syntézy řeči, které se používají také pro vytváření syntetické expresivní řeči, musíme uvést ještě jednu metodu. Ta však není samostatnou metodou syntézy řeči. Prvotním záměrem konverze hlasu, popsané např. v [62, 112], pro češtinu pak v [50, 51, 88], bylo modifikovat stávající systém syntézy řeči tak, aby produkoval řeč jiného řečníka, než pro který byl daný syntetizér původně vytvořen. Pro tento účel jsou popsány dva základní postupy:

- konverze původního řečového korpusu a následné vytvoření inventáře řečových jednotek (pro konkatenační syntézu) nebo natrénování HMM modelů (pro HMM syntézu) z takto nově vytvořeného korpusu;

- konverze výsledné syntetické řeči, tj. nejprve syntéza původním systémem TTS a poté transformace takto syntetizované promluvy pomocí konverzních funkcí.

Pokud je však možné konvertovat hlas jednoho řečníka na hlas jiného, vyvstává otázka, zda můžeme tuto konverzi aplikovat i na syntézu expresivní řeči. Tedy konvertovat buď a) neutrální řečový korpus na expresivní řečový korpus, nebo b) konvertovat neutrální syntetickou řeč pomocí konverzních funkcí do jiného, tentokráte expresivního, stylu. Metoda s využitím konverzních funkcí je zřejmě jako první popsána v [63], dále pak také v [126] nebo [118]. Systém produkce expresivní řeči tímto způsobem je pak znázorněn na obrázku 4.3.



Obrázek 4.3: Schéma konverze hlasu využívající metodu konverzních funkcí pro vytváření expresivní řeči.

4.2 Řečový korpus pro expresivní řeč

Pro syntézu neutrální řeči nejpoužívanějšími metodami (konkatenační syntézou, HMM syntézou) je zapotřebí, aby byl nejprve kvalitně nahrán a anoto-

ván řečový korpus. Rozsah takového korpusu pak záleží na použité metodě. Pro metodu dynamického výběru jednotek je zapotřebí, aby takový korpus byl rozsáhlý a obsahoval co nejvíce realizací jednotlivých řečových jednotek v různých kontextech a na různých pozicích ve větě, slově, slabice, apod. Pro konkatenační metodu s jednou realizací jednotky a HMM syntézu obvykle postačuje menší korpus.

Kvalitní řečový korpus je jednou z nejdůležitějších věcí pro budoucí syntézu. V případě syntézy expresivní řeči bychom měli postupovat podle toho, jakou metodu zvolíme – jeden korpus či více korpusů. V následujícím textu budou obecně popsány metody a principy nahrávání expresivního korpusu. Konkrétní postup nahrávání konkrétního korpusu pro naši úlohu, jeho zpracování, anotace a využití je pak popsáno v kapitole 7.

4.2.1 Metody nahrávání

Při použití konkatenační syntézy řeči s jedním zástupcem a syntézy HMM, kde dochází ke změně (modelování) prozodických charakteristik jednotlivých řečových jednotek i promluvy jako celku, je nejdůležitější vědět, jakým způsobem tyto změny provádět a jak se mají prozodické charakteristiky modelovat. Je tedy nutné provést důkladnou akustickou analýzu expresivní řeči. Za tímto účelem je potřeba nahrát expresivní korpus tak, aby promluvy v něm obsažené byly pokud možno co nejpřirozenější. Nahrávky by se tak měly pořizovat mimo laboratorní prostředí, během běžného života řečníka. Měly by zachytit jeho každodenní konverzaci a interakci s ostatními lidmi. Zároveň nesmí nahrávání řečníka jakkoliv ovlivňovat a obtěžovat. To vše by mělo vliv na jeho psychiku a tím samozřejmě také na přirozenost pořízeného korpusu. Zároveň by měly být nahrávky dostatečně kvalitní, aby je bylo možné analyzovat jak lingvisticky, tak prozodicky a spektrálně a zjistit potřebné akustické parametry.

Při použití metody výběru jednotek, kdy jsou jednotky pro syntézu vybírány dynamicky až při samotné syntéze, máme dvě možnosti jak sestavit inventář řečových jednotek[1]. V obou dvou případech je ovšem potřeba nahrát řečový korpus obsahující expresivní řeč, ať už má být tento korpus rozsáhlý, nebo podstatně menší. Protože tento korpus bude již přímo využit pro syntézu řeči, vysoká kvalita nahrávek je zde přímo nutností a má přednost před kvalitou vyjádřené expresivity. V úvahu zde tedy přichází pouze možnost nahrávání korpusu v nahrávacím studiu nebo podobném zařízení a především použití velmi kvalitních mikrofonů. Srovnání několika možností, jak zvolit

[1] Je to možnost více korpusů pro každou expresivní kategorii zvlášť, nebo jednoho korpusu obsahujícího promluvy s různými expresivními kategoriemi.

řečníka a jak zařídit, aby nahrávky zněly pokud možno co nejvíce přirozeně, přináší např. [15, 99, 75].

Co se týká akustických dat potřebných např. pro analýzu expresivní řeči nebo pro rozpoznávání expresivity/emocí v lidské řeči, představme zde pro ilustraci možnosti sběru takových dat pro japonštinu podrobněji popsané v [16].

První možností je, že řečník nosí po dlouhou dobu neustále přenosný mikrofon. Zaznamenává se tak jeho veškerá komunikace s okolím, včetně jakékoliv expresivní promluvy. Ta může být poté analyzována. Mikrofon musí být v tomto případě umístěn tak, aby řečníka co nejméně obtěžoval (pokud možno vůbec ne). Dále musí zaznamenávat pouze promluvu řečníka a nikoho jiného. Mohlo by zde jinak docházet k porušení práv jednotlivce, který by byl nahráván, aniž by o tom věděl. To jistě podmínky pro nahrávání ještě dále komplikuje.

Další možností je nahrávání telefonního rozhovoru, kdy je nahráván pouze jeden z účastníků. V tomto rozhovoru se předpokládá určité expresivní vyjadřování nahrávaného účastníka. Toto nahrávání může být provedeno za účelem zlepšení kvality i ve studiu.

Poslední variantou zmíněnou v [16] je umístění mikrofonu v domácnosti, nejlépe někde u stolu, kde se schází celá rodina. Nahrávky potom obsahují celé rozhovory členů domácnosti. Tyto nahrávky mají ovšem nejhorší kvalitu. To je způsobeno ruchem na pozadí a dále tím, že se jednotliví rodinní příslušníci mohou při konverzaci vzájemně překřikovat a jejich řeč se tak překrývá.

Podle [99] jsou postupy nahrávání dat určených přímo pro syntézu rozděleny do tří skupin:

přirozený hlasový projev Tento postup zahrnuje nahrávky pořízené např. v kokpitu letadla za obtížných situací [125], v televizním studiu při nějaké emotivní diskusi [40], při terapeutických sezeních [37] či při hraní počítačových her [60]. Problémem takovýchto nahrávek je pak horší kvalita, málo reprezentativních dat pro jednoho řečníka a téměř vůbec žádná kontrola nad vyjádřenými emocemi.

vyvolané emoce Přímým postupem, jak navodit u člověka afektivní stav, je použití psychoaktivních léků [52]. Další možností, jak vyvolat emoce, je dostat člověka do tíživé situace, dát mu obtížně řešitelný úkol nebo pomocí vizuální stimulace (promítání vhodných filmů nebo obrázků). Nevýhodou tohoto postupu je, že různé metody vyvolávání emocí mohou u různých lidí působit odlišně.

simulovaný expresivní hlasový projev Profesionální nebo i laický herec mají za úkol mluvit určitým expresivním stylem, k čemuž jim může

dopomoci například vhodný scénář. U tohoto postupu však hrozí nebezpečí, že herec bude expresivitu přehánět, což pak nebude znít přirozeně.

Vzhledem ke všem těmto (někdy i rozporuplným) požadavkům je těžké zvolit vhodné technické vybavení a metody, jak nahrání řečového korpusu provést. V minulosti bylo publikováno několik postupů nahrávání korpusu, který by měl sloužit ať už k akustické analýze nebo i následné syntéze expresivní řeči či naopak k rozpoznávání emocí nebo expresivních kategorií v mluvené řeči. Jako další příklady mimo již zmíněných uveďme alespoň [17, 129, 30, 106, 33, 95, 79, 7, 15].

Na následujících řádcích se tedy budeme snažit problematiku nahrávání řečových dat pro účely expresivní syntézy obecněji rozvést, a to hlavně otázku výběru řečníka a textů. Konkrétní výběr metody, postup nahrávání a zpracování expresivního korpusu v naší úloze je pak uveden v kapitole 7.

4.2.2 Řečník

Pro výběr řečníka samozřejmě existuje více variant. Když pomineme všeobecné požadavky, které musí splňovat (zejména hlas vhodný pro syntézu, srozumitelnost projevu, správná artikulace, apod.), je také nutné, aby měl předpoklady k expresivnímu vyjadřování. A to pokud možno přirozenou cestou.

Jednou z možností je výběr trénovaného herce, který je schopen potřebné expresivní vyjádření „zahrát“. Z různých divadelních a televizních představení by to pro něj teoreticky neměl být problém. Určitě bude schopen expresivní vyjadřování prezentovat přesvědčivě, ať bude text nahrávané promluvy jakýkoliv (o problémech výběru textu pojednává část 4.2.3). Jak to ovšem bude vypadat s přirozeností? Herec určitě bude mít zvláště v některých situacích snahu dané expresivní vyjádření přehánět. Nehledě na to, že základem přirozené expresivity jsou fyziologické změny v organismu (rychlejší nebo pomalejší dech, suché nebo mokré rty, atd.). Ty ovšem bude herec jen těžko simulovat. Expresivita tedy v jeho podání pravděpodobně nebude příliš přirozená, pokud budeme uvažovat standardní cestu nahrávání řečového korpusu pro syntézu. Možná by však stačilo, pokud by řečník budoucího posluchače dostatečně „oklamal“ a navodil u něho dojem, že pronášená řeč je skutečně expresivní. To je vlastně to, oč v zásadě jde – záleží totiž především na tom, jak danou situaci bude vnímat posluchač, především pro něj se expresivní řeč bude vytvářet.

Další variantou pro řečníka je výběr člověka–neherce. Zde ale vyvstává otázka, jak takového řečníka „přinutit“, aby své promluvě dodal expresivní

zabarvení, a aby se přitom zachovala přirozenost. Může to být vhodnou obrazovou nebo zvukovou stimulací, nebo například vhodným výběrem textů.

4.2.3 Texty

Ať už za řečníka bude vybrán herec nebo ne, je nutné ho nějakým způsobem „uvést“ do stavu, kdy bude přirozeně prezentovat svůj expresivní projev. Předpokládáme, že u řečníka-neherce je tento faktor mnohem důležitější než u řečníka-herce. Několik příkladů je uvedeno v [15]. Je zde popsáno jak výše zmíněné „hraní“, tak stimulace pomocí expresivně zabarveného textu, nebo simulace nějaké reálné situace prostřednictvím předem připraveného scénáře, který má za úkol řečníka tzv. „naladit na tu správnou vlnu“.

V případě stimulace pomocí expresivně zabarveného textu je volba obsahu tohoto textu jasná. Pro každou expresivní kategorii (resp. dialogový akt, pokud se budeme pohybovat v oblasti diskrétního rozdělení expresivity podle části 3.2) je vybrán text, který ji nějakým způsobem reprezentuje. Pro dialogové akty vyjadřující smutek to bude text připomínající nějakou smutnou událost, pro radostné např. text oznamující výhru v loterii, apod.

Pokud bude jako řečník určen herec, který bude expresivní řeč simulovat, je možné vybrat jako obsah textu prakticky cokoliv. Bylo by tedy dobré volit obsah pro všechny expresivní kategorie stejný, tedy expresivně neutrální. Další možností je výběr obsahově nesmyslných vět, ale tento přístup není pro obtížnou představivost a zapamatovatelnost příliš doporučován [12]. Pokud bychom mohli obsah zvolit, bylo by možné nejen objektivně porovnávat výstupní syntetickou řeč, ale také zvolit text vyvážený co do obsahu různých řečových jednotek. Tento postup má jistě svoje výhody a je zcela na řečníkovi, jak bude dané expresivní stavy reprezentovat.

Použití simulace nějaké reálné situace, tedy např. scénářů, je časově mnohem více náročnější než dva výše zmíněné postupy (použití expresivně zabarvených textů a použití obecných textů). A to jak jejich příprava, tak posléze i samotné nahrávání. Ve fázi přípravy jde o správný výběr textů, řazení jednotlivých promluv a vhodné zařazení té části scénáře, která má být začleněna do řečového korpusu. Ostatní části (především úvodní část) totiž do korpusu zařazeny být ani nemusí, mohou sloužit jen k „naladění“ řečníka. Ve fázi nahrávání pak časová náročnost souvisí s tím, že do korpusu je zařazena pouze část nahraných promluv, nepoužité promluvy se tedy nahrávají nadbytečně. K získání dostatečného množství použitelných dat je tak potřeba ve skutečnosti nahrávat mnohem déle.

Jako příklad výběru řečníka a textů uvedeme postup nahrávání korpusu podle [12]. Zde jsou jako řečníci vybráni lidé prostřednictvím inzerátu v novinách. Ti posléze projdou jakýmsi výběrem, kde o jejich vhodnosti rozhodují

zkušení fonetici na základě několika zkušebních vět. Zřejmě není náhodou, že drtivá většina vybraných řečníků v minulosti prošla „hereckou školou“, nicméně nejedná se o herce. Řečníkům jsou předloženy texty s neutrálním obsahem a jejich úkolem je zvolit vhodnou reprezentaci. Nutno podotknout, že korpus uvedený v této práci byl nahrán za účelem analýzy expresivní řeči.

Podrobnější popis návrhu řečového korpusu včetně všech faktů, které je třeba vzít v úvahu, uvádí např. [28]. Je zde uveden i seznam některých databází expresivní řeči, včetně postupů jakými byly získány.

4.2.4 Velikost korpusu

Pro kvalitní konkatenační syntézu řeči z textu použitím metody výběru jednotek je rozsáhlý korpus nezbytný. Měl by být vyvážený z hlediska různých řečových jednotek tak, aby se každá řečová jednotka v korpusu objevila nejméně jednou (nejlépe však samozřejmě vícekrát). Každá jednotka by se tam měla vyskytovat v různých kontextech, na různých pozicích ve frázích, slovech, slabikách a s různými prozodickými charakteristikami. Pokud by měl korpus obsahovat všechny jednotky se všemi možnými kontexty a prozodickými charakteristikami, byl by velmi obsáhlý a je otázkou, zda by vůbec bylo možné ho vytvořit. Proto je velmi důležitý správný kompromis mezi velikostí a obsahem. Obzvláště pokud by musel korpus existovat ve více verzích pro různé expresivní kategorie, viz část 4.1.2. I z tohoto důvodu je obecná syntéza expresivní řeči z textu velmi obtížnou úlohou.

Pro potřeby dialogového systému, který je popsán v kapitole 6, bude navržení korpusu poněkud jednodušší. Neutrální část korpusu bude shodná s korpusem používaným pro běžnou syntézu neutrální řeči. Expresivní část obsahující věty označené pomocí dialogových aktů popsaných v části 3.2 může být menší. Toto zjednodušení si můžeme dovolit, protože tento připravovaný korpus bude použit pouze pro konkrétní účel konkrétního dialogového systému. Důležitý je však samotný postup návrhu, nahrávání a zpracování expresivního korpusu, který je pak možno využít i v jiných situacích, resp. jiných konkrétních aplikacích, pro které bude plánováno využití syntézy expresivní řeči. Postup pro získání textů k nahrávání, tvorba scénářů, proces nahrávání a dalšího zpracování expresivního korpusu by totiž mohl být obdobný.

Kapitola 5

Cíle práce

V předchozích kapitolách jsme se věnovali především popisu metod používaných pro syntézu řeči, a to jak neutrální, tak expresivní. Představili jsme některé přístupy, které se používají pro popis expresivity, a uvedli jsme možné metody pro nahrávání řečových korpusů. V další části práce se budeme věnovat vlastnímu návrhu rozšíření metody dynamického výběru jednotek pro syntézu řeči tak, abychom byli schopni produkovat syntetickou expresivní řeč pro dialogový systém popsaný v části 6.1.
Mezi hlavní cíle této práce patří:

1. Návrh systému syntézy expresivní řeči v dialogu.
2. Získání reálných dat z dialogů, které budou nahrány za podobných podmínek, ve kterých by měl navržený systém posléze fungovat.
3. Návrh popisu expresivity, tedy definování vlastní množiny dialogových aktů na základě postupů, které se objevují v odborné literatuře.
4. Příprava a nahrávání expresivního korpusu, což zahrnuje především přípravu textů a scénářů, vývoj aplikace pro nahrávání, výběr řečníka a samotný proces nahrávání.
5. Zpracování nahraného expresivního korpusu, tedy zejména jeho anotace s využitím definovaných dialogových aktů za pomoci poslechových testů a akustická analýza expresivních řečových dat.
6. Definice penalizační matice, která bude reprezentovat rozdíly mezi jednotlivými expresivními kategoriemi (dialogovými akty) a bude dále využita při modifikaci výpočtu ceny cíle.
7. Modifikace metody dynamického výběru řečových jednotek:

   (a) úprava hodnotící funkce pro výpočet ceny cíle s využitím dialogových aktů a penalizační matice;

   (b) nastavení vah pro jednotlivé příznaky ovlivňující cenu cíle.

8. Vyhodnocení výsledků navrženého postupu pomocí poslechových testů. Hodnocení bude zahrnovat jak kvalitu syntetizované expresivní řeči, tak i vhodnost a přiměřenost expresivního vyjádření v dialogu. Zejména nás bude zajímat porovnání se systémem produkujícím neutrální řeč.

Syntetickou neutrální řečí (nebo neutrální syntézou) budeme v této práci označovat řeč produkovanou stávajícím systémem TTS ARTIC. To ve skutečnosti znamená syntézu „zpravodajského“ stylu, který slouží především k předávání informací. Nemá žádné ambice přenášet na posluchače nějakou expresivitu či emoce a budeme ho tak označovat za expresivně neutrální.

Kapitola 6

Návrh systému syntézy expresivní řeči v dialogu

Obecná syntéza expresivní řeči je velmi komplikovaná, a tak se současný výzkum zaměřuje spíše na konkrétnější použití hlasového výstupu. Alespoň pro představu uveďme několik zahraničních příkladů: komentáře fotbalových utkání [66], vývoj vojenských výcvikových aplikací [60] nebo poskytování turistických informací [129]. Z výše uvedeného důvodu je i náš výzkum zaměřen především na expresivní syntézu v omezené oblasti a to pro potřeby dialogového systému, který je popsán v části 6.1. V dialogovém systému, který je určen pro přirozenou hlasovou interakci člověka s počítačem je však více zásadních otázek než jen expresivní syntéza řeči (viz např. [84]). Takový systém by měl být schopen řešit následující problémy:

- rozpoznání expresivního/emotivního stavu lidského uživatele (tedy nejen rozpoznání promluvy jako takové, kterým se zabývá běžné automatické rozpoznávání řeči) – např. rozpoznat, zda je uživatel naštvaný, nebo třeba pospíchá a chce se rychle dozvědět informace;
- využít tuto informaci pro další plánování dialogu (tedy kromě určení cíle dialogu také určit, jak tohoto cíle dosáhnout) – pokud třeba uživatel pospíchá, volit krátké a co nejkonkrétnější fráze, aby se dosažení cíle co nejvíce urychlilo;
- vygenerovat vhodnou odpověď v závislosti na fázi dialogu a stavu uživatele (tj. kromě samotné obsahově významové informace volit i vhodnou slovní reprezentaci) – pokud je uživatel naštvaný, volit třeba slova omluvy;
- pomocí metod expresivní syntézy pak tuto odpověď prezentovat uživateli vhodným způsobem – pro našeho ukázkového pospíchajícího uži-

vatele třeba zrychlit syntézu, pro naštvaného volit také omluvný tón hlasu (nejen omluvná slova).

Existuje tedy také více možností, jak dát pomocí řeči najevo expresivitu, viz např. [84], kde se uvádí tři konkrétní faktory, které ovlivňují vnímání expresivity posluchačem a které by při vytváření expresivní syntézy měly být vzaty do úvahy. Prvním z těchto faktorů je verbální informace, tedy textový obsah promluvy nezávisle na charakteristice řeči. V [84] bylo dosaženo úspěšnosti 55 % v rozpoznání dané emoce lidským posluchačem pouze na základě textu. Přitom v té samé práci byla úspěšnost rozpoznání na základě mluvené řeči 70 %. Mluvenou řeč pak ovlivňují i dva zbývající faktory: prozodické charakteristiky (základní hlasivková frekvence, intenzita, apod.) a spektrální charakteristiky (např. kvalita hlasu). Dodejme, že v [84] se pracovalo s korpusem nahraných televizních filmů a diskrétními kategoriemi expresivních stavů (emocí), konkrétně strach, zlost a neutrální stav.

Dalšími faktory, které mohou ovlivňovat schopnost vnímání expresivity posluchačem, může být také například expresivní stav samotného posluchače, jeho sociální či kulturní postavení nebo psychický stav. Z výše uvedeného je zřejmé, že ani omezení syntézy expresivní řeči na konkrétní dialogový systém nemusí být dostatečně zjednodušující.

6.1 Dialogový systém

Náš další postup byl motivován vývojem dialogového systému, který by měl být zaměřen na rozhovor starších lidí (seniorů) s počítačem. Konkrétně se jedná o rozhovor o fotografiích ze života těchto seniorů. Dialogový systém by měl fungovat ve dvou jazykových verzích, české a anglické. Na našem pracovišti byla vyvíjena část automatického rozpoznávání řeči pro českou i anglickou verzi a část syntézy řeči pro českou verzi. Cílem bylo vytvořit virtuálního společníka (tzv. *Companion*, `http://www.companions-project.org`), který bude se seniory plynule hovořit o jejich fotografiích. Tento systém by v konečné fázi měl mít následující funkce:

rozpoznat spontánní řeč – Systém by měl rozumět spontánní řeči, když na něj uživatel bude hovořit. To zahrnuje využití automatického rozpoznání řeči (ASR, z anglického *Automatic Speech Recognition*, nebo také převod řeči na text) na základě metod, které jsou schopné rozpoznávat

mluvenou řeč, v našem případě nejlépe i bez adaptace na řečníka[1]. Protože spontánní řeč není zpravidla gramaticky správně a obsahuje různé „defekty“, je dalším krokem rekonstrukce řeči (*speech reconstruction*). Ta by měla výstup ASR transformovat tak, aby obsahoval gramaticky správný užitečný text (tj. například odstranění různých slovních výplní, opakování a oprav). Pak lze totiž použít běžné nástroje NLP (z anglického *Natural Language Processing*, tj. zpracování přirozeného jazyka) pro další zpracování.

porozumět tomu, co člověk říká – Porozumění mluvené řeči (NLU, z anglického *Natural Language Understanding*) je velmi obtížná a komplexní úloha. Zahrnuje mimo jiné syntaktickou a sémantickou analýzu textu, vytváření tzv. syntaktických stromů a anotace v tektogramatické rovině [46].

schopnost určit, co se na dané fotografii vyskytuje – Pomocí metod digitálního zpracování obrazu (DIP, z anglického *Digital Image Processing*) rozpoznat objekty a osoby vyskytující se na fotografiích. Systém by pak mohl sám vytvářet dotazy vztahující se k identifikovaným objektům, aniž by o nich musel uživatel předtím hovořit.

určovat tok dialogu – Jednou z hlavních funkcí systému by mělo být udržování dialogu a stimulace člověka k dalšímu vyprávění o fotografii. Tuto funkci zajišťuje tzv. správce dialogu (DM, z anglického *Dialogue Manager*). Jeho úkolem je na základě výstupu modulu NLU (případně modulu DIP) generovat vhodné reakce systému prostřednictvím modulu NLG (z anglického *Natural Language Generation*). Tedy například vytvářet rozumné otázky, které by blíže určovaly objekty zobrazené na fotografii a události, jež se k ní vztahují.

produkovat syntetickou řeč – Výstupem systému by měla být pokud možno co nejpřirozenější řeč vytvářená modulem TTS (z anglického *Text-To-Speech*), který navazuje na modul NLG. Řeč by měla být v závislosti na obsahu dialogu také expresivní a vyjadřovat tak „pocity“ či „nálady“ systému a přenášet je na uživatele.

modelovat virtuálního avatara – Virtuálního společníka by měl navenek reprezentovat 3D avatar. Tento úkol zahrnuje vytvoření virtuálního 3D avatara, který by měl pracovat v součinnosti s modulem TTS

[1]Pro zvýšení přesnosti rozpoznávání řeči byl však v rámci projektu vyvinut i takový systém adaptace na řečníka, který je uživateli skryt a nevyžaduje od něj žádnou explicitní interakci. Adaptace je provedena na základě několika prvních interakcí uživatele s počítačem v okamžiku, kdy je těchto interakcí dostatečné množství.

a ve skutečnosti tak zajišťovat audiovizuální syntézu řeči [120]. Důležitá je tedy synchronizace syntetické řeči s pohyby navenek viditelných artikulačních orgánů. Pro větší přirozenost virtuálního avatara je navíc zapotřebí zajistit simulaci různých gest či obličejových grimas.

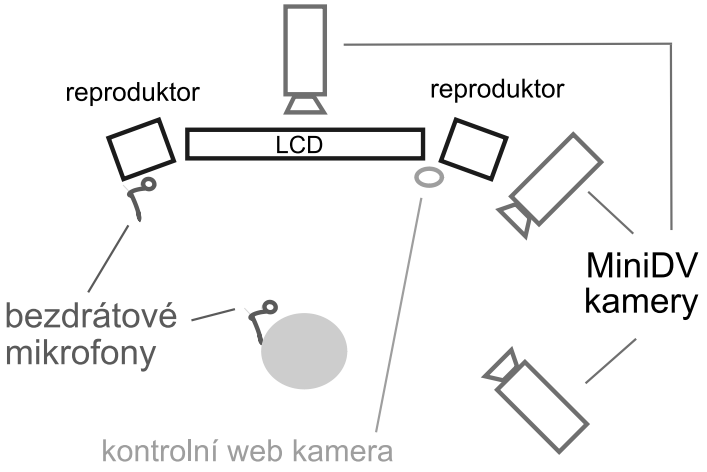
získat přehled o dané osobě – Systém by měl z dialogů získat celkový přehled o uživateli, např.: kdo je, jak se jmenuje, kde a kdy pracoval či pracuje, jaké jsou v jeho rodině příbuzenské vazby, jeho záliby a koníčky, apod.

uchovávat informace – Všechny informace získané během dialogů by měly být v systému uchovány pro pozdější použití (v rámci dalších dialogů). Systém by si tak měl „pamatovat“ např. rodinné vztahy, jména osob, se kterými se uživatel zná, města a místa, která uživatel navštívil nebo o nich hovořil, apod. To by mělo (společně s předchozím bodem) zvyšovat „inteligenci“ systému, aby byl schopný generovat vhodnější (příp. konkrétnější) reakce.

Naším úkolem bylo zajistit v tomto ohledu kvalitní a co nejpřirozenější syntézu řeči pro tento dialogový systém. Není tedy zapotřebí mít možnost syntetizovat nějaké obrovské množství expresivních promluv v neomezené oblasti, ale pouze jistá expresivní vyjádření ve velmi specifických rozhovorech, jejichž průběh lze částečně odhadnout a v podstatě také řídit. V těchto rozhovorech se tedy nejedná o obecnou syntézu expresivní řeči, ale o konkrétní využití dialogových aktů (komunikačních funkcí) popsaných v části 3.2.

6.2 Získání reálných dat

Abychom vůbec mohli vytvořit syntézu řeči v omezené oblasti, musíme tuto oblast nejprve znát. Pro tyto účely byl nahrán rozsáhlý audiovizuální korpus (vizuální část korpusu byla zachována pro případné pozdější využití). Během nahrávání bylo zaznamenáno 65 rozhovorů seniorů s počítačem, a to za použití metody *Wizard of Oz* [122], která bude popsána dále. Rozhovor tedy probíhá stejně jako kdyby si se seniorem povídal člověk za normálních okolností, např. u něj doma. Tato metoda má oproti přístupu běžně nahrávaného rozhovoru člověk–člověk tu výhodu, že nahrávaná osoba se v nahrávací místnosti cítí sama a může bez zábran hovořit o všem, o čem chce. Navíc výsledný dialogový systém bude také „pouze počítačový“, bude se tedy jednat o téměř shodnou situaci jako při nahrávání. Tím se chceme vyvarovat jakékoliv nepřirozenosti, která by se jinak mohla v rozhovoru objevit.



Obrázek 6.1: Schéma nahrávací místnosti.

Schéma nahrávací místnosti je zobrazeno na obrázku 6.1, snímek zachycující seniora (subjekt) při nahrávání je pak pro představu celé situace vyobrazen na obrázku 6.2.

Řečový signál z rozhovorů byl zaznamenáván dvěma bezdrátovými mikrofony (jeden pro nahrávaný subjekt, jeden pro produkovanou syntetizovanou řeč – pro usnadnění přípravných prací pro pozdější anotace, které jsou popsány v části 7.2). Abychom získali velmi kvalitní akustický materiál, byl použit externí zvukový zesilovač a externí zvuková karta Creative Sound Blaster Extigy. Všechny dialogy byly nahrány za použití vzorkovací frekvence 22kHz a se 16 bitovým rozlišením.

Video z rozhovorů bylo snímáno třemi miniDV videokamerami (čelní, boční a zadní pohled, jak je vidět na obrázku 6.2). Čelní pohled by mohl v budoucnu sloužit pro audio-vizuální rozpoznávání řeči, kdy by bylo možné využít tento úhel pohledu pro sledování pohybu rtů, a nebo také pro rozpoznávání expresivity u subjektu. Tento pohled, společně s bočním pohledem může být také využit pro pozdější 3D modelování hlavy. Protože boční pohled nezachycoval pouze hlavu subjektu, ale i celou horní část těla, lze pro něj v budoucnu jistě najít využití i pro rozpoznávání gest rukou, popř. celého těla. Zadní pohled pak zachycoval kromě subjektu i obrazovku, která sloužila pro interakci systému s uživatelem. Zde bychom pak mohli identifikovat situace, kdy subjekt např. ukazuje na určité místo na obrazovce prstem.

Technika Wizard of Oz (WoZ), která byla použita pro získání reálných dialogových dat, zjednodušeně znamená, že nahrávaný subjekt (v našem pří-

Obrázek 6.2: Snímek zachycující subjekt při nahrávání reálného dialogu; pohled ze všech tří kamer.

padě senior) se domnívá, že komunikuje se strojem ovládaným umělou inteligencí, který je reprezentován počítačem. Ve skutečnosti se za počítačem skrývá „wizard“ (operátor), tedy člověk (v našem případě dva lidé), který se subjektem komunikuje a vede s ním rozhovor prostřednictvím speciální a přesně pro tento účel vyvinuté aplikace. Ta se sestává ze dvou hlavních modulů, tzv. presenter a wizard.

První z těchto modulů, nazývaný *presenter*, je znázorněn na obrázku 6.3. Je určen pro interakci nahrávaného subjektu s počítačem, využívá stávající systém syntézy neutrální řeči s možností přehrání předdefinovaných neřečových událostí a 3D model mluvící hlavy [120]. Tzv. „avatar“ tedy představuje jinak abstraktní umělou inteligenci počítače schopnou autonomní komunikace s člověkem. Nahrávanému subjektu postupně předkládá a zobrazuje fotografie z jeho života, ptá se ho na doplňující otázky, povzbuzuje ho k vyprávění

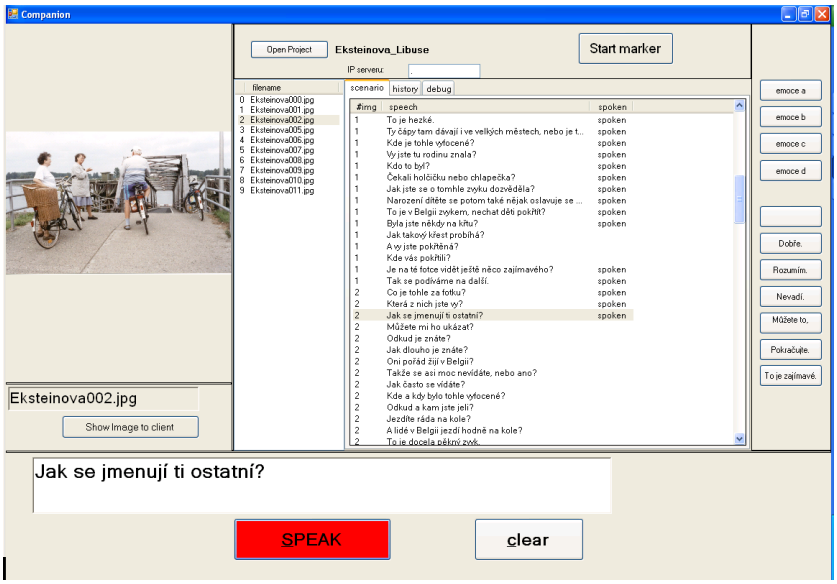
o lidech a událostech, které se k dané fotografii vztahují. Celkově se tedy snaží držet běh rozhovoru v nějaké rovině a také jej dále rozvíjet a posouvat. Jak je vidět z obrázku 6.3, na obrazovce byly zobrazeny i titulky k promluvě právě pronášené avatarem, které měly odstranit případnou nesrozumitelnost syntetizované řeči. Tato součást modulu byla po několika rozhovorech odstraněna, protože nebyly zaznamenány žádné případy, kdy by nebylo avatarovi z takového důvodu rozumět.

Obrázek 6.3: Aplikaci pro nahrávání metodou WoZ – presenter.

Na obrázku 6.4 je znázorněn druhý modul této aplikace, tzv. *wizard*, který poskytoval lidským operátorům různé možnosti jak komunikovat s nahrávaným subjektem. V levé části rozhraní je vidět fotografie, o které se právě diskutuje a která je zároveň zobrazena i v prvním modulu *presenter* a tedy na obrazovce sledované subjektem. V prostřední části je pak předem připravený scénář (viz dále) skládající se z přibližně 10 vět na jednu fotografii, ze kterých mohou operátoři snadno vybrat tu nejvhodnější větu pro právě probíhající část rozhovoru. Pokud bylo potřeba předem připravenou větu jakkoliv upravit, případně zareagovat mimo scénář, mohlo být využito textové pole zobrazené v dolní části rozhraní. Zde podotkněme, že této možnosti bylo během nahrávání hojně využíváno. Pravá část rozhraní poskytuje možnost vložit do rozhovoru buď neřečovou událost[2] nebo krátkou reakci na právě

[2]Pro účely nahrávání byly vybrány 4 neřečové události, u kterých předpokládáme časté využití v takovém dialogu – „aha“, přitakávací „ehm“, váhavé „mmm“ a úsměv. Všechny byly samozřejmě doprovázené korespondujícím vizuálním gestem avatara.

pronesenou větu subjektu. Operátoři pak měli možnost libovolně přepínat mezi fotografiemi a přestože bylo dodržováno předem stanovené pořadí fotografií, někdy byla využita i tato možnost – to v případě, kdy se subjekt během rozhovoru o nějaké fotografii zmínil o nějaké předcházející, popř. se k ní chtěl sám vrátit.



Obrázek 6.4: Aplikaci pro nahrávání metodou WoZ – wizard.

Důležitým a časově velmi náročným úkolem operátorů bylo vypůjčit si od seniorů, kteří se přihlásili k nahrávání, jejich rodinné fotografie, o kterých si tito lidé přáli hovořit, a zjistit a vyzvědět od nich co nejvíce relevantních informací vztahujících se ke každé fotografii. Na základě takto získaných informací pak bylo potřeba sestavit pravděpodobný scénář, kterým se rozhovor bude později ubírat. Ukázka takového scénáře je zobrazena na obrázku 6.5.

6.3 Statistika nahraných reálných dat

Celkem bylo tedy metodou WoZ nahráno 65 rozhovorů. Z toho bylo 37 žen a 28 mužů. Průměrná doba trvání jednoho rozhovoru je 56 minut (byly jsme omezeni technickým vybavením – maximální možnou délkou záznamu 1 hodiny na videopásku), za tu dobu se průměrně hovořilo o 8 fotografiích (maximum bylo 12, minimum 3). Průměrný věk činil 69 let a byl přibližně stejný jak pro muže, tak pro ženy (maximum bylo 86, minimum 54 let).

Bezprostředně po každém nahrávání byl se seniorem vždy vyplněn krátký dotazník týkající se jeho osobních údajů a názorů na právě proběhlý roz-

- Kde to bylo vyfocené?
- To bylo o prázdninách?
- Jezdili jste tam pravidelně?
- Ve kterém roce to tak mohlo být?
- Kdo je na téhle fotce vidět?
- Vaše dcera?
- Jak se jmenuje?
- Pamatujete si ještě i jméno toho jejího kamaráda?
- Kdo z nich je Veronika? *(dcera)*
- Kolik jí tenkrát mohlo být let?
- Líbilo se jí na horách? (už víme, že fotografie byla pořízena na horách)
- Také jste tam lyžovali?
- Na běžkách nebo na sjezdovkách?
- To muselo být nádherné.
- Určitě na to máte krásné vzpomínky.

Obrázek 6.5: Scénář připravený pro reálný dialog sestavený na základě informací od seniorů.

hovor. Z výsledků vyhodnocení dotazníků vyplynulo, že naprostá většina seniorů byla z Plzeňského kraje, několik jednotlivců pocházelo z Moravského kraje (moravské nářečí se od spisovné češtiny liší jak ve výslovnosti některých slov tak i přímo v použité slovní zásobě). Zhruba polovina nahrávaných seniorů vlastnila počítač, ale jen malá část z nich ho často využívala. Rozhovor byl seniory hodnocen jako přátelský a plynulý, syntetizovaná řeč pak jako srozumitelná a příjemná. Potěšující zprávou pak pro nás bylo, že většina dotázaných byla schopna představit si využití takového systému v domácnosti a někteří by ho i uvítali.

Získali jsme tedy více než 60 hodin řečových dat, ale především znalost o tom, co je obsahem rozhovorů a jakým směrem se takový rozhovor nad osobními fotografiemi ze života seniorů ubírá. Část přepisu jednoho z rozhovorů je pro ilustraci uveden v příloze F v tabulce F.1. Diskutovaná témata nám poskytují omezenou oblast, ve které bychom se měli při pozdější syntéze pohybovat. Je jasné, že řada rozhovorů se může, a zcela jistě bude, ubírat i jiným směrem a témata mohou být různá. Nicméně by mělo být zajištěno pokrytí těch nejdůležitějších frází v různých fázích dialogu. Mluvčím by měl být v tomto případě především člověk. Počítač je pouze posluchačem, který má dávat najevo, že mluvčího vnímá, případně ho stimulovat dalšími otázkami k povídání a k získání více informací. Je samozřejmě jasné, že syntéza

musí umět syntetizovat i všechny ostatní promluvy. Musí to být tedy obecný syntetizér, ve kterém však bude použito co nejvíce expresivně zabarvených promluv z řešené oblasti, jejichž součástí jsou samozřejmě i emotivní vyjádření.

K dosažení tohoto cíle by bylo zřejmě možné také nahrazovat určité promluvy, které jsou požadované k syntetizování. A to zejména takové, které se v našem korpusu vůbec nevyskytují, popř. by nějaká objektivní míra byla schopna určit, že jejich syntéza by byla nesrozumitelná či nekvalitní. Jako náhradní promluvy by pak byly vybrány takové, které v expresivním korpusu obsaženy jsou, popř. jejich syntéza by byla přirozenější a budou těm požadovaným velmi blízké. Musí se samozřejmě za všech okolností zachovat význam požadované promluvy. Tohoto prostředku by se však mělo využívat co možná nejméně a v našem systému s takovým mechanismem prozatím ani nepočítáme.

6.4 Použité dialogové akty

Na základě získané rozsáhlé audiovizuální databáze reálných rozhovorů člověka s počítačem (část 6.2) a znalostí získaných studiem různých schémat popisu dialogových aktů v dialogovém systému (část 3.2) jsme navrhli vlastní množinu dialogových aktů. V tabulce 6.1 je stručně uveden přehled těchto dialogových aktů a jejich symboly, jak jsou použity dále v této práci. Uvádíme i příklady, jak se tyto dialogové akty mohou vyskytnout v reálných dialozích. Tato množina dialogových aktů pak bude sloužit pro popis expresivity v expresivním korpusu určeném pro vlastní expresivní syntézu (kapitola 7).

Nyní představíme popis jednotlivých dialogových aktů, co mají vyjadřovat a v jaké podobě by se případně mohly objevit ve schématu dialogových aktů SWBD-DAMSL (uvedeno v hranatých závorkách). Tohoto „mapování“ by šlo v budoucnu případně využít v nějakém rozšíření množiny dialogových aktů i na jiné oblasti dialogů, popř. pokud by došlo k nějaké standardizaci dialogových aktů. Ne u všech námi navržených dialogových aktů je však toto mapování možné, a to ze dvou důvodů. Buď je námi navržený dialogový akt natolik specifický pro danou úlohu, že jeho protějšek ve schématu SWBD-DAMSL neexistuje (pak by zřejmě muselo dojít k přeformulování, případně k zařazení do nějaké nejbližší kategorie), nebo je jeho význam rozsáhlejší a může reprezentovat více různých kategorií (pak by muselo dojít k jeho nové anotaci s rozšířenou množinou dialogových aktů).

- pokyn (*DIRECTIVE*) – vyjadřuje pokyn směrem k druhému partnerovi v dialogu, aby pokračoval ve vyprávění; má blízko k *ENCOURAGE*, ale nemívá formu otázky; [dopředná funkce – *Action-directive*];

Tabulka 6.1: Množina použitých dialogových aktů.

dialogový akt	*symbol*	*příklad*
pokyn	*DIRECTIVE*	Řekněte mi to. Povídejte.
žádost	*REQUEST*	Vraťme se k tomu později.
vyčkávání	*WAIT*	Počkejte chvíli. Ještě moment.
omluva	*APOLOGY*	Promiňte. Omlouvám se.
pozdrav přivítání	*GREETING*	Ahoj. Dobrý den.
rozloučení	*GOODBYE*	Na shledanou. Ahoj.
poděkování	*THANKS*	Děkuji. Díky.
překvapení	*SURPRISE*	Opravdu máte 10 sourozenců?
empatie - smutek	*SAD-EMPATHY*	To je smutné. To slyším nerada. To je hrozné.
empatie - radost	*HAPPY-EMPATHY*	To je hezké. Výborně. To muselo být úžasné.
projevení zájmu	*SHOW-INTEREST*	Můžete mi o tom říct víc?
souhlas přitakání pochopení	*CONFIRM*	Ano. Aha. Rozumím. Dobře. Hmm.
nesouhlas nepochopení	*DISCONFIRM*	Ne. Tomu nerozumím. To nechápu.
povzbuzení pobídka	*ENCOURAGE*	Dobře. Například? A co vy?
nespecifikovaný	*NOT-SPECIFIED*	Slyšíte mě dobře? Jmenuji se Pavla.
jiný	OTHER	

- žádost (*REQUEST*) – vyjadřuje žádost, např. změnu tématu, posun v dialogu; [dopředná funkce];

- vyčkávání (*WAIT*) – reprezentuje žádost o chvíli strpení; má blízko k *REQUEST*;

- omluva (*APOLOGY*) – vyjadřuje omluvu za nějaký způsobený problém nebo špatné pochopení; [dopředná funkce – *Apology*];

- pozdrav (*GREETING*) – znamená zahájení rozhovoru, přivítání partnera v dialogu; [dopředná funkce – *Conventional-opening*];

- rozloučení (*GOODBYE*) – znamená ukončení rozhovoru; [dopředná funkce – *Conventional-closing*];
- poděkování (*THANKS*) – vyjadřuje poděkování, třeba za konverzaci nebo polichocení; [dopředná funkce – *Thanking*];
- překvapení (*SURPRISE*) – vyjadřuje překvapení nad některou právě oznámenou skutečností; [zpětná funkce];
- empatie - smutek (*SAD-EMPATHY*) – reprezentuje vcítění se do nějaké smutné situace, soucítění s partnerem v dialogu; [dopředná funkce – *Statement-opinion*];
- empatie - radost (*HAPPY-EMPATHY*) – reprezentuje vcítění se do nějaké radostné situace, sdílení radosti s partnerem v dialogu; [dopředná funkce – *Statement-opinion*];
- projevení zájmu (*SHOW-INTEREST*) – má přimět partnera v dialogu k podrobnějšímu vyprávění či popisu nějakého stavu či situace; má blízko k *ENCOURAGE*, ale většinou je ve formě otázky ano/ne přičemž se takováto stručná odpověď ve skutečnosti neočekává; [dopředná funkce – *Yes-No-Question*];
- souhlas / přitakání / pochopení (*CONFIRM*) – vyjadřuje souhlas s řečeným, pochopení s vysvětlenou situací nebo porozumění nějaké problematice; [zpětná funkce – *Agree/Accept*, *Response Acknowledgement*, *Acknowledge*, *Yes answers*, *Affirmative non-yes answers*];
- nesouhlas / nepochopení (*DISCONFIRM*) – vyjadřuje nesouhlas nebo nepochopení dané situace či problematiky; [zpětná funkce – S*ignal-non-understanding*, *No answers*, *Negative non-no answers*, *Other answers*];
- povzbuzení / pobídka (*ENCOURAGE*) – má pobízet či povzbuzovat k rozsáhlejší diskuzi o daném tématu; má blízko k *SHOW-INTEREST*, ale nemívá formu ano/ne otázky; [dopředná funkce];
- nespecifikovaný (*NOT-SPECIFIED*) – měl by označovat takové promluvy, které žádný dialogový akt nevyjadřují; z hlediska expresivity by měl mít stejnou funkci jako neutrální promluva [dopředná funkce – *Statement-non-opinion*];
- jiný (*OTHER*) – reprezentuje jakýkoliv jiný dialogový akt, který by se mohl vyskytnout, avšak nezapadá do žádné z předchozích kategorií.

Jak je vidět z uvedeného popisu dialogových aktů, většina z nich plní dopřednou funkci, tedy smysl takovýchto promluv spočívá v dalším rozvíjení dialogu. Mají přimět partnera v dialogu (uživatele, seniora) aby se více rozpovídal. Pokud uživatel sám o sobě vypráví a povídá, pak přichází na řadu dialogové akty zařazené v kategorii zpětné funkce, které by měly dát najevo, že systém uživateli naslouchá a vnímá ho.

Přestože schémata navržená pro popis dialogových aktů mají původně sloužit především pro popis různých fází dialogu, my předpokládáme, že v různých fázích dialogu může řečník určitým způsobem prezentovat svůj postoj k dané situaci či právě probíranému tématu nebo vyjadřovat své potřeby týkající se průběhu dialogu. Věříme tomu, že námi navržené dialogové akty budou sloužit nejenom k popisu různých fází dialogu, ale že budou prezentovat i postoj mluvčího a vyjadřovat jeho afektivní stav prostřednictvím expresivně zabarvené řeči. Použitím těchto dialogových aktů by se tak měla syntéza řeči v dialogu stát přirozenější pro posluchače, tedy partnera v dialogu, v našem případě seniora.

Kapitola 7

Vývoj korpusu pro syntézu expresivní řeči

Protože pro syntézu řeči bude v dialogovém systému použita metoda výběru jednotek popsaná v části 2.3.3, bylo potřeba vytvořit nový korpus obsahující expresivně vyjádřené promluvy. Ten poté může být sloučen se stávajícím neutrálním korpusem, nebo pouze vhodně doplněn určitými promluvami z neutrálního korpusu tak, abychom docílili úplného pokrytí všech možných jednotek, které se mohou v syntetizovaném textu vyskytnout. Jedná se o tzv. fonetické vyvážení, pro které lze použít například metody uvedené v [70]. Nově nahraný korpus musel být také samozřejmě řádně anotován vzhledem k použitým dialogovým aktům, které se tak staly jedním z příznaků pro výpočet ceny cíle (ta je obecně popsána v části 2.3.3, s modifikacemi pro využití v naší práci pak v části 8.2).

Proces návrhu vět, které budou obsaženy v expresivním korpusu je velmi důležitý. Vycházeli jsme tedy z přepisů již předtím pořízených nahrávek reálných dialogů, neboť právě ty by měly vystihovat podstatu dialogů a vymezovat omezenou oblast pro budoucí expresivní syntézu. Všechny promluvy (přesněji texty těchto promluv), které 3D avatar (mluvicí hlava) ve všech nahraných reálných dialozích pronesl (v celém audiovizuálním korpusu je takových promluv více než 7000), byly použity pro sestavení textů pro nahrávaní. To je zároveň také ten nejlepší možný případ, se kterým jsme se mohli setkat – korpus je dostatečně velký a bohatý a zachovalo se tak přirozené rozložení dialogových aktů. Pokud by však byl počet možných promluv příliš velký vzhledem ke zvažovanému způsobu nahrávání (viz dále), ať už z technických, finančních či jiných důvodů, bylo by potřeba tento počet nějakým způsobem zredukovat. Snížení počtu promluv by nebylo triviální záležitostí, nicméně lze použít například metodu vyřazení duplicitních promluv nebo metodu shlukování, kdy shlukovací algoritmus shlukne podobné věty do jedné

skupiny a vybere z ní pouze jednoho, nejvíce reprezentativního zástupce. Podobnost promluv může být dána jak jejich textovým obsahem, tak i jejich předpokládaným nebo odhadovaným expresivním vyjádřením za použití dialogových aktů, popř. vhodnou kombinací těchto přístupů. Tento algoritmus může být aplikován jak na původní, tak i na lematizované věty[1]. Je také možno brát v úvahu rozložení výskytů jednotlivých dialogových aktů, které může být v plánovaném korpusu buď rovnoměrné, nebo přirozené. Poté, co jsou vybrány texty (v našem případě všechny možné), lze pokračovat samotným nahráváním expresivního korpusu, což je popsáno v následující části 7.1.

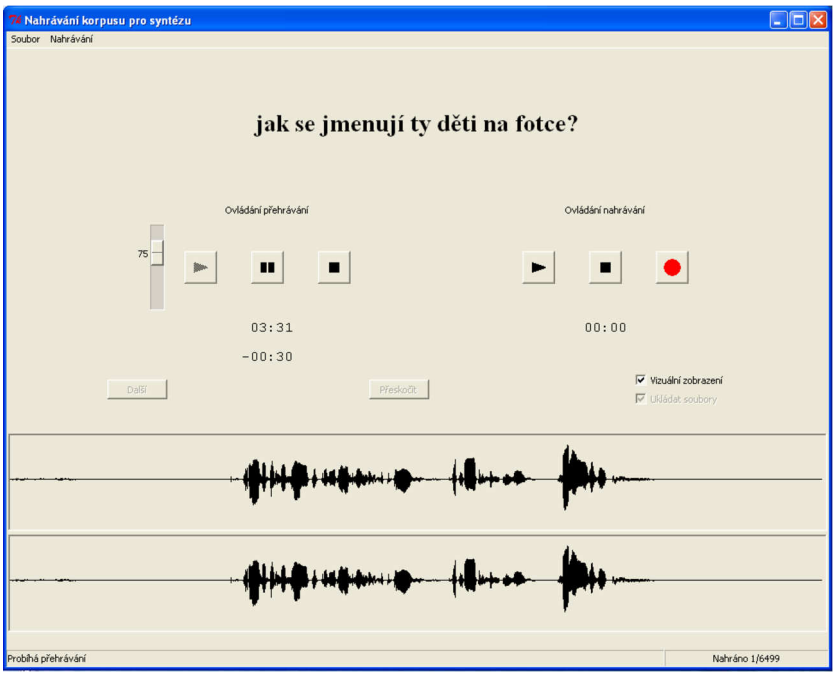
7.1 Nahrávání promluv

Pro nahrávání expresivního korpusu jsme zvolili metodu scénáře, zmíněnou v části 4.2.3. Jako řečníka jsme zvolili profesionální divadelní herečku, která již měla zkušenosti s nahráváním pro potřeby syntézy řeči – namlouvala jeden z neutrálních korpusů, které se v současnosti využívají v našem systému TTS. Tato skutečnost také koresponduje s potřebou mít k expresivnímu korpusu také korpus neutrální, abychom mohli provést vhodné sloučení korpusů. Nahrávání tedy probíhalo formou simulace reálného rozhovoru. Pro tento účel byla vyvinuta speciální aplikace, jejíž rozhraní je znázorněno na obrázku 7.1.

Řečníkovi vždy byla nejprve přehrána část původního reálného rozhovoru a jeho úkolem bylo podle předem daného scénáře vhodně reagovat a to pokud možno stejně jako v původním rozhovoru. Text, který měl přednést, byl zobrazen na obrazovce (převzat z původního dialogu), avšak řečník měl jistou volnost ve volbě přesných slov. Důležité bylo, aby byl zachován smysl věty. Pokud tedy řečník v určité situaci cítil, že by volil trochu jiná slova, popř. pro něj bylo snazší promluvu s upraveným textem přednést, bylo mu umožněno se od přesného znění odchýlit.

Přehrávaná reálná část rozhovoru měla řečníkovi poskytnout informaci o kontextu, v jakém se promluva objevila, a obsah nahrávané věty by měl řečníka stimulovat k expresivnímu vyjadřování. Aby kontext na řečníka působil co nejvíce, přehrávané části rozhovoru spolu ve většině případů souvisely. To znamená, že byly vybrány ucelené části rozhovorů, což byly v našem případě dialogy nad jednotlivými fotografiemi. Tento přístup se nám jeví jako lepší, než nahrávání izolovaných promluv bez vzájemné souvislosti. Řečník tak byl „vtažen“ do rozhovoru s konkrétní osobou o konkrétní fotografii. To ho mělo

[1] lematizace – proces, při kterém dochází k transformaci slov z kategorie ohebných slovních druhů na jejich základní tvar (např. otcův → otec, pracovali → pracovat).



Obrázek 7.1: Rozhraní aplikace pro nahrávání expresivního korpusu.

přimět reagovat přirozeně na vzniklé situace, přestože text rozhovoru byl předem daný.

Nahrávání probíhalo v odhlučněné nahrávací komoře za použití kvalitního mikrofonu, laryngografu[2], externí zvukové karty a s odborným dohledem. Vzorkovací frekvence nahrávek byla 48 kHz, pro pozdější použití k syntéze řeči byl řečový i hlasivkový signál převzorkován na 16 kHz.

Tímto způsobem jsme tedy získali materiál, který mohl být dále zpracován tak, aby byl vhodně spojen se stávající syntézou neutrální řeči a obohatil ji tak o expresivní řeč. Obecně by mělo být snahou pokrýt především ty věty, které se v reálných rozhovorech vyskytovaly častěji. Předpokladem totiž je, že i výsledný dialogový systém postavený na základě reálných dialogů bude častěji produkovat fráze vyskytující se v těchto dialozích. V našem postupu jsme otázku pokrytí řešit nemuseli, neboť jak již bylo uvedeno, v expresivním korpusu se objevily všechny věty z reálných dialogů. Samozřejmě je také nutné, aby výsledný systém byl schopný syntetizovat jakékoliv texty, proto tedy spojení s neutrální syntézou.

Díky metodě, kterou jsme zvolili pro nahrávání expresivního korpusu, jsme získali i malý soubor neřečových událostí (jako je například přitakání, úsměv, smích, apod.), které se někdy mimovolně, někdy cíleně, během na-

[2]Laryngograf je zařízení pro snímání hlasivkových pulzů.

hrávání objevily. Ty ale zatím nebyly využity, neboť jejich implementace do systému by znamenala velký zásah do samotného syntetizéru a není zcela triviální. Bylo by totiž potřeba nejprve analyzovat, v jakých situacích se objevují jaké neřečové události, ale především pak implementovat algoritmus, který by byl schopný neřečové události do syntetizované promluvy umístit ve vhodných okamžicích tak, aby jejich vložení působilo na posluchače přirozeně a aby se například často neopakovaly či ve výsledku nepůsobily naopak rušivě.

7.2 Anotace expresivního korpusu

K tomu, abychom mohli dialogové akty využít, je potřeba nahraná data anotovat. Pro anotaci expresivní řeči bylo navrženo několik postupů, které se liší podle zvoleného popisu expresivity. Několik jich ve stručnosti uvádí například [33]. Hlavním úkolem anotace v naší úloze je jednotlivé promluvy z expresivního korpusu označit v souladu s dialogovými akty tak, jak jsou uvedené v tabulce 6.1. Tohoto označení pak bude při syntéze využito jako příznaku pro všechny řečové jednotky vyskytující se v této promluvě.

Přestože například podle [21] je pro vnímání expresivity v řeči nutná znalost okolního kontextu a není tedy možné jednoznačně určit správnou expresivní kategorii pro izolované promluvy, rozhodli jsme se anotovat každou promluvu izolovaně. Efekt izolované promluvy mohl být částečně eliminován tím, že posluchači provádějící anotaci (anotátoři) byli nabádáni k poslechu vět tak, jak šly původně za sebou (avšak poslouchali pouze jednu stranu dialogu, tzv. avatara) a byly jim takto i přirozeně předkládány.

Pro anotace jsme využili stávající systém poslechových testů, který byl vyvinut na katedře kybernetiky a je běžně používán např. pro hodnocení kvality syntetizované řeči. Byl pochopitelně vhodně adaptován na anotační úlohu. Systém využívá webové rozhraní, které je zobrazeno na obrázku 7.2 a vzhledem k charakteru systému mohli anotátoři pracovat ze svých domovů ve svém volném čase.

Pro anotace expresivních kategorií v řečovém korpusu přichází v úvahu ještě možnost automatické (přesněji poloautomatické) anotace prezentované např. v [35, 31]. Zde byla malá část korpusu, nahraná v expresivním stylu (zdůrazněná slova), ručně anotována a na základě těchto ručních anotací byl vyvinut klasifikátor, který byl schopen v dalším (tentokráte mnohem větším) korpusu automaticky anotovat požadované expresivní styly. Tato konkrétní úloha byla oproti našemu problému poměrně snazší, protože klasifikátor se rozhodoval pouze mezi dvěma různými expresivními kategoriemi – zdůraz-

něné/nezdůrazněné. Nicméně takovýto postup by zcela jistě stál za bližší prozkoumání i v komplikovanějších situacích, jako je například ta naše.

Obrázek 7.2: Webové rozhraní pro anotace pomocí dialogových aktů.

Úloha ruční anotace takového množství promluv je jistě těžká a velmi časově náročná, přesto jsme získali 12 spolehlivých anotátorů (většina z nich pocházela z řad studentů). Anotátoři byli také finančně motivováni úlohu dokončit a to v rozumném čase 2 týdnů, bez nějakých náznaků podvodů či nespolehlivých anotací. Systém totiž disponuje možností odhalení takových praktik a anotátoři si byli vědomi toho, že v takovém případě by jim žádná odměna přiznána nebyla.

Anotátoři získali před vlastní anotací podrobné instrukce jak při anotacích postupovat. Měli také k dispozici několik příkladů jak anotace provádět, případně jak se rozhodnout a co dělat v případě, kdy by si správným přiřazením dialogového aktu k dané promluvě nebyli jisti. Poznamenejme ještě, že jim bylo výslovně řečeno, že neexistují žádné špatné a správné odpovědi, ale

že je vše na jejich vlastním rozhodnutí. Během anotací tedy byly každému anotátorovi postupně přehrány všechny věty z expresivního korpusu a jeho úkolem bylo vybrat ze seznamu dialogových aktů jeden, nebo více z nich, který se podle jeho přesvědčení k dané promluvě nejvíce hodí. Pokud anotátor shledal, že dialogových aktů lze přiřadit více, bylo jeho úkolem označit, zda se tyto dialogové akty v promluvě vyskytují a) „najednou“ nebo b) „za sebou“. To znamená, že v případě a) obsahuje promluva takový dialogový akt, který je kombinací jím označených dialogových aktů; v případě b) se promluva skládá z více částí (např. souvětí), kde každá část této promluvy představuje jiný dialogový akt. Nutno poznamenat, že z dalšího zpracování expresivního korpusu byly promluvy označené více dialogovými akty „za sebou“ vynechány, neboť by je bylo potřeba dále ručně zpracovat, tedy např. zkoumat, kde jsou v promluvě pravděpodobné přelomy mezi jednotlivými dialogovými akty a podle toho promluvu rozdělit na více částí.

Jak již bylo i v minulosti prokázáno, při anotaci expresivní nebo emotivní řeči více posluchači jsme vystaveni faktu, že různí posluchači vnímají různé projevy expresivity různě, tj. jednu a tu samou větu mohou posluchači označit různými dialogovými akty. Tato rozporuplnost se samozřejmě projevila i v námi anotovaných datech. Pro použití v expresivní syntéze bychom však potřebovali (bylo by to pro naši úlohu jednodušší), kdybychom měli každé promluvě přiřazený jeden dialogový akt. Existují dva základní způsoby, jak tohoto výsledku dosáhnout:

Metoda prosté většiny Tato jednoduchá metoda spočívá v tom, že promluvě bude přiřazen takový dialogový akt, na kterém se shodla většina posluchačů. Pokud neexistuje žádný dialogový akt, kterým by byla daná promluva označena více než 50 % posluchačů, bude anotace takové promluvy považována za nevěrohodnou a z dalších pokusů by měla být vyřazena.

Metoda maximální věrohodnosti Tato statistická metoda je založena na formulaci pravděpodobnostního modelu, který popisuje reálná data – anotace. Z daných dat se pak odhadují parametry takového modelu. Výpočet parametrů je v našem případě implementován použitím algoritmu EM [27]. Metoda i algoritmus jsou podrobněji popsány v přílohách A a B (částečně převzato z [97]). Metoda byla použita odděleně pro každý dialogový akt zvlášť. To znamená, že výsledky subjektivních anotací byly přeformulovány tak, aby všem promluvám v rámci daného dialogového aktu byla podle subjektivních anotací přiřazena binární hodnota 0/1, tj. zda promluva daný dialogový akt reprezentuje či nikoliv. Takto transformovaná data pak byla vstupem pro odhad parametrů pravděpodobnostního modelu. Stejným způsobem jsme

postupovali pro všechny dialogové akty. Získali jsme tak pravděpodobnostní modely odděleně pro každý dialogový akt. K tomuto postupu jsme se uchýlili z toho důvodu, že každá promluva mohla být označena více dialogovými akty zároveň (varianta „najednou“). Při znalosti parametrů pravděpodobnostních modelů můžeme jednotlivým promluvám automaticky přiřadit pro každý dialogový akt binární hodnotu 0/1. Přesnost takového odhadu je potom jedním z výstupů modelu. S využitím hodnoty přesnosti se pak můžeme rozhodnout, jaký dialogový akt přiřadíme promluvě, která podle pravděpodobnostního modelu „získala binární 1“ pro více dialogových aktů, popř. můžeme opět detekovat takové anotace, které označíme za nevěrohodné a měly by být z dalších pokusů vyřazeny. Takto získané anotace budeme v rámci uvažovaného postupu považovat pro další práci jako „objektivní“.

Použitím metody prosté většiny bychom z celkového počtu 7287 anotovaných expresivních promluv museli vyřadit 571 jako nevěrohodných a 138 označených příznakem „dialogové akty za sebou“. Průměrný poměr posluchačů, kteří označili stejný dialogový akt u dané promluvy (počítaný přes všechny věrohodně anotované promluvy) byl 81 %. Použitím metody maximální věrohodnosti bychom byli nuceni z 7287 anotovaných promluv vypustit pouze 35 jako nevěrohodných a 265 označených příznakem „dialogové akty za sebou“. Rozhodli jsme se tedy využít metodu maximální věrohodnosti, a to nejen z důvodu menšího počtu potenciálně nepoužitelných promluv, ale také proto, že podobný postup byl již v minulosti úspěchem použit [96, 97]. Příklady anotací některých vět včetně shody posluchačů, případně přesnosti pravděpodobnostního modelu uvádí tabulka F.2 v příloze F.

Protože jsme se rozhodli pro metodu maximální věrohodnosti, chtěli bychom také zjistit, s jakou mírou shody jednotliví anotátoři označovali promluvy dialogovými akty (a také variantu více po sobě jdoucích dialogových aktů v jedné promluvě). Pokud by nebyla zjištěna shoda mezi anotátory, mohli bychom:

(a) považovat subjektivní anotace jednotlivých anotátorů jako náhodné, což by mohlo značit jejich nespolehlivost – anotace takovýchto anotátorů bychom pak museli vypustit a celý proces objektivní anotace (s přetrénováním statistického modelu) opakovat;

(b) považovat množinu dialogových aktů jako špatně navrženou, pak by bylo nutné zrevidovat postup navržení dialogových aktů a celou anotaci opakovat.

K určení míry shody mezi anotátory existuje mnoho přístupů [90, 19], my jsme se rozhodli využít dvě statistické míry: Fleissovu kappu [36] a Cohenovu kappu [22].

První z nich, Fleissova kappa, je statistická míra pro posouzení spolehlivosti shody mezi pevným počtem hodnotitelů (anotátorů) při přiřazování kategorických hodnocení určitému počtu objektů nebo při klasifikaci objektů. Kategorické hodnocení je v našem případě přeformulované přiřazování dialogových aktů (pro každý dialogový akt binární 0/1), objektem je pak expresivní promluva. Provedli jsme tedy výpočet této míry mezi všemi posluchači, odděleně pro každý dialogový akt.

Tabulka 7.1: Slovní označení pro různé hodnoty kappa. Využívá se jak pro míru Cohenova kappa, tak pro míru Fleissova kappa.

kappa	míra shody
0,0 - 0,2	Nepatrná
0,2 - 0,4	Mírná
0,4 - 0,6	Střední
0,6 - 0,8	Značná
0,8 - 1,0	Téměř dokonalá

Hodnota Fleissovy kappy κ_F vždy leží v rozsahu 0 až 1, pokud je shoda vyšší než náhodná, a $\kappa_F < 0$, pokud je naměřená shoda pod úrovní náhodné shody. Z toho vyplývá, že čím vyšší hodnota, tím vyšší shoda. Na výsledné vyjádření, jakou míru shody vlastně hodnota *kappa* představuje (tj. kdy už je shoda významná), není žádný obecný mechanismus, nicméně obecně je přijímána charakteristika uvedená v tabulce 7.1. Výsledky výpočtu shody pomocí této míry v naší úloze jsou shrnuté v tabulce 7.2.

Druhým ukazatelem shody je Cohenova kappa, což je obdobná statistická míra jako Fleissova kappa, avšak určuje míru shody pouze mezi dvěma hodnotiteli (anotátory). Použili jsme tedy tuto míru k určení shody mezi objektivní anotací získanou metodou maximální věrohodnosti a každým jednotlivým anotátorem a poté jsme vypočítali střední hodnotu a směrodatnou odchylku. Výpočet Cohenovy kappy byl opět proveden odděleně pro každý dialogový akt. Tím bychom mohli odhalit anotátory, kteří prováděli anotace:

nekonzistentně – např. anotovali jeden dialogový akt v souladu s objektivní anotací a jiný zcela odlišně, což by ukazovalo na odlišné vnímání různých dialogových aktů;

Tabulka 7.2: Fleissova kappa jako míra shody mezi všemi anotátory pro jednotlivé dialogové akty.

dialogový akt	Fleissova kappa	Míra shody
DIRECTIVE	0,7282	Značná
REQUEST	0,5719	Střední
WAIT	0,5304	Střední
APOLOGY	0,6047	Značná
GREETING	0,7835	Značná
GOODBYE	0,7408	Značná
THANKS	0,8285	Téměř dokonalá
SURPRISE	0,2477	Mírná
SAD-EMPATHY	0,6746	Značná
HAPPY-EMPATHY	0,6525	Značná
SHOW-INTEREST	0,4485	Střední
CONFIRM	0,8444	Téměř dokonalá
DISCONFIRM	0,4928	Střední
ENCOURAGE	0,3739	Mírná
NOT-SPECIFIED	0,1495	Nepatrná
OTHER	0,0220	Nepatrná
střední hodnota	**0,5434**	**Střední**
po sobě jdoucí DA	0,5138	Střední

zcela náhodně – což by mělo za následek vyřazení jejich anotací a nový výpočet objektivní anotace.

Hodnoty Cohenovy kappy κ_C, stejně jako u předchozí míry, leží v rozsahu 0 až 1, pokud je shoda vyšší než náhodná, a $\kappa_C < 0$, pokud je naměřená shoda pod úrovní náhodné shody. Opět platí, že čím vyšší hodnota, tím vyšší shoda, viz tabulka 7.1. Naměřené výsledky shrnuje tabulka 7.3.

Střední hodnota Fleissovy kappy $\kappa_F = 0{,}5434$ znamená, že anotátoři dosáhli míry shody označené jako střední. Jak je vidět z tabulky 7.2, dialogové akty *OTHER* a *NOT-SPECIFIED* by mohli být označené jako špatně rozpoznatelné. U dialogového aktu *OTHER* to bude zřejmě způsobeno jeho podstatou, kdy tuto variantu volili anotátoři až jako poslední možnost, kdy v promluvě nějaké expresivní vyjádření cítili, avšak žádná z nabídnutých expresivních kategorií tomuto jejich pocitu neodpovídala. Vysvětlením pro nízkou hodnotu Fleissovy kappy u dialogového aktu *NOT-SPECIFIED* pak

Tabulka 7.3: Cohenova kappa jako míra shody mezi každým jednotlivým anotátorem a objektivní anotací. Je uvedena střední hodnota a směrodatná odchylka přes všechny anotátory.

dialogový akt	Cohenova kappa	Cohenova kappa (odchylka)	Míra shody
DIRECTIVE	0,8457	0,1308	Téměř dokonalá
REQUEST	0,7280	0,1638	Značná
WAIT	0,7015	0,4190	Značná
APOLOGY	0,7128	0,2321	Značná
GREETING	0,8675	0,1287	Téměř dokonalá
GOODBYE	0,7254	0,1365	Značná
THANKS	0,8941	0,1352	Téměř dokonalá
SURPRISE	0,4064	0,1518	Střední
SAD-EMPATHY	0,7663	0,0590	Značná
HAPPY-EMPATHY	0,7416	0,1637	Značná
SHOW-INTEREST	0,6315	0,3656	Značná
CONFIRM	0,9148	0,0969	Téměř dokonalá
DISCONFIRM	0,7153	0,1660	Značná
ENCOURAGE	0,5914	0,3670	Střední
NOT-SPECIFIED	0,3295	0,2292	Mírná
OTHER	0,0391	0,0595	Nepatrná
střední hodnota	**0,6632**		**Značná**
po sobě jdoucí DA	0,6570	0,2443	Značná

může být fakt, že někteří anotátoři byly v hodnocení expresivity v řeči více citliví, jiní méně. Pokud bychom z výsledků vyřadili tyto ze zřejmých důvodů pochopitelně nízké hodnoty, střední hodnota Fleissovy kappy by byla $\kappa_F = 0{,}6191$, což by znamenalo již značnou shodu.

Střední hodnota Cohenovy kappy $\kappa_C = 0{,}6632$ znamená, že můžeme pozorovat značnou shodu mezi subjektivními anotacemi jednotlivých anotátorů a objektivní anotací určenou námi pomocí metody maximální věrohodnosti. Pokud stejně jako v předchozím případě vyřadíme hodnoty pro dialogové akty *OTHER* a *NOT-SPECIFIED* (ze stejných důvodů jako předtím), získáme střední hodnotu Cohenovy kappy $\kappa_C = 0{,}7316$, což je však slovně klasifikováno stále jako značná shoda.

Jak je vidět z tabulky 7.2 a tabulky 7.3, byla měřena i shoda mezi anotátory co se týká označování dialogových aktů jako po sobě jdoucích v rámci jedné promluvy. Anotátoři na tomto příznaku dosáhli střední shody mezi se-

bou ($\kappa_F = 0{,}5138$) a značné shody s objektivní anotací ($\kappa_C = 0{,}6570$). Míry shody u všech ostatních dialogových aktů pak byly měřeny po vypuštění promluv, u nichž byla tato vlastnost indikována.

Ještě poznamenejme, že u žádného z anotátorů nebyla nalezena žádná významná odchylka co se týká shody s objektivní anotací, viz tabulka 7.4 – z důvodu rozsáhlosti této tabulky jsou naměřené hodnoty zaokrouhlené a nejsou uvedeny hodnoty pro dialogové akty *OTHER* a *NOT-SPECIFIED*. Pouze u jednoho anotátora (číslo 12) byla u dvou dialogových aktů (*SHOW-INTEREST* a *ENCOURAGE*) naměřena velmi nízká shoda s objektivní anotací – dokonce pod hranicí náhody. Vysvětlujeme si to tím, že anotátor mohl význam těchto dvou dialogových aktů pochopit obráceně. Dále můžeme u tří anotátorů (čísla 8, 10 a 11) pozorovat nízkou shodu u dialogového aktu *WAIT*. Pro konečné stanovení objektivní anotace jsme tedy u těchto dialogových aktů anotace příslušných anotátorů neuvažovali, avšak u ostatních jsme jejich anotace započítali, neboť jejich výsledky se vesměs shodují s objektivní anotací. Z uvedených výsledků tedy plyne, že nebylo nutné žádného z anotátorů kompletně vyřadit pro nevěrohodnost, neboť zmíněné odchylky u uvedených dialogových aktů a anotátorů byly výjimečné.

Bylo tedy prokázáno, že takto získané anotace můžeme považovat za věrohodné a můžeme s nimi dále pracovat. V tabulce 7.5 je pak uvedena relativní četnost jednotlivých dialogových aktů v celém expresivním korpusu (pro větší představu společně s příklady takových promluv). Součet všech relativních četností přesahuje hodnotu 100 %, což je způsobeno tím, že anotátoři mohli každé promluvě přiřadit více než jeden dialogový akt. Tato skutečnost se pak projevila i v objektivních anotacích. U promluv, kterým byl přiřazen více než jeden dialogový akt, jsme však pro další zpracování přiřadili ten dialogový akt, jehož přesnost (z hlediska modelu popsaného u metody maximální věrohodnosti) byla nejvyšší, pokud ovšem překročila stanovenou mez. Tu jsme stanovili jako 0,5, přičemž přesnost se teoreticky může pohybovat v intervalu $\langle 0; 1\rangle$. Pokud přesnost odhadu tuto mez nepřekročila, promluvu jsme vyřadili jako nevěrohodně anotovanou. Takovéto případy se většinou překrývaly s promluvami, ve kterých se dle objektivní anotace vyskytly dialogové akty za sebou, které jsme taktéž vyřadili. Ještě poznamenejme, že přesnost odhadu nejpravděpodobnějšího (a tedy použitého) dialogového aktu se u jednotlivých promluv téměř vždy pohybovala v intervalu $\langle 0{,}99; 1{,}00\rangle$.

Je vidět, že některé dialogové akty se v expresivním korpusu vyskytují významně častěji než jiné, a to především *SHOW-INTEREST*, *ENCOURAGE*, popř. ještě *CONFIRM*. Jiné se naopak vyskytují velmi zřídka, jako např. *DISCONFIRM*, *APOLOGY*, *THANKS* nebo *WAIT*. Tyto výsledky odpovídají povaze rozhovorů, kdy v takovém dialogovém systému je nesouhlas s uživatelem výjimečný stejně tak jako poděkování či omluva. Naopak proje-

Tabulka 7.4: Cohenova kappa pro jednotlivé anotátory, tj. shoda mezi daným anotátorem a objektivní anotací pro jednotlivé dialogové akty (zaokrouhlené hodnoty). Dialogové akty *NOT-SPECIFIED* a *OTHER* nebyly z důvodů uvedených na straně 65 do tabulky zahrnuty.

dialogový akt	anotátor	1	2	3	4	5	6	7	8	9	10	11	12	**stř. hod.**
DIRECTIVE		0,91	0,98	0,91	0,87	0,92	0,93	0,91	0,66	0,92	0,69	0,56	0,88	**0,85**
REQUEST		0,50	0,94	0,80	0,74	0,80	0,68	0,78	0,82	0,88	0,63	0,36	0,80	**0,73**
WAIT		0,97	1,00	1,00	0,75	0,94	1,00	0,82	0,07	0,97	0,00	0,00	0,91	**0,70**
APOLOGY		0,45	0,43	0,96	0,94	0,74	0,40	0,56	0,51	0,71	0,96	0,93	0,96	**0,71**
GREETING		0,69	0,98	0,69	0,98	1,00	0,99	0,96	0,76	0,82	0,92	0,93	0,68	**0,87**
GOODBYE		0,60	0,65	0,91	0,87	0,64	0,60	0,59	0,95	0,83	0,80	0,61	0,64	**0,73**
THANKS		0,99	0,99	1,00	0,70	0,71	1,00	0,99	0,99	1,00	0,91	0,70	0,74	**0,89**
SURPRISE		0,28	0,33	0,22	0,31	0,20	0,55	0,50	0,39	0,53	0,72	0,42	0,41	**0,41**
SAD-EMPATHY		0,85	0,82	0,75	0,76	0,72	0,76	0,78	0,74	0,88	0,77	0,67	0,70	**0,77**
HAPPY-EMPATHY		0,88	0,78	0,74	0,82	0,84	0,85	0,85	0,67	0,76	0,74	0,26	0,70	**0,74**
SHOW-INTEREST		0,69	0,88	0,68	0,97	0,91	0,96	0,32	0,91	0,38	0,68	0,48	-0,28	**0,63**
CONFIRM		0,97	0,93	0,97	0,97	0,97	0,98	0,98	0,76	0,96	0,87	0,68	0,96	**0,91**
DISCONFIRM		0,85	0,46	0,79	0,84	0,93	0,85	0,47	0,79	0,75	0,44	0,71	0,72	**0,72**
ENCOURAGE		0,76	0,86	0,79	0,91	0,88	0,88	0,27	0,89	0,13	0,60	0,29	-0,17	**0,59**
střední hodnota		**0,74**	**0,79**	**0,80**	**0,82**	**0,80**	**0,81**	**0,70**	**0,71**	**0,75**	**0,70**	**0,54**	**0,62**	**0,73**

Tabulka 7.5: Četnost výskytu jednotlivých dialogových aktů v expresivním korpusu společně s ukázkami vět.

dialogový akt	relativní četnost výskytu	příklad
DIRECTIVE	2,4%	Řekněte mi to. Povídejte.
REQUEST	4,4%	Vraťme se k tomu později.
WAIT	0,7%	Počkejte chvilku. Chvíli strpení.
APOLOGY	0,6%	Promiňte. Omlouvám se.
GREETING	1,4%	Ahoj. Dobrý den.
GOODBYE	1,6%	Na shledanou. Uvidíme se později.
THANKS	0,7%	Děkuji vám. Díky.
SURPRISE	4,2%	Opravdu máte 10 sourozenců?
SAD-EMPATHY	3,4%	To je mi líto. To je smutné.
HAPPY-EMPATHY	8,6%	To je hezké. Skvělé.
SHOW-INTEREST	34,9%	Můžete mi o tom ještě něco říct?
CONFIRM	13,2%	Ano. Jo. Aha.
DISCONFIRM	0,2%	Ne. Nerozumím tomu.
ENCOURAGE	29,4%	Dobře, a co vy, jak se tam líbilo vám?
NOT-SPECIFIED	7,4%	Jmenuji se Petra.
po sobě jdoucí DA	3,7%	

vování zájmu o další informace či povzbuzování v pokračování vyprávění je velice žádoucí.

7.3 Akustická analýza expresivní řeči

Před tím, než použijeme získaný expresivní korpus anotovaný pomocí dialogových aktů v oblasti syntézy české expresivní řeči, jsme se rozhodli nejprve provést akustickou analýzu expresivního řečového signálu. Inspirovaly nás k tomu podobné analýzy, které byly provedeny např. pro japonštinu [57, 79], italštinu [134], němčinu [64], angličtinu [64, 26] a dále samozřejmě i pro ostatní jazyky, včetně námi provedené akustické analýzy české emotivní řeči [44]. Výsledky akustické analýzy pak bývají využity buď ve

výzkumu expresivní/emotivní řeči, v rozpoznávání expresivity/emocí z akustického signálu nebo, jako tomu bude i v našem případě, ve výzkumu syntézy expresivní řeči. Na základě výsledků analýzy totiž budeme v části 8.1.2 vytvářet akustickou penalizační matici, která bude posléze použita v algoritmu dynamického výběru jednotek při určování ceny cíle.

Jedním z cílů akustické analýzy je zjistit, jaké akustické parametry řeči ovlivňují její vnímání posluchačem jako expresivní. V naší úloze jde tedy o to nalézt tzv. *akustické koreláty* různých expresivních kategorií (dialogových aktů) v řečovém signálu a zjistit, zda se jejich hodnoty pro různé expresivní kategorie liší a jak moc. Výsledky akustické analýzy pak mohou být v rámci úlohy syntézy expresivní řeči metodou dynamického výběru jednotek použity pro určení penalizace (viz část 2.3.3 a rovnice 2.1) příznaku „dialogový akt“. Postup výpočtu konkrétních koeficientů penalizační matice z akustické analýzy je pak uveden v části 8.1.2. Předpokládáme, že takto určené akustické koreláty a jejich využití při dynamickém výběru jednotek pomůže vytvářet přirozenější syntetickou řeč.

Závěry vyvozené z akustické analýzy však nelze v žádném případě zobecňovat, neboť expresivní korpus byl nahrán pouze za přispění jednoho řečníka s jasně definovanými dialogovými akty pro specifickou úlohu syntézy řeči v dialogu seniorů s počítačem. Pro účely takto definované syntézy expresivní řeči jsou však dostačující, protože syntetický hlas plně napodobuje právě hlas řečníka, který korpus nahrál a to i v těch parametrech, které ovlivňují vnímání řeči jako expresivní. Analýza dat řečníka, jehož hlas bude použit pro syntézu, je tedy opodstatněná a získané výsledky důležité, nikoliv však obecné.

V průběhu akustické analýzy expresivní řeči byly měřeny následující akustické parametry:

- průměrná hodnota $F0$ pro všechny znělé fonémy;
- průměrná hodnota $F0$ pro celou promluvu (započítány pouze znělé úseky);
- maximální hodnota $F0$ pro celou promluvu (započítány pouze znělé úseky);
- minimální hodnota $F0$ pro celou promluvu (započítány pouze znělé úseky);
- rozsah hodnoty $F0$ pro celou promluvu (rozdíl mezi maximem a minimem pro danou promluvu);
- průměrná hodnota doby trvání jednotlivých fonémů;

- průměrná hodnota RMS řečového signálu jako energie RMS pro jednotlivé fonémy;

- průměrná hodnota formantových frekvencí ($F1$, $F2$, $F3$) pro všechny české krátké samohlásky (/a/, /e/, /i/, /o/, /u/).

Fonémové akustické parametry byly počítány zvlášť pro různé skupiny fonémů (znělé, neznělé, samohlásky, souhlásky, znělé souhlásky, neznělé souhlásky, apod.) i pro jednotlivé fonémy. Stejně tak větné akustické parametry byly počítány z různých skupin fonémů v dané větě, toto však již nebylo předmětem našeho dalšího zkoumání. Využili jsme především hodnoty vypočtené pro vhodnou skupinu fonémů – pro parametry vztahující se k $F0$ jsme využili všech znělých fonémů, pro výpočet doby trvání a hodnoty RMS pak všech dostupných fonémů (vyloučili jsme segmenty reprezentující pauzy a rázy). Dále zde uvádíme i naměřené hodnoty pro foném /e/, který jsme vybrali jako zástupce jednoho z nejčastěji se vyskytujících fonémů v českém jazyce. Hodnoty získané pro tento foném nám budou sloužit k ověření, zda hodnoty naměřené na větší skupině fonémů (znělé fonémy, resp. všechny fonémy) jsou dostatečně reprezentativní, tedy že nevyváženost fonémů v promluvách označených jednotlivými dialogovými akty nijak zásadně neovlivňuje námi prezentované výsledky. Toto ověření provedeme pomocí korelačních koeficientů – pokud bude dostatečná korelace mezi hodnotami naměřenými pro velkou skupinu fonémů a zvlášť pro foném /e/, můžeme považovat výsledky pro velké skupiny fonémů jako reprezentativní (viz část 7.3.7). Formantová analýza je pak samozřejmě specifická, neboť hodnoty všech pozorovaných formantů se pro různé samohlásky liší.

Lze namítnout, že tyto zvolené segmentální charakteristiky řečového signálu mohou být expresivním zabarvením ovlivněny různě v různých částech jedné promluvy. Jako příklad uveďme domněnku, že nejvýrazněji bude expresivitou ovlivněna první, nebo naopak poslední část promluvy, nebo pouze ty části, na nichž lze identifikovat přízvuk. Rozhodli jsme se však nebrat tento fakt v úvahu[3] a to z následujících důvodů:

velké množství dat – Pro většinu dialogových aktů máme k dispozici velké množství dat (segmentů) a předpokládáme, že vypočtené statistické parametry měřených veličin nebudou uvedeným faktem ovlivněné, tedy že i tak budou vhodně data reprezentovat.

[3]Zde se otevírá možnost pro případné budoucí vylepšení tohoto postupu – mohli bychom zjistit, kde v promluvě se expresivita projevuje nejsilněji a podle toho pak upravit i algoritmus dynamického výběru jednotek.

konzistentní chyba – Pokud se ignorováním tohoto faktu dopouštíme nějaké chyby, předpokládáme, že se stejné chyby dopouštíme konzistentně u všech dialogových aktů, a pro získání akustické penalizační matice (tedy nějakého rozdílu mezi charakteristikami různých dialogových aktů) opět získáme vhodnou reprezentaci dat.

anotace dat – Expresivní korpus je anotován tak, že každé větě je přiřazen jeden dialogový akt (ne tedy jen nějakým určitým částem věty) a následný algoritmus dynamického výběru jednotek tak zachází se všemi jednotkami z dané věty, jako by všechny reprezentovaly ten stejný dialogový akt.

Akustickou analýzu jsme provedli jak pro expresivní korpus s anotovanými dialogovými akty, tak pro neutrální korpus stejného řečníka (resp. pro jeho část, vybrali jsme z něj 4000 vět), abychom mohli provést vzájemné porovnání. Zajímaly nás také výsledky porovnání neutrální řeči (dále označované jako dialogový akt *NEUTRAL*) s promluvami označenými dialogovým aktem *NOT-SPECIFIED*, který by měl taktéž reprezentovat neutrální řeč. Je však pravdou, že expresivní korpus a neutrální korpus byly nahrány s mírným časovým rozestupem[4], stejně tak nastavení nahrávacího zařízení nebylo zcela identické (z technických důvodů). Tento fakt může ovlivnit takové porovnání naměřených hodnot, zejména pak co se týká hodnot RMS akustického signálu.

Ve výsledcích analýz neuvádíme pro lepší přehlednost hodnoty pro dialogový akt *OTHER*, neboť pro něj zpravidla není dostatek reprezentativních dat. Ve výpočtu penalizační matice však již zahrnut je, neboť i jednotky označené tímto dialogovým aktem mohou být použity během syntézy expresivní řeči.

7.3.1 Zpracování dat

Hodnoty jednotlivých akustických parametrů jsme získali z expresivního řečového korpusu, postupy získání dat pro jednotlivé analýzy jsou popsány v následujících částech. Avšak před tím, než jsme získaná data mohli použít k extrakci akustických parametrů a jejich charakteristik, bylo potřeba je určitým způsobem předzpracovat.

Vzhledem k tomu, že všechna data jsou získávána na základě automatických metod, které fungují s určitou chybovostí, a dat je poměrně velké množství (alespoň pro většinu dialogových aktů), musíme z nich odstranit tzv. odlehlé hodnoty („outliery“, z anglického *outliers*). To jsou data, která

[4]Časový rozestup mezi nahrávkami byl přibližně půl roku.

jsou nekonzistentní s ostatními daty a mohli by další výsledky experimentů zkreslit. Pro identifikaci takových hodnot v datech a vypořádání se s nimi existuje několik statistických postupů, které využívají různé statistické veličiny (charakteristiky) vypočtené ze souboru dat X. Některé dále použité statistické charakteristiky jsou popsány v příloze C. Uveďme tedy některé používané postupy pro odstranění odlehlých prvků v jednorozměrném prostoru[5]:

(a) použití metody modifikovaného Thompsonova τ (popsána v příloze D);

(b) pomocí 10% a 90% percentilu, tj. odstranění takových hodnot $x_i \in X$, pro které platí:
$x_i > Q_{0,90}$ nebo $x_i < Q_{0,10}$;

(c) pomocí střední hodnoty μ a směrodatné odchylky σ, tj. odstranění takových hodnot $x_i \in X$, pro které platí:
$x_i > \mu + k\sigma$ nebo $x_i < \mu - k\sigma$,
kde k je zvolená konstanta, obvykle $k = 3$;

(d) pomocí mezikvartilového rozpětí, tj. odstranění takových hodnot $x_i \in X$, pro které platí:
$x_i > Q_{0,75} + w(Q_{0,75} - Q_{0,25})$ nebo $x_i < Q_{0,25} - w(Q_{0,75} - Q_{0,25})$,
kde w je zvolená konstanta, obvykle $w = 1,5$.

Těmito postupy můžeme identifikovat odlehlé hodnoty pro jednorozměrné veličiny, v našem případě tedy pro jednotlivé naměřené akustické parametry zvlášť (F0, doba trvání, RMS, atd.). Musíme však vzít v úvahu i možnost, že prvek ve vícerozměrném prostoru[6] může být outlier i přesto, že v každé jednotlivé dimenzi outlier není (a naopak, přestože je prvek v nějaké dimenzi outlier ještě nemusí vždy nutně znamenat, že je outlier také ve vícerozměrném prostoru). Pro detekci outlierů ve vícerozměrném prostoru pak lze využít například algoritmy založené na „Wilksově metodě“ [124, 94] popsané

[5]Všechny uvedené metody lze ještě v našem případě modifikovat tak, že nejprve odstraníme prvky s hodnotou 0, neboť ty jsou v našem souboru dat zcela jistě chybné.

[6]Vícerozměrným prostorem zde rozumíme prostor, jehož dimenze jsou tvořeny použitými akustickými parametry, tedy prvek (segment, foném) v tomto prostoru je charakterizován vektorem souřadnic, např. $[f0, dur, rms]$, kde $f0$ značí hodnotu $F0$, dur hodnotu doby trvání a rms hodnotu RMS (dimenze prostoru samozřejmě záleží na použitých akustických parametrech). V tomto případě je nutné, aby si naměřené hodnoty odpovídaly, tedy aby x-tá hodnota v souboru dat s hodnotami $F0$ patřila stejnému segmentu jako x-tá hodnota v souboru s hodnotami doby trvání, atd. Toho lze však dosáhnout jen pro znělé segmenty, pokud budeme chtít využít akustické parametry vztahující se k $F0$. Jinou variantou je pak pro neznělé segmenty uvažovat, že $F0 = 0$.

v příloze E. Samozřejmě lze tuto metodu využít i pro detekci outlierů v jednorozměrném prostoru.

Pro určení základních statistických parametrů (střední hodnota, směrodatná odchylka, apod.) zobrazených v tabulkách u jednotlivých akustických analýz jsme se rozhodli využít dvě metody pro odstranění outlierů. První z nich je metoda (a), tedy metoda modifikovaného Thompsonova τ obohacená o předchozí odstranění nulových hodnot. Ta je určena pouze pro jednorozměrné veličiny – z důvodu možnosti využití všech dostupných dat pro určitý akustický parametr (pro analýzu $F0$ pouze znělé segmenty, pro analýzu doby trvání a hodnot RMS všechny segmenty). Dále ji budeme označovat jako metodu TT. Jako druhou jsme použili Wilksovu metodu (použitelná jak pro vícerozměrné, tak pro jednorozměrné veličiny) s předchozím odstraněním nulových hodnot – dále označovaná jako metoda WILKS. Tato metoda bude totiž později použita pro odstraňování outlierů ve vícerozměrných datech, ze kterých se bude vytvářet akustická penalizační matice (popsaná v části 8.1.2). U každé akustické analýzy je pak vždy uveden jak celkový počet prvků (segmentů), které se v korpusu vyskytují (N_P), tak i počet (poměr) odlehlých hodnot (P_O), které byly z dat těmito metodami odstraněny.

Jako výsledky akustických analýz jsou pak prezentovány statistické charakteristiky (podrobněji popsané v příloze C) vypočtené z dat po odstranění outlierů:

- střední hodnota μ (míra polohy náhodné veličiny);
- směrodatná odchylka σ (míra statistické variability náhodné veličiny);
- koeficient šikmosti γ_1 (popisuje nesymetrii rozdělení pravděpodobnosti náhodné veličiny);
- koeficient špičatosti γ_2 (porovnává pravděpodobnostní rozdělení dané veličiny s normálním rozdělením pravděpodobnosti).

Předpokládáme, že tyto charakteristiky vhodně popisují rozložení pravděpodobnosti jednotlivých hodnot akustických parametrů a budou tedy (po transformaci popsané v části 8.1.2) využity i při vytváření akustické penalizační matice. Koeficienty šikmosti a špičatosti jsou použity i v jiných studiích zabývajících se analýzou řečového signálu, například [86, 85].

7.3.2 Analýza F0 pro fonémy

K výpočtu hodnot $F0$ pro jednotlivé segmenty byl využit glotální signál, který byl při nahrávání zaznamenáván společně s řečovým signálem. Byl použit algoritmus detekce hlasivkových pulzů, tzv. pitchmarků (z anglického *pitchmarks*), tedy bodů v řečovém signálu, které korespondují s uzavřením

hlasivek. Tento algoritmus je podrobněji popsaný v [68] nebo [69]. Z této posloupnosti pitchmarků byla pro každý znělý segment získána průměrná hodnota $F0$ následujícím způsobem: nejprve jsme získali lokální odhad $F0$ jako medián inverzních hodnot vzdáleností mezi čtyřmi po sobě jdoucími pitchmarky a poté byla tato posloupnost lokálních odhadů $F0$ vyhlazena použitím mediánového filtru řádu tři. Z celého souboru dat pak byly odstraněny odlehlé hodnoty metodami TT a WILKS, viz část 7.3.1. Nakonec byly vypočteny statistické charakteristiky pro každý dialogový akt odděleně. Výsledky po odstranění outlierů metodou TT shrnují tabulky 7.6 a 7.7, výsledky pro metodu WILKS jsou uvedeny v příloze F v tabulce F.3.

Tabulka 7.6: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty $F0$ znělých fonémů, metoda odstranění outlierů TT.

	Všechny znělé fonémy				
Dialogový akt	$\mu \pm \sigma$ [Hz]	γ_1	γ_2	N_P	P_O
APOLOGY	194 ± 17	- 0,17	- 0,24	892	15%
CONFIRM	188 ± 29	- 0,08	- 0,05	3916	29%
DIRECTIVE	191 ± 28	0,29	- 0,10	2208	18%
DISCONFIRM	179 ± 27	- 0,46	- 0,39	111	17%
ENCOURAGE	202 ± 25	- 0,01	- 0,26	33 807	9%
GOODBYE	192 ± 23	- 0,21	- 0,23	1368	15%
GREETING	182 ± 30	- 0,10	- 0,05	2107	12%
HAPPY-EMPATHY	185 ± 19	- 0,15	- 0,16	4969	15%
NOT-SPECIFIED	194 ± 25	- 0,13	- 0,38	4010	12%
REQUEST	201 ± 22	- 0,12	- 0,30	4415	11%
SAD-EMPATHY	181 ± 19	- 0,20	- 0,11	1712	15%
SHOW-INTEREST	201 ± 29	- 0,03	- 0,08	32 765	14%
SURPRISE	194 ± 27	- 0,03	- 0,40	1426	9%
THANKS	191 ± 18	- 0,24	- 0,52	771	7%
WAIT	180 ± 28	0,13	- 0,37	1368	17%
NEUTRAL	197 ± 43	0,03	- 0,27	160 988	9%

Z výsledků lze vyčíst, že hodnoty $F0$ se v závislosti na označeném dialogovém aktu mění a mezi jednotlivými dialogovými akty jsou různé rozdíly. Provedli jsme test významnosti ANOVA, jehož výsledkem je zjištění, že rozdíly mezi všemi dialogovými akty jsou statisticky významné (na hladině významnosti $\alpha = 0{,}05$). Tento test významnosti však může být ovlivněn velkým počtem dialogových aktů. Na obrázku F.1 v příloze F lze tedy pozo-

Tabulka 7.7: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty $F0$ fonému */e/*, metoda odstranění outlierů TT.

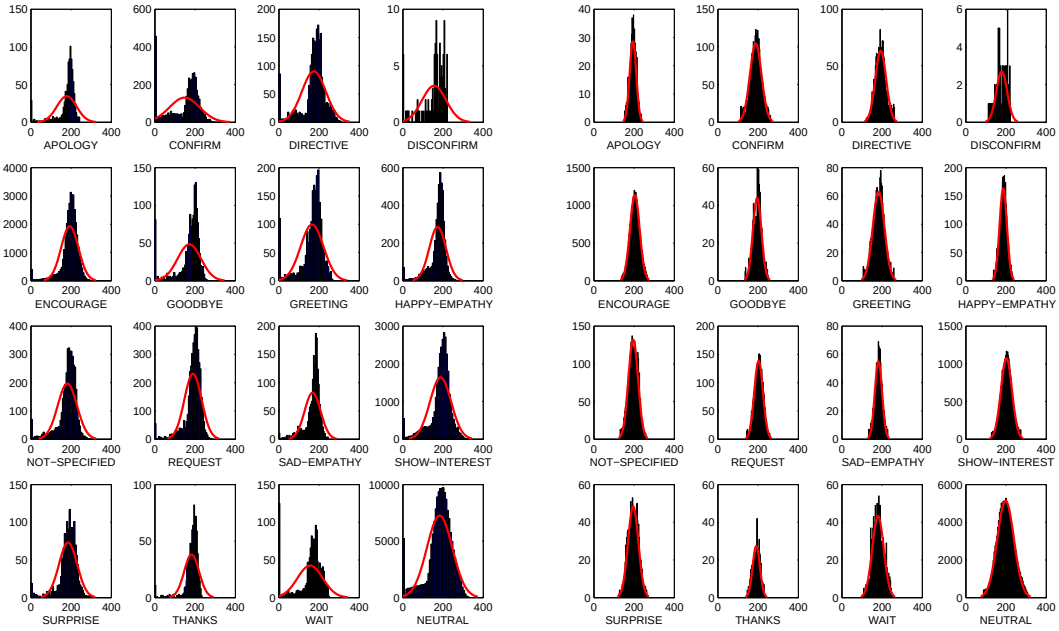
	Foném */e/*				
Dialogový akt	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O
APOLOGY	202 ± 12	0,34	- 0,50	131	13%
CONFIRM	114 ± 52	0,43	- 0,80	355	1%
DIRECTIVE	209 ± 32	0,35	- 0,06	475	12%
DISCONFIRM	187 ± 22	- 0,06	- 1,16	28	14%
ENCOURAGE	213 ± 24	0,03	- 0,23	6227	7%
GOODBYE	209 ± 20	- 0,08	- 0,29	158	2%
GREETING	189 ± 18	- 0,11	- 0,53	230	13%
HAPPY-EMPATHY	189 ± 14	0,14	- 0,18	904	15%
NOT-SPECIFIED	200 ± 21	0,09	- 0,35	375	12%
REQUEST	214 ± 21	- 0,08	- 0,43	880	4%
SAD-EMPATHY	189 ± 14	0,16	0,02	232	7%
SHOW-INTEREST	210 ± 26	0,21	0,16	5093	14%
SURPRISE	203 ± 26	- 0,01	- 0,42	185	5%
THANKS	202 ± 12	0,03	- 0,16	144	1%
WAIT	197 ± 28	- 0,34	- 0,49	190	6%
NEUTRAL	206 ± 41	0,10	- 0,37	19 002	5%

rovat i rozdíly mezi jednotlivými dialogovými akty znázorněné pomocí tzv. boxplotu.

U fonému */e/* lze pro střední hodnotu $F0$ u dialogového aktu *CONFIRM* pozorovat velký rozdíl oproti ostatním dialogovým aktů. Ten si lze vysvětlit například chybou výpočtu $F0$, přestože počet analyzovaných segmentů pro foném */e/* u tohoto dialogového aktu je poměrně velký. Dalším vysvětlením může být také chyba způsobená nepřesnou segmentací. Tento jev by jistě stál za další zkoumání. Pro skupinu všech fonémů pak je tento rozdíl výrazně menší, avšak z poměru odstraněných outlierů (29% pro metodu TT, 12% pro metodu WILKS) je patrné, že velká část dat byla právě pomocí tohoto mechanismu odstraněna. Je tedy nasnadě vyslovit hypotézu, že většina odstraněných outlierů ze souboru dat všech znělých fonémů byly právě hodnoty reprezentující foném */e/*. Toto zjištění přispívá k tomu, abychom se snažili využít co největšího počtu dat (fonémů) k získání reprezentativních výsledků a ke konstrukci akustické penalizační matice. Různé chyby totiž

mohou být při velkém množství dat odstraněny právě eliminací outlierů, při malém množství dat bychom získali nereprezentativní výsledky.

Pokud budeme zkoumat podobnost statistických charakteristik pro dialogový akt *NOT-SPECIFIED* a *NEUTRAL*, které by měly reprezentovat neutrální řeč, objevíme sice nevelké, ale přesto postřehnutelné rozdíly. Pokud se tyto rozdíly projeví i u dalších zkoumaných akustických charakteristik, bude nutné tyto dva dialogové akty rozlišovat a nebudeme moci řeč označenou dialogovým aktem *NOT-SPECIFIED* považovat za neutrální.



Obrázek 7.3: Porovnání histogramů hodnot $F0$ všech znělých fonémů bez odstranění outlierů (vlevo) a při použití metody TT pro odstranění outlierů (vpravo). Hodnoty jsou uvedené v [Hz].

Jaký vliv má odstranění outlierů na rozložení hodnot $F0$ u všech znělých fonémů pak demonstruje histogram na obrázku 7.3, kde červená křivka označuje, jaké rozložení četností by měla tato veličina, pokud by měla normální rozdělení pravděpodobnosti s danou střední hodnotou a směrodatnou odchylkou. Lze vidět, že po odstranění outlierů metodou TT je rozdělení pravděpodobnosti pro hodnoty $F0$ a pro dostatečně reprezentované dialogové akty (z hlediska počtu dostupných segmentů) téměř normální. Poněkud odlišná je situace pro metodu WILKS (jejíž histogram je zobrazen v příloze F na obrázku F.2), která neodstraňuje takové množství outlierů a více tak

zachovává původní rozdělení pravděpodobnosti. Tento fakt je patrný i z uvedených koeficientů šikmosti a špičatosti v tabulce F.3, kde se jejich hodnoty odklánějí od nulových hodnot platných pro normální rozdělení výrazně více, než je tomu v případě metody TT.

7.3.3 Analýza F0 pro věty

Pro akustickou analýzu $F0$ pro celé věty platí podobný postup jako v případě zkoumání hodnot $F0$ pro fonémy s tím rozdílem, že tentokrát jsme vypočítali průměrné, maximální a minimální hodnoty $F0$ pro celé věty a také rozsah hodnot $F0$ (tedy rozdíl maximální a minimální hodnoty pro každou větu). Do výpočtu samozřejmě vstupovaly pouze znělé části řeči (pro skupinu všech znělých fonémů) nebo pouze fonémy */e/*. V příloze F jsou v tabulkách F.4 – F.7 zobrazeny stručné výsledky této analýzy dosažené s využitím metody TT pro odstranění outlierů.

Za povšimnutí stojí poměr odstraněných outlierů pro rozsah hodnot $F0$ v rámci vět pro fonémy */e/* u dialogového aktu *CONFIRM*, který činí 97%. Je to způsobeno tím, že většina hodnot byla nulových. V těchto především krátkých větách se totiž foném */e/* vyskytoval velmi málo, většinou pouze jednou, tedy průměrná hodnota tohoto jediného fonému ve větě tvořila zároveň průměrnou hodnotu celé věty, jakožto i maximální a minimální hodnotu (je vidět i z tabulek). Rozsah hodnot $F0$ ve větě pro tento foném, jako rozdíl maximální a minimální hodnoty, byl tedy nulový. I přesto, že je to tedy správný údaj, nemůžeme ho považovat za dostatečně reprezentativní a jeho odstranění jako outlier je korektní. Zde se tedy projevil problém nedostatečného počtu dat.

Stejně jako u hodnot $F0$ pro fonémy, i u těchto akustických parametrů se pochopitelně projevila pravděpodobná chyba při určování $F0$ pro fonémy */e/* u dialogového aktu *CONFIRM*. Průměrné, minimální i maximální hodnoty pro věty počítané z fonému */e/* se výrazně odlišují od hodnot pro ostatní dialogové akty. Nicméně v tomto případě ani nedošlo k nějakému výraznému odstranění outlierů z původního souboru dat všech znělých fonémů (s výjimkou rozsahu $F0$), jako tomu bylo u $F0$ pro fonémy.

Jinak si lze povšimnout poměrně výrazné variability těchto akustických parametrů v rámci jednotlivých dialogových aktů, nicméně zde bychom si nedovolili vytvářet žádné další závěry, neboť počet vět v korpusu byl pro většinu dialogových aktů nereprezentativní, maximálně v řádu desítek či několika stovek (samozřejmě s výjimkami více zastoupených dialogových aktů). Avšak rozdíly mezi jednotlivými dialogovými akty u těchto akustických parametrů jsou v souladu s rozdíly naměřenými pro fonémové hodnoty $F0$, jsou s nimi tedy podle očekávání výrazně korelované. Z těchto důvodů tyto

naměřené charakteristiky nebudeme při tvorbě akustické penalizační matice využívat a zde je uvádíme spíše pro ilustraci.

7.3.4 Analýza doby trvání

Určení doby trvání fonémů bylo provedeno na základě výsledků automatické segmentace korpusu [72] s využitím nástroje HTK (*Hidden Markov Model Toolkit [131]*). Každému fonému je přiřazen čas začátku a čas konce v rámci promluvy, dobu trvání lze pak jednoduše určit odečtením těchto dvou hodnot. Analýze jsme podrobili segmenty odpovídající jak všem fonémům, tak fonému */e/*. Výsledky pro metodu TT jsou uvedeny v tabulkách 7.8 a 7.9, pro metodu WILKS pak v příloze F v tabulce F.8.

Tabulka 7.8: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro dobu trvání všech fonémů, metoda odstranění outlierů TT.

	Všechny fonémy				
Dialogový akt	$\mu \pm\sigma$ [ms]	γ_1	γ_2	N_P	P_O
APOLOGY	79 ± 28	0,75	0,14	1189	7%
CONFIRM	97 ± 35	0,63	- 0,17	4635	14%
DIRECTIVE	69 ± 20	0,37	- 0,10	3003	19%
DISCONFIRM	99 ± 49	0,93	- 0,00	145	6%
ENCOURAGE	80 ± 25	0,48	- 0,23	47 224	9%
GOODBYE	72 ± 24	0,33	- 0,73	1743	8%
GREETING	75 ± 27	0,73	- 0,25	2732	10%
HAPPY-EMPATHY	79 ± 26	0,46	- 0,26	6891	12%
NOT-SPECIFIED	72 ± 22	0,42	- 0,29	5465	11%
REQUEST	70 ± 20	0,67	- 0,05	6049	19%
SAD-EMPATHY	78 ± 26	0,47	- 0,33	2408	14%
SHOW-INTEREST	79 ± 24	0,47	- 0,33	44 394	11%
SURPRISE	78 ± 24	0,38	- 0,14	1870	11%
THANKS	91 ± 34	0,50	- 0,81	989	5%
WAIT	70 ± 25	0,21	- 0,45	2033	4%
NEUTRAL	78 ± 24	0,44	- 0,31	218 700	10%

Pro lepší přehlednost dosažených výsledků jsme opět využili boxplot, který je zobrazen v příloze F na obrázku F.3. Lze opět snadno vidět výrazný rozdíl pro dialogový akt *CONFIRM*, především u fonému */e/*. S ohledem

Tabulka 7.9: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro dobu trvání fonému */e/*, metoda odstranění outlierů TT.

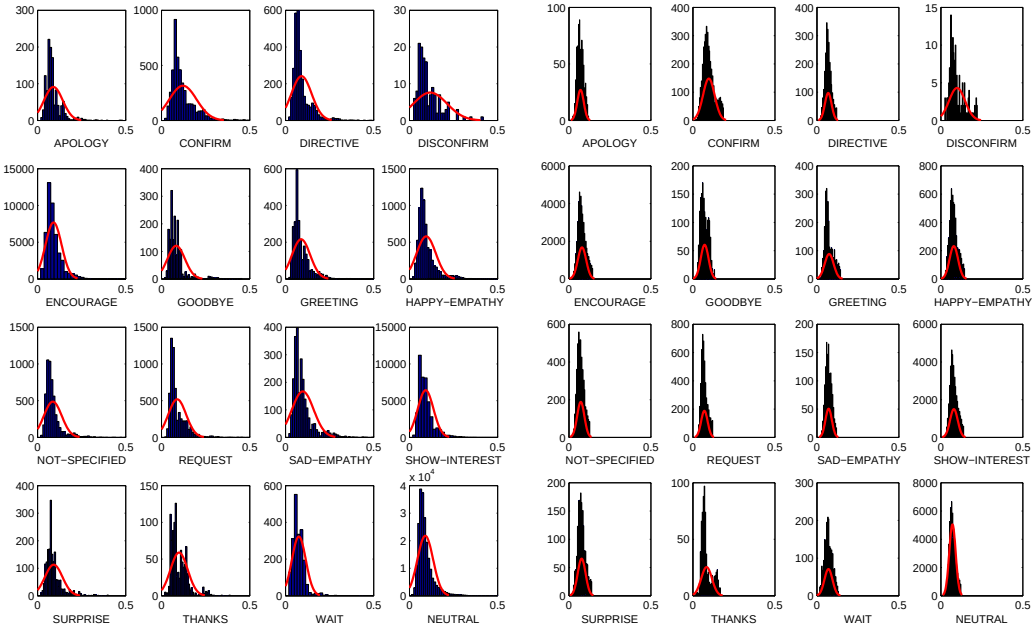
	Foném /e/				
Dialogový akt	$\mu \pm\sigma$ [ms]	γ_1	γ_2	N_P	P_O
APOLOGY	65 ± 13	- 0,08	- 0,73	131	5%
CONFIRM	144 ± 40	- 0,74	0,04	355	2%
DIRECTIVE	61 ± 15	- 0,40	- 0,33	475	13%
DISCONFIRM	102 ± 58	0,80	- 0,81	28	7%
ENCOURAGE	63 ± 14	0,20	- 0,37	6227	16%
GOODBYE	51 ± 12	- 0,11	- 0,94	158	1%
GREETING	92 ± 51	1,25	- 0,04	230	1%
HAPPY-EMPATHY	65 ± 17	0,15	- 0,48	904	11%
NOT-SPECIFIED	61 ± 15	0,35	- 0,49	375	11%
REQUEST	60 ± 12	0,00	- 0,36	880	9%
SAD-EMPATHY	60 ± 16	0,21	- 0,36	232	7%
SHOW-INTEREST	70 ± 16	0,11	- 0,20	5093	14%
SURPRISE	67 ± 17	0,24	- 0,17	185	9%
THANKS	70 ± 13	0,26	- 0,25	144	1%
WAIT	49 ± 15	0,43	- 0,71	190	4%
NEUTRAL	67 ± 15	0,24	- 0,24	19 002	10%

na tyto výsledky se nám jeví jako nejpravděpodobnější varianta vysvětlení takové odlišnosti pro tento dialogový akt u fonému */e/* nepřesná segmentace.

Jak je vidět z výsledků (uvažujme metodu TT pro odstranění outlierů), doba trvání fonémů je u dialogového aktu *NEUTRAL*, reprezentujícího neutrální řeč, rovna 78 ms. Pokud bychom měli porovnat naše výsledky s jinými analýzami, například [57] nebo [134], dojdeme k poněkud překvapivému zjištění. V uvedených studiích byly vyvozeny závěry, že doba trvání fonémů ve smutných větách je delší než pro neutrální věty, a ve veselých větách naopak kratší. K porovnání s námi dosaženými výsledky tedy můžeme použít dialogové akty *SAD-EMPATHY* a *HAPPY-EMPATHY*, které by jistým způsobem mohly zastupovat smutné, resp. veselé věty[7]. Pro tyto dialogové akty však závěry prezentované v uvedených studiích neplatí. Doba trvání fonémů pro dialogový akt *SAD-EMPATHY* je totiž 78 ms a pro *HAPPY-EMPATHY*

[7]Pravdou ovšem je, že v případě dialogových aktů se jedná o expresivní nahrávky pořízené formou dialogů (scénářů), zatímco u většiny emotivních vět se jedná o hrané či předstírané emoce. To samozřejmě také může porovnání ovlivňovat.

79 ms, což jsou hodnoty srovnatelné s neutrální řečí (pokud využijeme metodu WILKS pro odstranění outlierů dojdeme k podobnému závěru). To by mohlo ukazovat na mírnou odlišnost českého jazyka od ostatních jazyků, pro které byly tyto studie provedeny, nebo na specifickou vlastnost našeho řečníka. Nicméně hypotézu o vlastnosti řečníka může částečně vyvrátit fakt, že k podobnému závěru jako v této práci (i když ne naprosto shodnému) jsme dospěli i my v předchozí práci [44], resp. [41] pro jiného řečníka, kde se však skutečně jednalo o analýzu emotivní řeči a ne expresivní řeči označené pomocí dialogových aktů. I tam se projevil určitý rozpor s citovanými studiemi.



Obrázek 7.4: Histogram pro hodnoty doby trvání všech fonémů bez odstranění outlierů (vlevo) a při použití metody TT pro odstranění outlierů (vpravo). Hodnoty jsou uvedené v [s].

Na závěr analýzy doby trvání fonémů ještě uvedeme, podobně jako u dalších analýz, histogram reprezentující rozložení četnosti výskytu jednotlivých hodnot tohoto akustického parametru. Na obrázku 7.4 můžeme opět pozorovat, jaký vliv má na rozložení četnosti předzpracování dat, tedy odstranění odlehlých hodnot. Červená křivka opět ukazuje, jak by vypadalo normální rozdělení pravděpodobnosti s danou střední hodnotou a směrodatnou odchylkou. Je zřejmé, že ani předzpracované hodnoty tohoto akustického parametru se normálním rozdělením neřídí, což je patrné i z hodnot koeficientů šikmosti a špičatosti uvedených v tabulce 7.8. Výjimkou pak zřejmě je neutrální řeč.

7.3.5 Analýza RMS

Hodnotu RMS[8] (zkratka z anglického *Root Mean Square*, kvadratická střední hodnota) akustického signálu lze počítat podle rovnice

$$RMS = \sqrt{\frac{\sum_{i=1}^{n} s(i)^2}{n}}, \tag{7.1}$$

kde $s(i)$ je i-tý vzorek signálu a n počet vzorků. Pro každý analyzovaný segment se hodnota RMS vypočítá z celého jeho signálu.

Získané výsledky analýzy RMS pro metodu odstranění outlierů TT jsou uvedeny v tabulkách 7.10 a 7.11, pro metodu WILKS pak v příloze F v tabulce F.9. Boxplot je zobrazen také v příloze F na obrázku F.4. I pro tento akustický parametr lze sledovat mírné rozdíly mezi jednotlivými dialogovými akty. Provedený test ANOVA prokázal, že rozdíly jsou statisticky významné (p-hodnota byla téměř nulová, hladina významnosti $\alpha = 0{,}05$).

Velký rozdíl (zejména pro foném */e/*) u dialogového aktu *NEUTRAL*, reprezentujícího neutrální řeč, lze vysvětlit odlišným nastavením nahrávacího zařízení. Jak již bylo uvedeno dříve, mezi nahráváním neutrálního a expresivního korpusu byla, byť krátká, časová prodleva a z technických důvodů se nastavení z předchozího nahrávání nezachovalo. Opět můžeme pozorovat velký rozdíl u střední hodnoty fonému */e/* pro dialogový akt *CONFIRM*, stejně jako u předchozích akustických parametrů.

Na obrázku 7.5 je vidět, že pro akustický parametr hodnota RMS nebyla zřejmě metoda TT pro detekci odlehlých hodnot příliš úspěšná[9]. Je to patrné z počtu odstraněných outlierů, když jen nepatrné procento z celého souboru dat bylo pomocí tohoto algoritmu odstraněno.

7.3.6 Analýza formantů

Hodnoty formantových frekvencí jednotlivých českých samohlásek byly získány pomocí nástroje Speech Filing System (SFS)[10], konkrétně programem `formanal`. Ten je v současnosti označován jako nejlepší prostředek pro získání formantových frekvencí v rámci nástroje SFS[11]. Pro další analýzu jsme

[8] Statistický nástroj pro měření amplitudy veličiny, jejíž hodnota se v čase mění. Obzvláště užitečné, pokud veličina nabývá jak kladných, tak záporných hodnot, např. právě řečový signál.

[9] Podobný výsledek platí i pro metodu WILKS, při jejímž použití nebyly dokonce detekovány téměř žádné odlehlé hodnoty.

[10] Speech Filing System – http://www.phon.ucl.ac.uk/resource/sfs.

[11] Samozřejmě existují i další počítačové programy pro extrakci formantů, například hojně využívaný PRAAT [9].

Tabulka 7.10: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty RMS všech fonémů, metoda odstranění outlierů TT.

	Všechny fonémy				
Dialogový akt	$\mu \pm \sigma$	γ_1	γ_2	N_P	P_O
APOLOGY	0,19 ±0,10	- 0,00	- 0,69	1189	0,5%
CONFIRM	0,17 ±0,11	0,47	- 0,63	4878	1,7%
DIRECTIVE	0,16 ±0,10	0,32	- 0,88	3006	1,3%
DISCONFIRM	0,17 ±0,09	0,09	- 0,85	147	1,4%
ENCOURAGE	0,17 ±0,10	0,12	- 0,86	47 292	0,6%
GOODBYE	0,18 ±0,10	- 0,02	- 0,66	1744	0,1%
GREETING	0,15 ±0,08	0,11	- 0,60	2734	1,9%
HAPPY-EMPATHY	0,17 ±0,10	0,11	- 1,01	6896	0,1%
NOT-SPECIFIED	0,16 ±0,10	0,22	- 0,82	5468	0,7%
REQUEST	0,19 ±0,10	0,03	- 0,89	6058	0,3%
SAD-EMPATHY	0,19 ±0,11	0,20	- 0,86	2408	0,3%
SHOW-INTEREST	0,18 ±0,10	0,11	- 0,94	44 481	0,4%
SURPRISE	0,19 ±0,11	0,05	- 0,90	1874	0,4%
THANKS	0,19 ±0,09	0,06	- 1,03	989	0,1%
WAIT	0,14 ±0,09	0,27	- 0,90	2037	0,6%
NEUTRAL	0,13 ±0,08	0,23	- 0,76	218 700	1,4%

využili hodnot pro první tři formanty ($F1, F2, F3$) všech českých krátkých samohlásek (*/a/*, */e/*, */i/*, */o/*, */u/*).

Při odstraňování outlierů pro soubor dat obsahující hodnoty formantů jsme postupovali poněkud odlišným způsobem, než v předchozích případech. Využili jsme sice metodu TT pro každý formant zvlášť, ale posléze jsme ze souboru dat odstranili hodnoty všech formantů každého segmentu (fonému), který byl alespoň v jedné formantové frekvenci označen jako outlier. V tom případě je totiž možné, že i ostatní formanty budou nepřesné, neboť jsme si ověřili, že v případě špatné detekce jednoho z formantů bývají chybně detekované frekvence i ostatních[12]. Je pravděpodobné, že by takto špatně přiřazené formantové frekvence byly v rámci detekce outlierů stejně odstraněny, nicméně nám přišlo vhodné takové situaci předejít a vyvarovat se tak zbytečných chyb. Výsledkem je samozřejmě i to, že poměr odstraněných outlierů je pro každý formant totožný.

[12]Pokud je například špatně určený formant $F1$, je pravděpodobné, že skutečná hodnota formantu $F1$ bude přiřazena formantu $F2$. Tento problém popisuje a řeší např. [39].

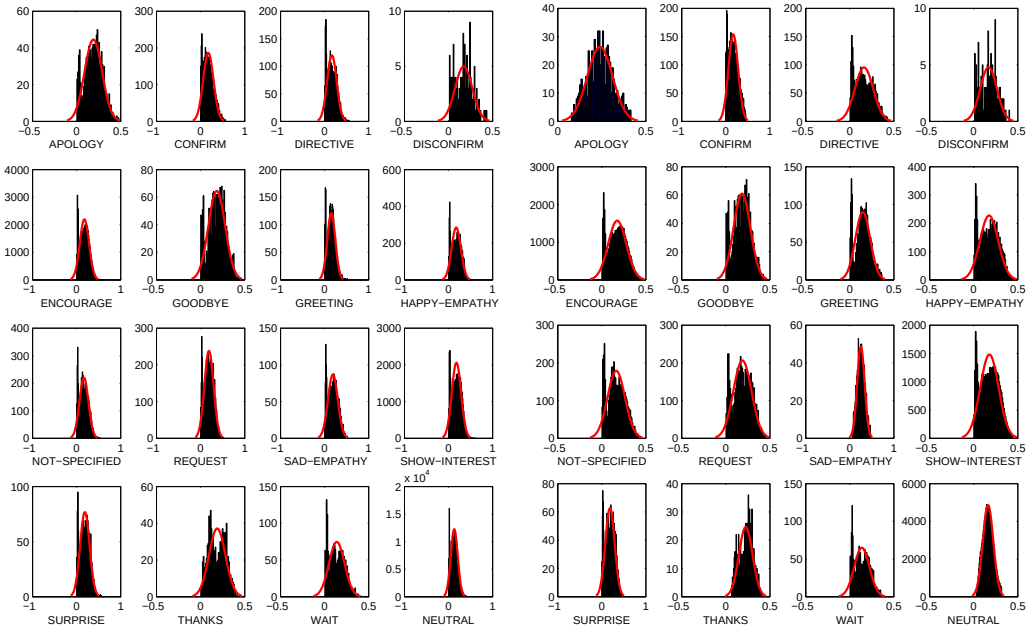
Tabulka 7.11: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty RMS fonému */e/*, metoda odstranění outlierů TT.

	Foném *e*				
Dialogový akt	$\mu \pm \sigma$	γ_1	γ_2	N_P	P_O
APOLOGY	0,27 ±0,06	0,04	- 0,57	131	0,8%
CONFIRM	0,12 ±0,06	0,73	- 0,39	355	5,9%
DIRECTIVE	0,24 ±0,09	- 0,38	- 0,10	475	0,2%
DISCONFIRM	0,23 ±0,08	0,30	- 0,58	28	0,0%
ENCOURAGE	0,24 ±0,07	- 0,11	- 0,43	6227	0,3%
GOODBYE	0,27 ±0,06	0,40	- 0,28	158	0,6%
GREETING	0,23 ±0,07	0,52	- 0,71	230	0,0%
HAPPY-EMPATHY	0,24 ±0,08	- 0,39	- 0,31	904	0,0%
NOT-SPECIFIED	0,26 ±0,07	- 0,40	- 0,35	375	2,7%
REQUEST	0,27 ±0,06	- 0,15	- 0,24	880	1,9%
SAD-EMPATHY	0,27 ±0,08	0,06	- 0,44	232	0,0%
SHOW-INTEREST	0,25 ±0,08	- 0,27	- 0,49	5093	0,2%
SURPRISE	0,26 ±0,06	- 0,07	- 0,68	185	1,1%
THANKS	0,28 ±0,06	0,05	- 0,54	144	0,0%
WAIT	0,24 ±0,06	0,27	- 0,40	190	1,6%
NEUTRAL	0,18 ±0,06	0,05	- 0,13	19 002	1,8%

Výsledky formantové analýzy jsou pro jejich rozsáhlost uvedeny v příloze F. Hodnoty uvedené v tabulkách F.10 – F.14 reprezentují střední hodnoty a směrodatné odchylky formantových frekvencí, kdy pro každý foném byly formantové frekvence určeny z prostřední části fonému (nezapočítávaly se hodnoty naměřené v první a poslední čtvrtině daného fonému). Tím by měl být jejich odhad přesnější a neovlivněný tzv. *tranzienty*, tedy přechody mezi předcházejícím a následujícím fonémem. Lze si povšimnout, že pro foném */u/* nebyl nalezen žádný segment reprezentující dialogový akt *GOODBYE* (proto nejsou pro tento dialogový akt v tabulce uvedené žádné hodnoty) a pouze jeden segment reprezentující *DISCONFIRM*.

Z naměřených hodnot jsou patrné rozdíly (někdy i velmi výrazné) mezi jednotlivými dialogovými akty. Pro lepší znázornění těchto rozdílů jsme ještě v příloze F na obrázcích F.5 – F.9 zobrazili získané hodnoty v rovině $F1 \times F2$, kde na ose x jsou hodnoty formantových frekvencí $F1$, na ose y pak $F2$.

Na obrázcích F.5 – F.9 v příloze F lze také pozorovat výraznou odlišnost dialogového aktu *NEUTRAL* od ostatních. To přikládáme již zmíněnému



Obrázek 7.5: Histogram pro hodnoty RMS všech fonémů bez odstranění outlierů (vlevo) a při použití metody TT pro odstranění outlierů (vpravo).

faktu, že neutrální korpus a expresivní korpus byly nahrávány s mírným časovým rozestupem, lehce odlišným nastavením nahrávacího systému a samozřejmě se změněnou hlasovou dispozicí řečníka, což mohlo způsobit mírnou změnu barvy hlasu.

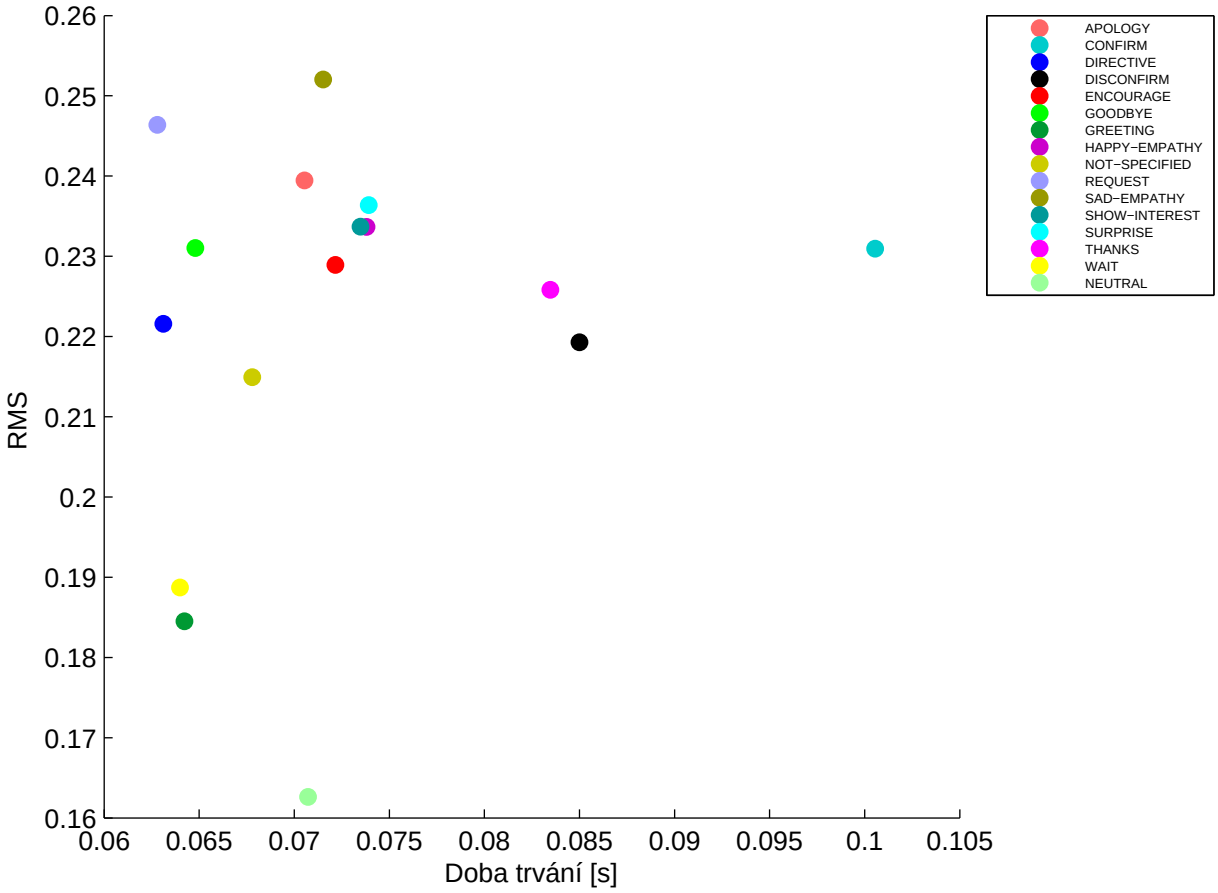
7.3.7 Shrnutí akustické analýzy

V předcházejících částech byly prezentovány výsledky rozsáhlé akustické analýzy neutrálních i expresivních vět dle jednotlivých dialogových aktů. V následujících částech je stručně shrneme a ukážeme, že výsledky získané pro velkou skupinu fonémů nejsou příliš ovlivněny fonetickou nevyvážeností expresivního korpusu. Dále také ukážeme naměřené korelace mezi jednotlivými akustickými parametry.

Stručné shrnutí

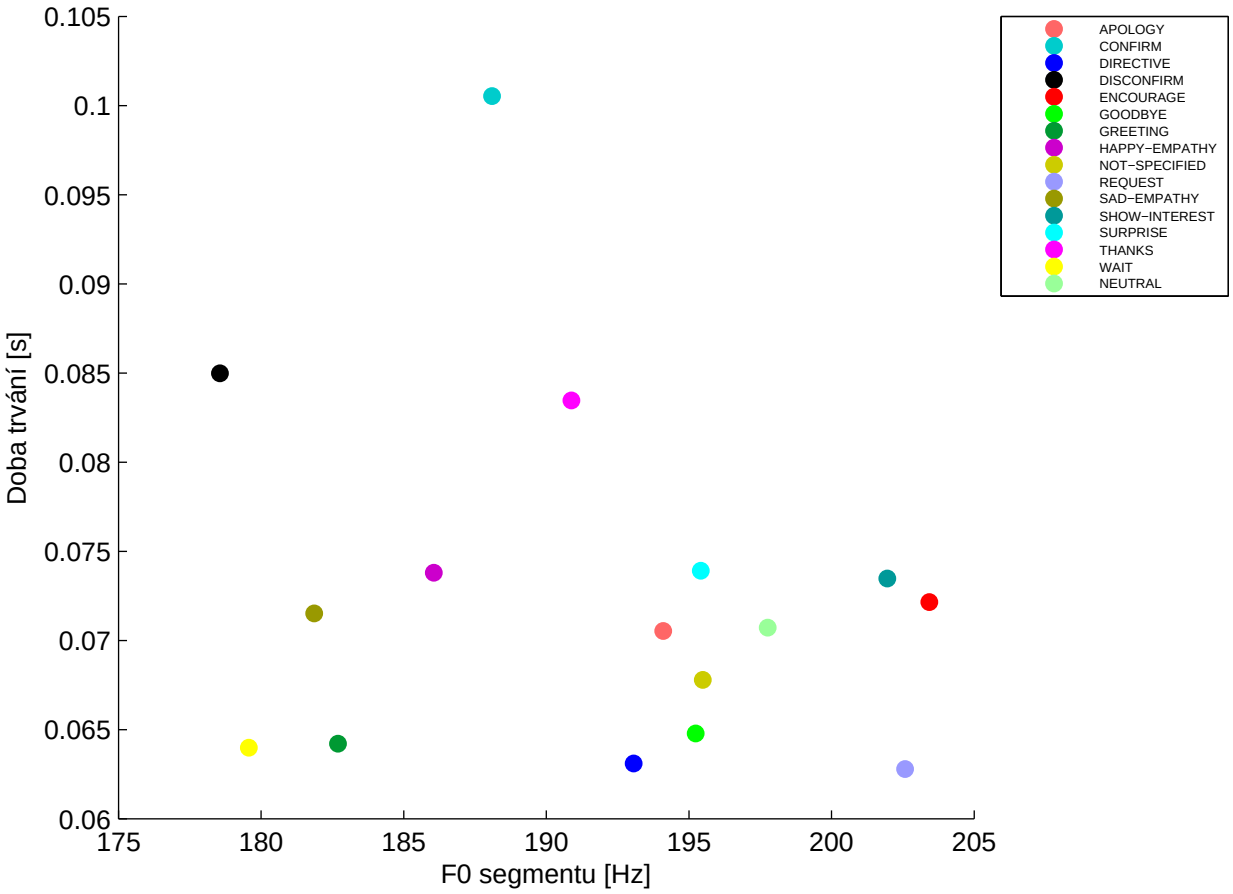
Na obrázcích 7.6 – 7.8 jsou graficky zobrazeny střední hodnoty pro tři základní akustické parametry ($F0$ pro fonémy, doba trvání, RMS) ve třech

různých rovinách – doba trvání × RMS, $F0$ fonému × doba trvání a $F0$ fonému × RMS – pro všechny znělé fonémy po odstranění outlierů metodou TT.



Obrázek 7.6: Zobrazení středních hodnot akustických parametrů všech znělých fonémů v rovině doba trvání × RMS.

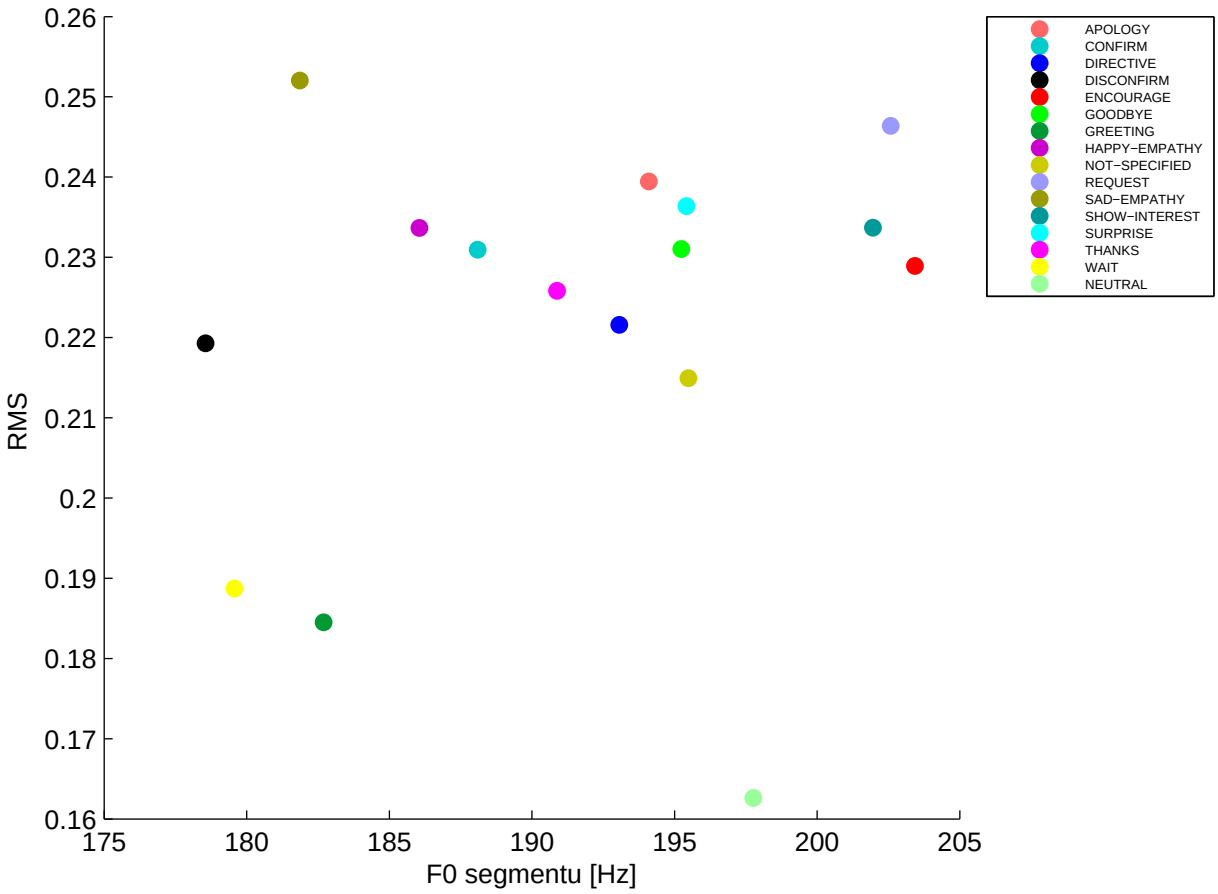
Z vyobrazených dat lze snadno pozorovat rozdíly mezi dialogovými akty, v některé rovině větší, v jiné menší. Pokud však budeme pohlížet na střední hodnoty těchto tří základních akustických parametrů jako na vektor určující polohu dialogového aktu v třírozměrném prostoru se systémem souřadnic reprezentovaným právě těmito akustickými parametry, rozdíly budou jistě patrné. Významnost rozdílů pro jednotlivé akustické parametry byla vždy



Obrázek 7.7: Zobrazení středních hodnot akustických parametrů všech zněných fonémů v rovině $F0$ fonému $\times$ doba trvání.

testována pomocí analýzy ANOVA s hladinou významnosti $\alpha = 0{,}05$ a to i přes všechny možné dvojice dialogových aktů[13]. Mezi dvojicemi pak v některých případech byla pro různé akustické parametry zjištěna statistická nevýznamnost rozdílu středních hodnot. Naměřené p-hodnoty jsou zobrazeny v tabulkách F.15 – F.17 v příloze F, kde zvýrazněné hodnoty (větší

[13]Při testování přes všechny dialogové akty jsme vždy dospěli k výsledku, že rozdíly mezi dialogovými akty jsou významné, což však také mohlo být způsobeno poměrně velkým počtem dialogových aktů a analyzovaných segmentů v rámci jednotlivých dialogových aktů.



Obrázek 7.8: Zobrazení středních hodnot akustických parametrů všech zněl-ých fonémů v rovině $F0$ fonému $\times$ RMS.

než 0,05) značí statistickou nevýznamnost rozdílu mezi středními hodnotami pro danou dvojici dialogových aktů.

Právě rozdílů mezi vektory reprezentujícími různé dialogové akty[14] bude využito při výpočtu akustické části penalizační matice, která je popsána v části 8.1.2. Tyto rozdíly totiž budou určitým způsobem částečně definovat penalizaci jednotek označených jiným dialogovým aktem, než který bude vyžadován pro syntézu expresivní řeči.

[14]Vektory reprezentující dialogové akty však budou tvořeny poněkud odlišně a budou více než třírozměrné.

Korelace mezi skupinami fonémů

Analýzy pro všechny dříve uvedené akustické parametry, z nichž některé budou dále využity pro tvorbu penalizační matice (viz část 8.1.2), byly provedeny jak pro velkou skupinu fonémů (všechny fonémy a znělé fonémy, dle akustického parametru), tak i pro jeden vybraný foném */e/*. Naším cílem bylo zjistit, zda data získaná pro velkou skupinu fonémů jsou dostatečně reprezentativní napříč všemi dialogovými akty. Expresivní korpus totiž obsahuje věty foneticky nevyvážené, tj. věty zastupující dialogový akt X mohou obsahovat jinou skladbu fonémů než věty zastupující dialogový akt Y, což může zkreslovat dosažené výsledky. Analýzou pouze fonému */e/* jsme tento nedostatek nevyváženosti eliminovali a zajímá nás, zda jsou výsledky pro tento foném v souladu s výsledky pro velkou skupinu fonémů. K tomu jsme využili korelační koeficienty, jejichž hodnoty porovnávající výsledky všech dialogových aktů pro různé akustické parametry jsou uvedeny v tabulce 7.12.

Tabulka 7.12: Korelační koeficienty mezi velkou skupinou fonémů a fonémem */e/* napříč všemi dialogovými akty pro různé akustické parametry.

akustický parametr	porovnávaná skupina fonémů	korelace	korelace (bez *CONFIRM*)
doba trvání	všechny	0,76	0,68
RMS	všechny	0,52	0,83
F0 pro fonémy	znělé	0,43	0,89
F0 pro věty	znělé	0,64	0,86
F0 max pro věty	znělé	0,59	0,85
F0 min pro věty	znělé	0,50	0,76
F0 rozsah pro věty	znělé	0,79	0,86
střední hodnota		**0,60**	**0,82**

Byla zjištěna poměrně vysoká korelace ($\rho = 0{,}60$) mezi výsledky pro velkou skupinu fonémů a výsledky získanými pouze pro foném */e/*. Pokud nebudeme započítávat výsledky pro dialogový akt *CONFIRM*, jehož výsledky akustických analýz pro foném */e/* byly zřejmě v některých případech ovlivněny špatnou segmentací a v některých i nedostatečným počtem dat (viz část 7.3), dojdeme k hodnotě korelace $\rho = 0{,}82$, což již značí vysokou míru lineární závislosti mezi naměřenými hodnotami.

Z výše uvedeného lze tedy odvodit, že výsledky získané pro velkou skupinu fonémů jsou dostatečně reprezentativní, a fakt, že expresivní korpus je v rámci různých dialogových aktů foneticky nevyvážený, by nám do dal-

šího využití výsledků nemusel zanášet nikterak velkou chybu. Získáváme tak výhodu velkého počtu dat, což by mělo přispět k robustnosti celého postupu.

Korelace mezi akustickými parametry

Zajímavým údajem může také být korelace mezi jednotlivými akustickými parametry, resp. jejich středními hodnotami. Tyto korelační koeficienty jsou pro skupinu všech znělých fonémů uvedené v tabulce 7.13 (metoda odstranění outliers TT). Korelační koeficienty v absolutní hodnotě větší než 0,5 jsou zvýrazněné, měly by ukazovat na poměrně silnou lineární závislost těchto akustických parametrů.

Tabulka 7.13: Korelační koeficienty mezi jednotlivými akustickými parametry pro skupinu všech znělých fonémů.

akustický parametr	$F0$ (fonémy)	doba trvání	RMS	$F0$ (věty)	$F0_{\text{MAX}}$ (věty)	$F0_{\text{MIN}}$ (věty)	$F0_{\text{ROZSAH}}$ (věty)
$F0$ (fonémy)	**1,00**	−0,46	0,04	**0,95**	**0,51**	0,23	0,20
doba trvání	−0,46	**1,00**	0,16	**−0,55**	−0,31	0,25	−0,36
RMS	0,04	0,16	**1,00**	0,04	−0,40	**0,75**	**−0,70**
$F0$ (věty)	**0,95**	**−0,55**	0,04	**1,00**	**0,65**	0,21	0,31
$F0_{\text{MAX}}$ (věty)	**0,51**	−0,31	−0,40	**0,65**	**1,00**	−0,30	**0,83**
$F0_{\text{MIN}}$ (věty)	0,23	0,25	**0,75**	0,21	−0,30	**1,00**	**−0,78**
$F0_{\text{ROZSAH}}$ (věty)	0,20	−0,36	**−0,70**	0,31	**0,83**	**−0,78**	**1,00**

Lze si povšimnout samozřejmé závislosti průměrné $F0$ pro fonémy a průměrné $F0$ pro věty a dále pak také negativní korelace mezi průměrnou $F0$ pro věty a dobou trvání, což vyjadřuje rychlejší tempo řeči, pokud je zvýšená $F0$ (pro průměrnou $F0$ fonémů je tato korelace lehce pod prahem, který jsme stanovili). Zajímavá je také korelace mezi hodnotou RMS a minimální $F0$ pro věty, resp. negativní korelace s rozsahem $F0$ pro věty. Těchto výsledků nicméně v práci nijak nevyužíváme a uvádíme je spíše jen pro ilustraci závislostí mezi akustickými parametry.

Kapitola 8

Modifikace metody výběru jednotek

Z přirozených dialogů seniorů s počítačem jsme získali přehled o oblasti, ve které se bude náš (dialogový) systém syntézy expresivní řeči používat. Na jejich základě jsme také nadefinovali expresivní kategorie (dialogové akty). Tyto znalosti jsme pak využili pro nahrání expresivního řečového korpusu, který bude využit pro syntézu expresivní řeči. S využitím dialogových aktů jsme pomocí poslechového testu celý expresivní korpus nechali anotovat několika posluchači. Na základě těchto subjektivních anotací jsme objektivně každé větě z korpusu přiřadili jeden dialogový akt (viz definice objektivní anotace na straně 63). Takto popsaný korpus jsme podrobili rozsáhlé akustické analýze, která odhalila rozdíly různých akustických parametrů mezi definovanými dialogovými akty. Dalším naším úkolem tedy je takto získané informace využít při modifikování stávajícího systému syntézy neutrální řeči tak, aby byl schopný pro danou oblast (rozhovory seniorů o fotografiích) produkovat navíc i řeč expresivní.

8.1 Penalizační matice

Modifikace algoritmu výběru jednotek, který je použit v našem stávajícím systému, bude spočívat především v poněkud komplexnějším výpočtu ceny cíle, která je jedním z parametrů ovlivňující výběr jednotek z inventáře. Popis změn ve výpočtu pro expresivní syntézu uvádí část 8.2, avšak abychom mohli přistoupit k takovým změnám, je třeba nejprve určit a nadefinovat numerické rozdíly mezi jednotlivými expresivními kategoriemi (dialogovými akty). Rozhodli jsme se tedy nadefinovat tzv. *penalizační matici*, obecněji popsanou v části 4.1.2, která by tyto numerické rozdíly měla reprezentovat.

Podobná penalizační matice byla již v minulosti využita například pro určení rozdílů mezi takovými řečovými jednotkami, které byly označeny příznakem „zdůrazněné“ a takovými, které byly označené jako „nezdůrazněné“ [35]. Vzhledem k získaným znalostem a informacím jsme založili penalizační matici jak na rozdílném vnímání jednotlivých dialogových aktů posluchači při anotacích (*percepční část penalizační matice*), tak na rozdílných hodnotách různých akustických parametrů získaných prostřednictvím akustické analýzy (*akustická část penalizační matice*). Tyto dva faktory jsme pak vhodně zkombinovali a získali výslednou penalizační matici popsanou v části 8.1.3.

8.1.1 Percepční penalizační matice

Základem percepční části penalizační matice je tzv. matice záměn uvedená v tabulce 8.1. V této matici každý prvek a_{ij} reprezentuje, v kolika případech byla věta označená objektivní anotací (viz část 7.2) dialogovým aktem j, subjektivně anotována posluchači jako věta představující dialogový akt i. Měla by tedy vyjadřovat percepční podobnost jednotlivých dialogových aktů.

Největší hodnoty pro jednotlivé dialogové akty se v matici záměn samozřejmě vyskytují na diagonále a představují „úspěšnost“ anotátorů při anotacích (hodnoceno dle objektivní anotace). Průměrná úspěšnost, určená tímto způsobem, pak byla 68 %. Pokud budeme uvažovat jen čtyři nejčastěji se v korpusu vyskytující dialogové akty[1], úspěšnost by byla 75 %. Pohledem na matici záměn také zjistíme, že anotátoři nejvíce zaměňovali dialogový akt *ENCOURAGE* se *SHOW-INTEREST* a dialogový akt *APOLOGY* se *SAD-EMPATHY*. Vysoké hodnoty pak také zaznamenáme v řádku reprezentujícím objektivní anotaci dialogového aktu *NOT-SPECIFIED*, který byl poměrně často zaměňován s ostatními.

Abychom tuto matici záměn mohli využít ke stanovení percepčních vzdáleností mezi jednotlivými dialogovými akty, provedli jsme transformaci této matice podle rovnice 8.1 tak, abychom získali koeficienty p'_{ij}:

$$p'_{ij} = \mathrm{abs}(\log(\frac{a_{ij}}{\max_i})), \tag{8.1}$$

kde a_{ij} jsou koeficienty z matice záměn a $\max_i$ reprezentuje maximální hodnotu z a_{ij} pro pevné i a všechna j, tj. největší počet záměn dialogového aktu i s ostatními dialogovými akty (v našem případě to vždy byly hodnoty na diagonále matice záměn).

Pro případy, kdy logaritmus není definován (pokud je jeho argument roven nule, tedy žádná percepční podobnost mezi dialogovými akty nebyla

[1] *SHOW-INTEREST*, *ENCOURAGE*, *CONFIRM* a *HAPPY-EMPATHY*, viz tabulka 7.5.

Tabulka 8.1: Matice záměn jako základ pro určení percepční penalizační matice.

subjektivní \ objektivní	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	OTHER	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT
APOLOGY	290	1	7	6	0	34	0	1	7	3	0	160	2	1	0	2
CONFIRM	0	8423	1	5	2	0	0	181	85	6	0	123	31	26	2	3
DIRECTIVE	0	2	1760	1	130	0	8	3	21	2	47	4	11	0	0	2
DISCONFIRM	7	4	0	131	1	0	0	7	24	8	1	6	1	1	0	1
ENCOURAGE	0	105	43	5	17472	0	10	245	442	437	583	168	5571	566	1	52
GOODBYE	48	5	26	0	2	845	1	110	145	15	0	66	6	0	176	3
GREETING	0	1	0	0	1	0	1139	3	54	1	0	0	2	0	0	2
HAPPY-EMPATHY	28	936	2	32	306	24	0	4305	424	109	3	102	228	120	24	11
NOT-SPECIFIED	29	515	118	49	790	239	204	518	1831	293	206	230	727	246	160	68
OTHER	0	0	2	0	0	1	0	0	2	7	0	0	0	0	0	0
REQUEST	0	0	111	0	412	0	116	7	65	104	2747	4	92	0	0	4
SAD-EMPATHY	203	208	26	12	139	55	0	26	112	37	4	2020	181	39	0	3
SHOW-INTEREST	0	53	113	7	5754	0	0	101	1099	52	446	165	22410	236	0	29
SURPRISE	0	112	2	11	892	0	0	207	252	74	9	70	796	1004	1	4
THANKS	0	3	0	0	1	5	0	0	0	0	0	0	0	0	619	0
WAIT	3	1	111	0	32	0	0	1	1	0	49	1	9	0	0	448

nalezena), stanovili jsme koeficient $p'_{ij} = K$. Pro hodnotu konstanty K musí platit

$$K \geq \max_{\forall i,j}(\text{abs}(\log(\frac{a_{ij}}{\max_i})), \tag{8.2}$$

kde $\max_{\forall i,j}$ je maximum přes všechna i, j, pro která logaritmus definován je.

Tato konstanta slouží k definici největší možné vzdálenosti mezi dialogovými akty. V naší práci jsme hodnotu konstanty stanovili jako $K = 5$, což byla zhruba dvojnásobná hodnota oproti minimální možné hodnotě K dle rovnice 8.2.

Logaritmus byl ve výpočtu použit proto, abychom zdůraznili rozdíly mezi dialogovými akty a dále předpokládáme, že lidské vnímání (nejen) akustických signálů je také logaritmicky založeno (viz Weber-Fechnerův psychofyzikální zákon[2]).

Koeficienty p_{ij} výsledné percepční penalizační matice $\mathbf{P}$ jsme pak získali normalizací koeficientů p'_{ij} konstantou K, tedy

$$p_{ij} = \frac{p'_{ij}}{K}. \tag{8.3}$$

Výsledná percepční penalizační matice $\mathbf{P}$ je pak zobrazena v tabulce 8.2 a reprezentuje tedy percepční vzdálenost mezi jednotlivými dialogovými akty. Na diagonále této matice se objevují pouze nulové hodnoty, zatímco hodnoty 1 značí největší možnou vzdálenost mezi dialogovými akty.

Lze vidět, že minimální vzdálenosti se zachovaly mezi těmi dialogovými akty, které posluchači nejvíce zaměňovali při anotacích, což je i smyslem percepční matice. Pro dialogový akt *NOT-SPECIFIED* jsme pak detekovali nejnižší průměrnou vzdálenost od ostatních dialogových aktů (0,18), zatímco pro *THANKS* byla tato průměrná vzdálenost nejvyšší (0,84). Celková průměrná vzdálenost naměřená mezi dialogovými akty byla 0,54. Percepční penalizační matice není již z podstaty symetrická, je tedy třeba rozlišovat, co přesně jednotlivé koeficienty znamenají (sloupce reprezentují objektivní anotaci, řádky subjektivní anotace posluchačů).

Dialogový akt *NEUTRAL* samozřejmě nemůže být v percepční penalizační matici zahrnut vůbec, neboť neutrální věty neprošly stejnou anotací, jako věty z expresivního korpusu. Při konstrukci celkové penalizační matice se pak s tímto faktem budeme muset vypořádat (viz část 8.1.3).

[2]Weber-Fechnerův psychofyzikální zákon zjednodušeně říká: Mění-li se fyzikální podněty působící na naše smysly řadou geometrickou, vnímáme jejich změnu v řadě aritmetické [123], [92]. Což znamená, že změna fyziologického vjemu v je úměrná relativní změně jeho fyzikální příčiny p, platí tedy $dv = K \cdot \frac{dp}{p}$, odkud $v = K \cdot \ln \frac{p}{p_0}$, kde p_0 je referenční hodnota veličiny hodnotící příčiny vjemu.

Tabulka 8.2: Konečná percepční penalizační matice **P**.

subjektivní \ objektivní	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	OTHER	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	**str. hod.**
APOLOGY	0,00	0,49	0,32	0,34	1,00	0,19	1,00	0,49	0,32	0,40	1,00	0,05	0,43	0,49	1,00	0,43	**0,50**
CONFIRM	1,00	0,00	0,79	0,65	0,72	1,00	1,00	0,33	0,40	0,63	1,00	0,37	0,49	0,50	0,72	0,69	**0,64**
DIRECTIVE	1,00	0,59	0,00	0,65	0,23	1,00	0,47	0,55	0,38	0,59	0,31	0,53	0,44	1,00	1,00	0,59	**0,58**
DISCONFIRM	0,25	0,30	1,00	0,00	0,42	1,00	1,00	0,25	0,15	0,24	0,42	0,27	0,42	0,42	1,00	0,42	**0,47**
ENCOURAGE	1,00	0,44	0,52	0,71	0,00	1,00	0,65	0,37	0,32	0,32	0,30	0,40	0,10	0,30	0,85	0,51	**0,49**
GOODBYE	0,25	0,45	0,30	1,00	0,53	0,00	0,59	0,18	0,15	0,35	1,00	0,22	0,43	1,00	0,14	0,49	**0,44**
GREETING	1,00	0,61	1,00	1,00	0,61	1,00	0,00	0,52	0,26	0,61	1,00	1,00	0,55	1,00	1,00	0,55	**0,73**
HAPPY-EMPATHY	0,44	0,13	0,67	0,43	0,23	0,45	1,00	0,00	0,20	0,32	0,63	0,33	0,26	0,31	0,45	0,52	**0,40**
NOT-SPECIFIED	0,36	0,11	0,24	0,31	0,07	0,18	0,19	0,11	0,00	0,16	0,19	0,18	0,08	0,17	0,21	0,29	**0,18**
OTHER	1,00	1,00	0,11	1,00	1,00	0,17	1,00	1,00	0,11	0,00	1,00	1,00	1,00	1,00	1,00	1,00	**0,77**
REQUEST	1,00	1,00	0,28	1,00	0,16	1,00	0,27	0,52	0,33	0,28	0,00	0,57	0,30	1,00	1,00	0,57	**0,58**
SAD-EMPATHY	0,20	0,20	0,38	0,45	0,23	0,31	1,00	0,38	0,25	0,35	0,54	0,00	0,21	0,34	1,00	0,57	**0,40**
SHOW-INTEREST	1,00	0,53	0,46	0,70	0,12	1,00	1,00	0,47	0,26	0,53	0,34	0,43	0,00	0,40	1,00	0,58	**0,55**
SURPRISE	1,00	0,19	0,54	0,39	0,01	1,00	1,00	0,14	0,12	0,23	0,41	0,23	0,02	0,00	0,60	0,48	**0,40**
THANKS	1,00	0,46	1,00	1,00	0,56	0,42	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00	0,00	1,00	**0,84**
WAIT	0,43	0,53	0,12	1,00	0,23	1,00	1,00	0,53	0,53	1,00	0,19	0,53	0,34	1,00	1,00	0,00	**0,59**

8.1.2 Akustická penalizační matice

Akustická část penalizační matice je založená na výsledcích akustické analýzy popsané v části 7.3, které ukazují různé statistické charakteristiky různých akustických parametrů expresivního řečového signálu v návaznosti na anotace pomocí dialogových aktů.

Výběr akustických parametrů a jejich charakteristik

Při přípravě akustické penalizační matice jsme se rozhodli vzít v úvahu následující akustické parametry všech znělých fonémů:

- hodnoty $F0$ pro fonémy;
- hodnoty doby trvání pro fonémy;
- hodnoty RMS pro fonémy.

Tyto parametry jsme ze všech akustických parametrů vybrali ze dvou důvodů. Jednak jsou k dispozici pro všechny znělé fonémy, a jednak jsou určeny z dostatečného počtu dat. Hodnoty $F0$ pro věty (ať už maximální, minimální, či rozsah) nebudou využity, neboť počet prvků (vět), ze kterých byly tyto hodnoty získány, byl pro většinu dialogových aktů nízký a tudíž nelze tyto výsledky považovat za reprezentativní. Hodnoty formantových frekvencí jsou pak svázané s konkrétními fonémy a jsou tedy jen obtížně použitelné při tvorbě jediné penalizační matice. Důvodem využití pouze znělých fonémů i pro dobu trvání a hodnotu RMS[3] je fakt, že takto můžeme pro každý znělý foném získat vektor hodnot $[f0, dur, rms]$ reprezentující umístění daného fonému ve třírozměrném prostoru, se souřadnicemi $F0$ ($f0$), doba trvání (dur) a hodnota RMS (rms). Pro odstranění odlehlých hodnot jsme pak mohli využít metody WILKS pro vícerozměrné veličiny[4], popsané v příloze E.

Po odstranění vícerozměrných outlierů jsme pro uvedené akustické parametry využili následující statistické charakteristiky:

- střední hodnota μ;
- směrodatná odchylka σ;
- koeficient šikmosti γ_1;
- koeficient špičatosti γ_2.

[3] V části 7.3 byla provedena akustická analýza i pro skupinu všech fonémů a dále také i pro jednotlivé fonémy.

[4] Narozdíl od akustické analýzy, kde byly uvedeny především výsledky pro metodu odstranění outlierů TT, příp. i metodu WILKS pro jednorozměrné veličiny.

Tyto statistické charakteristiky[5] by měly popisovat rozdělení pravděpodobnosti hodnot akustických parametrů. Jednotlivé charakteristiky byly transformovány do intervalu $\langle 0, 1\rangle$ tak, jak bude vysvětleno dále, abychom se vyhnuli nerovnoměrným absolutním rozdílům jednotlivých hodnot[6]. Každý dialogový akt bude tedy reprezentován celkem 12 hodnotami (3 akustické parametry $\times$ 4 statistické charakteristiky).

Transformace statistických charakteristik

Abychom mohli získané výsledky akustické analýzy (statistické charakteristiky pro jednotlivé dialogové akty) využít pro tvorbu penalizační matice, je potřeba získaná data transformovat. Protože naším cílem je získat akustické rozdíly mezi daty reprezentujícími jednotlivé dialogové akty, je potřeba odstranit absolutní rozdíly mezi jednotlivými vypočtenými charakteristikami. Pro transformaci dat (vektoru hodnot $\mathbf{x}$) máme na výběr dva základní přístupy:

(a) **standardizace** (z-score) – vektor hodnot $\mathbf{x}$ se transformuje na vektor $\mathbf{x_s}$ podle vztahu

$$\mathbf{x}_s = \frac{\mathbf{x} - \hat{\mu_x}}{\hat{\sigma_x}}, \tag{8.4}$$

kde $\hat{\mu_x}$ je odhad střední hodnoty $\mathbf{x}$ a $\hat{\sigma_x}$ je odhad směrodatné odchylky. Výsledkem je vektor $\mathbf{x_s}$, jehož hodnoty mají střední hodnotu $\mu = 0$ a směrodatnou odchylku $\sigma = 1$.

(b) **normalizace** – vektor hodnot $\mathbf{x}$ se transformuje na vektor $\mathbf{x_n}$ podle vztahu

$$\mathbf{x}_n = \frac{\mathbf{x} - \min_x}{\max_x - \min_x}, \tag{8.5}$$

kde $\min_x$ je minimální hodnota $\mathbf{x}$ a $\max_x$ je maximální hodnota $\mathbf{x}$. Výsledkem je vektor $\mathbf{x_n}$, jehož hodnoty leží v intervalu $\langle 0, 1\rangle$.

Z výsledných hodnot budeme vycházet při vytváření penalizační matice, která bude následně využita při výpočtu ceny cíle v úloze syntézy expresivní řeči. Při výpočtu ceny cíle se bere v úvahu více příznaků jednotlivých řečových jednotek, přičemž rozsah hodnot těchto příznaků se v našem systému TTS pohybuje v intervalu $\langle 0, 1\rangle$ a konečná významnost daného příznaku je dána až následně použitými váhami (viz část 2.1). Abychom dodrželi tento

[5] Využili jsme tedy všechny statistické charakteristiky, které jsme určovali i při akustické analýze.

[6] Například rozdíl středních hodnot $F0$, který se pohybuje v řádu jednotek či desítek, by převážil rozdíl středních hodnot RMS, který se pohybuje v řádu setin či desetin.

systém, zvolili jsme metodu normalizace (b), která transformuje vstupní data na požadovaný interval. I přesto se však při výpočtu rozdílu (vzdálenosti) mezi jednotlivými dialogovými akty dostaneme mimo uvedený interval, neboť pro tento výpočet využijeme Euklidovu vzdálenost. Vypočtený rozdíl pak tedy bude třeba opět obdobným způsobem transformovat do patřičných mezí.

Výpočet koeficientů

Pro určení prvotních koeficientů a'_{ij} akustické penalizační matice $\mathbf{A}'$ (uvedené v příloze F v tabulce F.18) jsme využili Euklidovské vzdálenosti mezi transformovanými 12-rozměrnými vektory hodnot jednotlivých dialogových aktů, tedy

$$a'_{ij} = \mathbf{d}_i - \mathbf{d}_j, \tag{8.6}$$

kde $\mathbf{d}_i$ představuje 12-rozměrný vektor čtyř statistických charakteristik pro tři různé akustické parametry dialogového aktu i.

Přestože jsme transformovali statistické charakteristiky do intervalu $\langle 0, 1\rangle$, po aplikování Euklidovské vzdálenosti se hodnoty rozdílů mezi dialogovými akty ocitly mimo tento interval. Musíme tedy koeficienty a'_{ij} matice $\mathbf{A}'$ opět transformovat. To provedeme podle rovnice

$$a_{ij} = \frac{a'_{ij}}{\max_{ij}}, \tag{8.7}$$

kde $\max_{ij}$ je maximální hodnota vyskytující se v matici $\mathbf{A}'$, v našem případě $\max_{ij} = 2{,}36$. Hodnotu minima dle rovnice 8.5 aplikovat nemusíme, neboť minimální hodnota ze všech koeficientů a'_{ij} je vždy nulová[7].

Transformované koeficienty a_{ij} tvořící výslednou akustickou penalizační matici $\mathbf{A}$ jsou zobrazeny v tabulce 8.3.

Největší průměrnou vzdálenost od ostatních dialogových aktů jsme naměřili pro dialogový akt *OTHER* (0,77), nejmenší pak pro *SURPRISE* (0,36). Průměrná vzdálenost mezi dialogovými akty pak byla 0,48. Do akustické penalizační matice jsme již mohli zahrnout i dialogový akt *NEUTRAL*, protože akustické parametry neutrálních vět, které tento dialogový akt zastupují, můžeme určit, na rozdíl od percepční části, kde jsme neměli potřebná data k dispozici. Tato penalizační matice je také na rozdíl od svého percepčního protějšku symetrická, platí tedy $a_{ij} = a_{ji}$.

[7] Všechny hodnoty a'_{ii} jsou nulové, neboť reprezentují vzdálenost mezi stejným dialogovým aktem.

Tabulka 8.3: Konečná akustická penalizační matice **A**.

	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	OTHER	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	**str. hod.**
APOLOGY	0,00	0,83	0,60	0,58	0,34	0,44	0,59	0,45	0,48	0,81	0,36	0,51	0,51	0,46	0,33	0,56	**0,49**
CONFIRM	0,83	0,00	0,49	0,40	0,66	0,78	0,69	0,61	0,71	1,00	0,73	0,54	0,55	0,61	0,78	0,71	**0,63**
DIRECTIVE	0,60	0,49	0,00	0,37	0,34	0,44	0,43	0,33	0,29	0,85	0,37	0,36	0,20	0,24	0,68	0,58	**0,41**
DISCONFIRM	0,58	0,40	0,37	0,00	0,44	0,48	0,45	0,34	0,46	0,89	0,51	0,32	0,39	0,40	0,60	0,55	**0,45**
ENCOURAGE	0,34	0,66	0,34	0,44	0,00	0,32	0,47	0,27	0,27	0,77	0,19	0,36	0,24	0,17	0,45	0,58	**0,37**
GOODBYE	0,44	0,78	0,44	0,48	0,32	0,00	0,59	0,23	0,22	0,95	0,28	0,32	0,40	0,30	0,60	0,72	**0,44**
GREETING	0,59	0,69	0,43	0,45	0,47	0,59	0,00	0,54	0,46	0,87	0,56	0,59	0,48	0,49	0,73	0,52	**0,53**
HAPPY-EMPATHY	0,45	0,61	0,33	0,34	0,27	0,23	0,54	0,00	0,25	0,84	0,26	0,17	0,30	0,20	0,50	0,61	**0,37**
NOT-SPECIFIED	0,48	0,71	0,29	0,46	0,27	0,22	0,46	0,25	0,00	0,89	0,25	0,35	0,29	0,22	0,64	0,65	**0,40**
OTHER	0,81	1,00	0,85	0,89	0,77	0,95	0,87	0,84	0,89	0,00	0,76	0,94	0,74	0,82	0,69	0,56	**0,77**
REQUEST	0,36	0,73	0,37	0,51	0,19	0,28	0,56	0,26	0,25	0,76	0,00	0,36	0,24	0,18	0,50	0,60	**0,38**
SAD-EMPATHY	0,51	0,54	0,36	0,32	0,36	0,32	0,59	0,17	0,35	0,94	0,36	0,00	0,37	0,28	0,56	0,68	**0,42**
SHOW-INTEREST	0,51	0,55	0,20	0,39	0,24	0,40	0,48	0,30	0,29	0,74	0,24	0,37	0,00	0,16	0,58	0,55	**0,38**
SURPRISE	0,46	0,61	0,24	0,40	0,17	0,30	0,49	0,20	0,22	0,82	0,18	0,28	0,16	0,00	0,55	0,62	**0,36**
THANKS	0,33	0,78	0,68	0,60	0,45	0,60	0,73	0,50	0,64	0,69	0,50	0,56	0,58	0,55	0,00	0,53	**0,55**
WAIT	0,56	0,71	0,58	0,55	0,58	0,72	0,52	0,61	0,65	0,56	0,60	0,68	0,55	0,62	0,53	0,00	**0,56**
NEUTRAL	0,78	0,72	0,41	0,54	0,55	0,61	0,42	0,58	0,46	0,84	0,59	0,67	0,45	0,51	0,86	0,63	**0,60**

Na závěr této části musíme poznamenat, že výběr akustických parametrů byl inspirován také ostatními studiemi zabývajícími se akustickou analýzou expresivní řeči, kde jsou právě tyto parametry považovány za velmi důležité z hlediska přenosu expresivního vyjádření od řečníka k posluchači. Je samozřejmě možné, že i jiné akustické parametry (např. spektrální sklon, formantové frekvence a další) mohou expresivitu v řeči ovlivňovat.

8.1.3 Celková penalizační matice

Celkovou penalizační matici **M**, která bude obsahovat konečné vzdálenosti mezi jednotlivými dialogovými akty tak, jak budou později používány při výběru řečových jednotek z inventáře, sestavíme z dříve připravené percepční a akustické penalizační matice. Pro výpočet koeficientů m_{ij} matice **M** platí rovnice

$$m_{ij} = \frac{w_p \cdot p_{ij} + w_a \cdot a_{ij}}{w_p + w_a}, \tag{8.8}$$

kde p_{ij}, resp. a_{ij} jsou koeficienty percepční, resp. akustické penalizační matice, w_p a w_a jsou konstanty reprezentující váhy příslušných matic. Pro naši úlohu jsme volili $w_p = 3$ a $w_a = 1$, přičemž jsme vyzkoušeli i další kombinace těchto vah[8]. Ze všech těchto kombinací se jako subjektivně nejlepší jevila právě námi zvolená kombinace, přestože rozdíly mezi výslednou syntetickou expresivní řečí produkovanou s různým nastavením vah byly velmi malé.

Jak je vidět z rovnice 8.8, pro určení koeficientů matice **M** (zobrazenou v tabulce 8.4) jsme využili vážený průměr koeficientů matic **P** a **A**, přičemž větší váhu (při naší volbě $w_p > w_a$) má percepční penalizační matice. Věříme totiž, že způsob vnímání expresivity posluchači by měl více ovlivňovat následnou syntézu expresivní řeči než výsledky akustické analýzy, proto je na percepci kladen větší důraz. Je však samozřejmě možné více spoléhat na akustickou analýzu, potom bychom volili $w_a > w_p$.

Pro dialogový akt *NEUTRAL*, jenž není zahrnut v percepční penalizační matici, jsme určili koeficienty matice **M** pouze z akustické penalizační matice **A**, a to tak, že $m_{\text{NEUTRAL},j} = a_{\text{NEUTRAL},j}$. Tímto způsobem sice do celkové penalizační matice zavádíme jistou chybu a nerovnováhu, nicméně z důvodu nedostupnosti potřebných dat tak postupovat musíme. Při expresivní (případně i neutrální) syntéze řeči se pak totiž musíme s tímto dialogovým aktem vypořádat, neboť v inventáři řečových jednotek máme kromě jiného uloženy jednotky právě s příznakem dialogového aktu *NEUTRAL*.

[8] $w_p = w_a = 1$; $w_p = 1, w_a = 3$; $w_p = 1, w_a = 0$ a $w_p = 0, w_a = 1$.

Tabulka 8.4: Celková penalizační matice **M**.

	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	OTHER	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	NEUTRAL
APOLOGY	0,00	0,58	0,39	0,40	0,83	0,25	0,90	0,48	0,36	0,50	0,84	0,17	0,45	0,48	0,83	0,47	0,78
CONFIRM	0,96	0,00	0,71	0,58	0,71	0,95	0,92	0,40	0,48	0,72	0,93	0,41	0,50	0,53	0,74	0,70	0,72
DIRECTIVE	0,90	0,56	0,00	0,58	0,26	0,86	0,46	0,50	0,36	0,65	0,33	0,49	0,38	0,81	0,92	0,59	0,41
DISCONFIRM	0,34	0,33	0,84	0,00	0,43	0,87	0,86	0,28	0,23	0,40	0,44	0,28	0,42	0,42	0,90	0,45	0,54
ENCOURAGE	0,83	0,50	0,48	0,64	0,00	0,83	0,60	0,35	0,31	0,43	0,27	0,39	0,14	0,27	0,75	0,52	0,55
GOODBYE	0,30	0,53	0,34	0,87	0,47	0,00	0,59	0,19	0,17	0,50	0,82	0,25	0,42	0,82	0,25	0,55	0,61
GREETING	0,90	0,63	0,86	0,86	0,58	0,90	0,00	0,52	0,31	0,68	0,89	0,90	0,53	0,87	0,93	0,54	0,42
HAPPY-EMPATHY	0,44	0,25	0,58	0,41	0,24	0,39	0,89	0,00	0,21	0,45	0,54	0,29	0,27	0,28	0,46	0,54	0,58
NOT-SPECIFIED	0,39	0,26	0,25	0,35	0,12	0,19	0,26	0,15	0,00	0,34	0,20	0,22	0,13	0,18	0,32	0,38	0,46
OTHER	0,95	1,00	0,29	0,97	0,94	0,36	0,97	0,96	0,30	0,00	0,94	0,99	0,94	0,95	0,92	0,89	0,84
REQUEST	0,84	0,93	0,30	0,88	0,17	0,82	0,35	0,45	0,31	0,40	0,00	0,51	0,28	0,80	0,87	0,58	0,59
SAD-EMPATHY	0,28	0,28	0,37	0,41	0,26	0,32	0,90	0,33	0,28	0,50	0,49	0,00	0,25	0,33	0,89	0,59	0,67
SHOW-INTEREST	0,88	0,53	0,39	0,62	0,15	0,85	0,87	0,43	0,27	0,58	0,32	0,41	0,00	0,34	0,90	0,57	0,45
SURPRISE	0,86	0,29	0,47	0,40	0,05	0,82	0,87	0,15	0,14	0,37	0,35	0,24	0,06	0,00	0,59	0,51	0,51
THANKS	0,83	0,54	0,92	0,90	0,53	0,46	0,93	0,88	0,91	0,92	0,87	0,89	0,90	0,89	0,00	0,88	0,86
WAIT	0,47	0,58	0,24	0,89	0,32	0,93	0,88	0,55	0,56	0,89	0,29	0,57	0,39	0,90	0,88	0,00	0,63
NEUTRAL	0,78	0,72	0,41	0,54	0,55	0,61	0,42	0,58	0,46	0,84	0,59	0,67	0,45	0,51	0,86	0,63	0,00

Jen pro úplnost dodejme, že celková penalizační matice nemusí být symetrická, neboť tuto vlastnost přebírá z percepční penalizační matice. Je tedy opět nutné rozlišovat prvek m_{ij} od prvku m_{ji} (dále viz část 8.2).

Pro zajímavost ještě uvedeme korelaci mezi percepčními a akustickými výsledky pro jednotlivé dialogové akty. Korelační koeficienty jsou uvedené v tabulce 8.5.

Tabulka 8.5: Korelace mezi percepčními a akustickými výsledky.

dialogový akt	korelace
APOLOGY	0,01
CONFIRM	0,64
DIRECTIVE	0,52
DISCONFIRM	0,21
ENCOURAGE	0,32
GOODBYE	0,11
GREETING	0,49
HAPPY-EMPATHY	0,30
NOT-SPECIFIED	0,40
OTHER	0,32
REQUEST	0,29
SAD-EMPATHY	0,41
SHOW-INTEREST	0,65
SURPRISE	0,35
THANKS	0,51
WAIT	0,37
střední hodnota	**0,37**

Střední hodnota korelačního koeficientu $\rho = 0{,}37$ sice není příliš přesvědčivá, nicméně obzvláště pro některé dialogové akty je korelační koeficient poměrně vysoký. Z uvedených korelací by se dalo usuzovat, že pro některé dialogové akty (zejména *CONFIRM* a *SHOW-INTEREST*) existuje silnější vazba mezi jejich akustickými parametry a vnímáním posluchači. Tyto korelace pro nás nicméně nejsou zas až tak příliš důležitým parametrem, neboť celková penalizační matice se skládá z obou částí, jak akustické tak percepční. Pokud bychom zjistili příliš velkou korelaci mezi těmito částmi, mohli bychom využít jen jednu z nich. Avšak dodejme také, že korelace je schopna odhalit pouze závislost lineární, tedy i přes poměrně nízké hodnoty korelačních koeficientů může mezi oběma částmi penalizační matice existovat jiná závislost, kterou bychom mohli odhalit například pokročilejšími statistickými metodami.

8.2 Výpočet ceny cíle

Se znalostí celkové penalizační matice můžeme v samotném algoritmu dynamického výběru jednotek snadno určit potřebnou cenu cíle pro cílovou jednotku a kandidáty z inventáře pomocí již dříve uvedené rovnice 2.1

$$C^{\mathrm{t}}(t_i, u_i) = \sum_{k=1}^{p'} w_k^{\mathrm{t}} C_k^{\mathrm{t}}(t_i, u_i), \tag{8.9}$$

kde t_i je cílová jednotka, u_i je jednotka v inventáři odpovídající cílové jednotce t_i, w_k^{t} je vektor vah, p' je počet všech příznaků (množina příznaků je oproti rovnici 2.1 rozšířena o příznak DA reprezentující dialogový akt) a C_k^{t} je míra vzdálenosti k-tého příznaku, přičemž míra $C_{\mathrm{DA}}^{\mathrm{t}}$ reprezentující vzdálenost mezi dialogovými akty je určena dle celkové penalizační matice.

Protože penalizační matice není symetrická, je nutné rozlišovat jednotlivé prvky. Pokud určujeme cenu cíle mezi cílovou jednotkou t_i s požadovaným dialogovým aktem D_t a kandidátem z inventáře řečových jednotek u_i označeným dialogovým aktem D_u, pak odpovídajícím prvkem v penalizační matici je prvek m_{D_t,D_u}, neboli řádky penalizační matice (objektivní anotace z percepční části) jsou přiřazeny cílové jednotce, zatímco sloupce (subjektivní anotace – percepce expresivity) jsou přiřazeny kandidátu z inventáře.

Výpočet ceny řetězení se nijak nemění, protože ta má za úkol měřit lokální nespojitosti při řetězení jednotek a zajistit hladkost spojení, což by nemělo být expresivitou ovlivněno.

8.3 Nastavení vah pro příznak expresivity

Nastavení váhy pro příznak dialogového aktu není jednoduchým úkolem, neboť samotné nastavování vah pro jakýkoliv příznak je poměrně obtížné. Byly vyvinuty i techniky, jak tyto váhy nastavovat automaticky [67, 34], nicméně v našem systému ARTIC se používá nastavení uvedené v tabulce 8.6, které se dosud osvědčilo ve výzkumných i praktických aplikacích. Přesto je nutno nějakým způsobem určit pokud možno co nejlepší váhu pro příznak dialogového aktu tak, aby vhodně doplňovala váhy ostatních příznaků, a aby negativním způsobem neovlivnila funkci a výstup celého algoritmu.

Hodnotu této váhy jsme tedy určili experimentálně. Vygenerovali jsme dostatečné množství expresivních syntetických vět se zastoupením všech dialogových aktů pomocí různého nastavení váhy pro příznak dialogového aktu. Pro experiment jsme vybrali následující hodnoty: $1, 2, 3, 7, 10, 12, 15, 20, 25, 30, 40, 50$. Subjektivně jsme posoudili, které nastavení již degraduje produkovaný řečový signál (vysoké hodnoty), a které

Tabulka 8.6: Váhy jednotlivých příznaků použitých v systému ARTIC a v systému expresivní syntézy řeči.

příznak	váha
pozice v prozodickém slově	7,0
levý fonémový kontext	3,0
pravý fonémový kontext	3,0
typ prozodému	14,0
shoda znělosti – na začátku	8,5
shoda znělosti – na konci	8,5
dialogový akt	*12,0*

naopak zbytečně potlačuje použití jednotek s patřičným dialogovým aktem na úkor ostatních příznaků (nízké hodnoty). Zároveň jsme vzali v úvahu i váhy pro ostatní příznaky uvedené v tabulce 8.6 tak, aby nastavení bylo pokud možno vyvážené. Po tomto experimentálním posouzení jsme se nakonec rozhodli použít váhu $w_{DA} = 12$, neboť se nám syntetické věty s tímto nastavením subjektivně jevili jako nejpřirozenější.

8.4 Shrnutí navrženého systému

V naší práci jsme vyvinuli doménově omezený systém syntézy expresivní řeči využívající metodu dynamického výběru jednotek se zaměřením na použití v dialogovém systému pro komunikaci mezi člověkem a strojem na dané téma (konverzace o osobních fotografiích). Vývoj tohoto dialogového systému byl součástí mezinárodního projektu Companions (`http://www.companions-project.org`). V následujících bodech shrneme postup celé práce.

- Základním kamenem tohoto vývoje bylo nahrání přirozených reálných dialogů mezi člověkem a počítačem metodou Wizard of Oz. Tato rozsáhlá audiovizuální databáze čítající 65 dialogů a zhruba 60 hodin konverzace byla podrobena analýze a na jejím základě jsme připravili texty a stanovili způsob nahrávání expresivního korpusu určeného pro pozdější syntézu řeči.

- Na základě reálných dialogů jsme nadefinovali různé expresivní kategorie (dialogové akty), které by měly vystihovat a popisovat expresivitu ve vymezené oblasti.

- Vyvinuli jsme speciální nahrávací software, který vyhovoval požadavkům na způsob nahrávání formou připravených scénářů (dialogů). Dále jsme vybrali vhodného řečníka, který bude expresivní korpus namlouvat. Vybrali jsme stejného řečníka (herce), který již v minulosti namlouval neutrální korpus pro syntézu neutrální řeči. Pomocí vytvořené aplikace jsme v nahrávacím studiu nahráli expresivní korpus formou dialogů. Tyto dialogy byly podobné reálným rozhovorům člověka s počítačem.

- Expresivní korpus byl poté ručně anotován několika anotátory prostřednictvím webového poslechového testu pomocí nadefinovaných dialogových aktů. Ze získaných subjektivních anotací jsme metodou maximální věrohodnosti získali anotace objektivní, přičemž míra objektivity a shody s lidskými anotátory byla ověřena několika statistickými metodami. Na expresivním korpusu jsme provedli rozsáhlou akustickou analýzu, při které jsme zkoumali variabilitu různých akustických parametrů a jejich statistických charakteristik napříč všemi dialogovými akty. Na základě výsledků akustické analýzy jsme stanovili numerické rozdíly mezi jednotlivými dialogovými akty a vytvořili jsme akustickou penalizační matici, reprezentující jejich akustické vzdálenosti. Tu jsme vhodně zkombinovali s percepční penalizační maticí získanou z porovnání mezi subjektivními anotacemi a objektivní anotací expresivního korpusu a získali jsme tak celkovou penalizační matici.

- Pomocí neformálních poslechových testů[9] jsme určili váhu pro příznak dialogového aktu, který se, jako jeden z mnoha, využívá při syntéze řeči metodou dynamického výběru jednotek. S použitím této váhy a celkové penalizační matice jsme pak modifikovali definici funkce pro výpočet ceny cíle, což je jedna ze dvou hodnotících funkcí ovlivňujících výběr vhodných jednotek z inventáře.

- S použitím nově vyvinutého systému syntézy expresivní řeči jsme pak vygenerovali několik testovacích vět, které mají sloužit pro ověření úspěšnosti tohoto systému, a to jak z hlediska kvality syntézy, tak z hlediska percepce produkované expresivity v syntetické řeči. Výsledky tohoto ověření jsou popsány v kapitole 9.

Popsaný postup byl sice připraven především pro potřeby dialogového systému popsaného v kapitole 6, nicméně poskytuje metodiku, jak postupovat v případě návrhu jiného dialogového systému v jakékoliv jiné omezené

[9]Neformálním poslechovým testem myslíme test, kterého se zúčastnil malý počet posluchačů.

oblasti s využitím libovolných jiných expresivních kategorií či dokonce i jinak zvoleného popisu expresivity (zde už by však byla nejspíš nutná revize tohoto postupu v závislosti na popisu expresivity).

Důležitým předpokladem pro syntézu expresivní řeči tak, jak zde byla navržena, je správná funkce dialogového systému jako celku. V této práci totiž předpokládáme, že expresivní kategorie, která má být použita pro syntézu, je apriori známa. Je tedy nutné, aby takový dialogový systém tuto funkcionalitu podporoval a takové informace na základě svého stavu (a fázi dialogu) poskytoval. Lze uvažovat i o začlenění takového modulu před vlastní systém syntézy řeči, který by byl schopný sám určit expresivní kategorii jen ze vstupního textu, třeba i bez nějaké další vnější informace, viz například [103].

Kapitola 9

Vyhodnocení navrženého systému

Pro systém navržený v této práci musíme provést ověření, zda splnil požadavky na něj kladené. Zejména tedy budeme zjišťovat, zda posluchači vnímají řeč produkovanou systémem syntézy expresivní řeči skutečně jako expresivní („úspěšnost rozpoznání expresivity“ uvedená v části 9.2.1) a také zda se změnila kvalita syntetické řeči (v porovnání s původním systémem syntézy neutrální řeči, viz část 9.2.2). Protože námi navržená syntéza expresivní řeči je zaměřena na použití v dialogovém systému popsaném v části 6.1, budeme také v části 9.4 zkoumat, jak je syntetická expresivní řeč vhodná pro takový dialog.

Během návrhu našeho systému jsme zjistili, že některé dialogové akty se v expresivním korpusu vyskytují poměrně výrazně častěji než jiné, některé se naopak vyskytují minimálně. Zastoupení jednotlivých dialogových aktů je uvedeno v tabulce 7.5. Pokud uvážíme tento fakt, a dále také rozsáhlost hodnocení pomocí poslechových testů a potřebu reprezentativního počtu hodnotících posluchačů, rozhodli jsme se provést vyhodnocení pouze na nejčastěji se vyskytujících a některých vybraných dialogových aktech. Těmito hodnocenými dialogovými akty jsou:

- *SHOW-INTEREST* – relativní četnost výskytu 34,9 %;
- *ENCOURAGE* – relativní četnost výskytu 29,4 %;
- *CONFIRM* – relativní četnost výskytu 13,2 %;
- *HAPPY-EMPATHY* – relativní četnost výskytu 8,6 %;
- *SAD-EMPATHY* – přidán z důvodu předpokládaného opačného expresivního významu oproti dialogovému aktu *HAPPY-EMPATHY*; relativní četnost výskytu 3,4%;
- *NOT-SPECIFIED* – kromě toho, že patří mezi pět nejčastějších dialogových aktů, by měl částečně reprezentovat neutrální řeč, přestože

z penalizační matice uvedené v části 8.1.3 to úplně zřejmé není; relativní četnost výskytu 7,4 %;

- *NEUTRAL* – toto není skutečný dialogový akt, pouze by měl zastupovat standardní syntézu „neutrálního“ stylu[1], kterou produkuje náš stávající systém ARTIC.

Předpokládáme, že výběr dialogových aktů[2] navržených k hodnocení by za optimálních podmínek neměl nijak ovlivňovat výsledky vyhodnocení. Optimálními podmínkami v tomto případě rozumíme dostatečný počet dat pro všechny dialogové akty, kterých by se v budoucí práci dalo dosáhnout například lepším výběrem vět při nahrávání expresivního korpusu, případně redukcí počtu dialogových aktů sloučením na základě jejich podobnosti.

Všechny dále uvedené poslechové testy byly provedeny pomocí stejného systému jako tomu bylo v případě anotací expresivního korpusu. Popis tohoto systému je tedy uveden v části 7.2, grafické rozhraní je pak podobné tomu na obrázku 7.2. Pouze schéma otázek a voleb je pro vyhodnocovací poslechové testy samozřejmě jiné.

Většina posluchačů, která se zúčastnila poslechových testů, byla z řad odborníků v oboru zpracování řeči, část posluchačů byla tvořena studenty.

9.1 Vnímání expresivity v přirozené řeči

Předtím než začneme hodnotit vnímání expresivity v syntetické řeči, je potřeba stanovit nějaký základ, se kterým můžeme dosažené výsledky porovnávat. Tento základ nám přinese poslechový test založený na hodnocení expresivity v přirozené řeči (náhodně vybraných vět z expresivního i neutrálního korpusu).

Může se zdát, že je tento test zbytečný, protože dialogové akty máme v expresivním korpusu označené také na základě poslechového testu. Nicméně test vnímání expresivity v přirozené řeči byl definován poněkud odlišně než anotace expresivního korpusu. Neměl za úkol kategorizovat promluvy, ale pouze ohodnotit, zda posluchači expresivitu vnímají či nikoliv. Skupina posluchačů byla také odlišná od skupiny anotátorů, která expresivní korpus původně anotovala.

V nově provedeném poslechovém testu byly všem posluchačům předkládány přirozené věty z korpusů a jejich úkolem bylo označit, zda se jim promluva subjektivně jeví jako expresivní, resp. jestli cítí jakékoliv expresivní

[1] Definice neutrálního stylu je uvedená na straně 42 v kapitole 5.

[2] Prvních pět hodnocených dialogových aktů budeme dále označovat také jako „expresivní dialogové akty“, protože jejich smyslem je vyjadřovat expresivitu, zatímco zbylé dva řadíme do kategorie „neutrálních“.

zabarvení řeči. Posluchači měli také možnost vyznačit, že nejsou schopni rozhodnout. Testu se zúčastnilo 14 posluchačů a souhrnné výsledky jsou zobrazeny v tabulce 9.1, výsledky pro jednotlivé zkoumané dialogové akty pak v tabulce 9.2. Bylo testováno celkem 34 promluv, 4 pro každý zkoumaný expresivní dialogový akt a 7 pro neutrální dialogové akty. Promluvy byly z korpusů (expresivního i neutrálního) vybrány náhodně.

Tabulka 9.1: Vyhodnocení vnímání expresivity v přirozené řeči posluchači na základě poslechového testu – souhrnné výsledky.

dialogový akt	úspěšnost vnímání expresivity	nelze rozhodnout
expresivní	52 %	9 %
neutrální	39 %	7 %

Tabulka 9.2: Vyhodnocení vnímání expresivity v přirozené řeči posluchači na základě poslechového testu – výsledky jednotlivě dle dialogových aktů.

dialogový akt	úspěšnost vnímání expresivity	nelze rozhodnout
CONFIRM	38 %	3 %
ENCOURAGE	61 %	7 %
HAPPY-EMPATHY	77 %	4 %
SAD-EMPATHY	73 %	6 %
SHOW-INTEREST	18 %	11 %
střední hodnota	**53** %	**6** %
NOT-SPECIFIED	42 %	13 %
NEUTRAL	36 %	3 %

Výsledky jsou poměrně překvapivé. Ve 36 % případů pro dialogový akt *NEUTRAL* (42 % pro *NOT-SPECIFIED*) posluchači vnímali v řeči expresivitu, přestože se jedná o neutrální věty[3]. To by znamenalo, že i v neutrálním korpusu je obsažen nějaký druh expresivity a posluchači jsou na její vnímání citliví a i v zamýšlené neutrální řeči ji rozpoznají. Tento fakt také může souviset s obsahovou stránkou promluv, neboť jak bylo zjištěno [43], samotný obsah promluv může posluchače velmi ovlivnit při rozhodování, zda něja-

[3]Neutrálními větami rozumíme věty, které by neměli přenášet žádnou expresivitu. Buď byly s tímto záměrem nahrány (dialogový akt *NEUTRAL*) nebo tak byly anotovány (*NOT-SPECIFIED*).

kou expresivitu vnímá či nikoliv. To jsme se však snažili eliminovat právě výběrem náhodných vět.

V případě expresivních dialogových aktů pak velmi záleží na konkrétním dialogovém aktu, neboť například *HAPPY-EMPATHY* a *SAD-EMPATHY* vykazují poměrně vysokou míru vnímání expresivity, zatímco *SHOW-INTEREST* velmi nízkou (na druhou stranu je zde nejvyšší procento odpovědí, že posluchači nejsou schopni rozhodnout). V tomto případě je již otázka obsahu promluv bezpředmětná, neboť pravděpodobně většina promluv v expresivním korpusu označených dialogovými akty jsou obsahově expresivní.

Tyto výsledky nám poskytují základ pro vyhodnocení úspěšnosti syntetické expresivní řeči, neboť pokud u některých dialogových aktů posluchači nevnímají expresivitu ani v přirozené řeči, neměli bychom pravděpodobně očekávat žádné vysoké hodnoty ani u syntetické řeči.

9.2 Vyhodnocení syntetické expresivní řeči

Vyhodnocení expresivní syntetické řeči bylo provedeno obdobným způsobem jako u přirozené řeči. U syntetické řeči jsme však zkoumali jak schopnost posluchačů vnímat expresivitu, tak také kvalitu syntetické řeči, která bude pravděpodobně zanesením prvku expresivity ovlivněna.

Testu se tentokrát zúčastnilo 13 posluchačů. Bylo testováno celkem 30 promluv, 4 pro každý zkoumaný dialogový akt (syntetická řeč) a 2 pro přirozenou neutrální řeč (aby bylo možné porovnat kvalitu přirozené řeči s kvalitou syntetické řeči). Syntetické expresivní promluvy jsou dostupné v příloze G.1.

Musíme ještě podotknout, že texty vět určených pro syntézu nebyly obsaženy v žádném z korpusů přesně tak, jak byly syntetizovány. Chtěli jsme se totiž vyhnout zkreslení výsledků v důsledku pouhého přehrávání vět z korpusu, které by zejména z hlediska kvality řeči byly zřejmě hodnoceny velmi pozitivně. Pro poslechový test byly texty skutečných vět z korpusu upraveny tak, aby byly pouze podobné a měly stejný nebo alespoň přibližně stejný význam.

9.2.1 Vnímání expresivity

Souhrnné výsledky pro test úspěšnosti vnímání expresivity posluchači jsou zobrazeny v tabulce 9.3, výsledky pro jednotlivé zkoumané dialogové akty pak v tabulce 9.4.

Tabulka 9.3: Vyhodnocení vnímání expresivity v syntetické řeči posluchači na základě poslechového testu – souhrnné výsledky.

dialogový akt	úspěšnost vnímání expresivity	nelze rozhodnout
expresivní	54 %	6 %
neutrální	13 %	0 %
přirozená řeč (neutrální)	42 %	4 %

Tabulka 9.4: Vyhodnocení vnímání expresivity v syntetické řeči posluchači na základě poslechového testu – výsledky jednotlivě dle dialogových aktů.

dialogový akt	úspěšnost vnímání expresivity	nelze rozhodnout
CONFIRM	69 %	4 %
ENCOURAGE	42 %	8 %
HAPPY-EMPATHY	50 %	10 %
SAD-EMPATHY	63 %	4 %
SHOW-INTEREST	46 %	4 %
střední hodnota	**54** %	**6** %
NOT-SPECIFIED	10 %	0 %
NEUTRAL	15 %	0 %
přirozená řeč (neutrální)	42 %	4 %

Opět můžeme pozorovat překvapivý výsledek pro přirozenou neutrální řeč, nicméně je to výsledek v souladu s předchozím testem. Pro syntetickou neutrální řeč pak již posluchači většinou žádnou expresivitu nevnímali. V expresivní řeči opět záleží na jednotlivých dialogových aktech. Výsledky však již nejsou v rámci různých dialogových aktů tak rozporuplné (rozkolísané) jako v případě přirozené řeči a průměrná dosažená „úspěšnost rozpoznání expresivity“ dosahující 54 % dokonce lehce převyšuje úspěšnost dosaženou pro přirozenou řeč (52 %). Výsledek pro dialogový akt *HAPPY-EMPATHY* je v případě syntetické řeči výrazně horší než u přirozené řeči (50 % oproti 77 % v přirozené řeči), avšak zase jsme zde zaznamenali největší nárůst posluchačů, kteří nejsou schopni rozhodnout.

Abychom ověřili, že dosažené výsledky nejsou pouze náhodné, provedli jsme výpočet statistické míry reprezentující shodu mezi posluchači, a také míry *precision* (přesnost), *recall* (úplnost), jejich kombinaci *F1 míru* a *accu-*

racy (přesnost) (jsou definovány vztahy 9.1 – 9.4). Ty se využívají především pro hodnocení úspěšnosti klasifikátorů při klasifikační úloze. Pokud se však zamyslíme nad naším poslechovým testem, můžeme jej také považovat za klasifikační úlohu, kdy klasifikátory jsou posluchači a klasifikují v zásadě do dvou tříd, vnímám/nevnímám expresivitu (odpovědi „nelze rozhodnout“ jsme pro toto ověření neuvažovali).

Statistickou mírou reprezentující shodu posluchačů je například Fleissova kappa κ_F, jejíž kladná hodnota značí shodu nad hranicí náhody. Jakou míru shody hodnota κ_F představuje je uvedeno v tabulce 7.1. V našem experimentu jsme naměřili $\kappa_F = 0{,}37$, což značí mírnou shodu.

Hodnoty pro *precision* (P), *recall* (R), *F1 míru* a *accuracy* (Acc) jsou definovány následujícími vztahy:

$$\mathrm{P} = \frac{tp}{tp + fp}, \tag{9.1}$$

$$\mathrm{R} = \frac{tp}{tp + fn}, \tag{9.2}$$

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}, \tag{9.3}$$

$$\mathrm{Acc} = \frac{tp + tn}{tp + tn + fp + fn}, \tag{9.4}$$

kde tp značí počet vět označených jako expresivní, když expresivitu opravdu vyjadřovat měly, tn reprezentuje počet vět, které nebyly označené jako expresivní a expresivitu opravdu neobsahovaly, fp představuje počet vět označených jako expresivní, přestože expresivitu vyjadřovat neměly a fn znamená počet vět, které nebyly označené jako expresivní, přestože byly syntetizovány jako expresivní.

Hodnoty vypočtené pro výsledky poslechového testu jsou uvedeny v tabulce 9.5. Pro porovnání uvádíme tyto hodnoty i pro případ, kdy by posluchači hodnotili v tomto testu expresivitu zcela náhodně[4].

Z ověření výsledků plyne, že dosažená „úspěšnost rozpoznání expresivity“ v syntetické řeči není náhodná. Je třeba si také uvědomit, že existují dvě základní skutečnosti, které výsledky vnímání expresivity ovlivňují. První z nich

[4]Výpočet těchto hodnot jsme provedli pomocí simulace, kdy jsme simulovali posluchače, který by poslechový test vyplňoval náhodně. Promluvy byly automaticky a zcela náhodně označovány jednou z kategorií vnímám/nevnímám expresivitu, jak by to dělal takový posluchač. Nasimulovali jsme 13 posluchačů (stejný počet, jaký skutečně testy prováděl) a tento postup jsme 10 000× zopakovali.

Tabulka 9.5: Úspěšnost vnímání (klasifikace) expresivity v syntetické řeči posluchači na základě poslechového testu a porovnání se situací, kdy by posluchači hodnotili zcela náhodně.

míra	posluchači	náhoda
precision	0,92	0,72
recall	0,58	0,50
F1 míra	0,71	0,59
accuracy	0,66	0,50

je syntetizér, potažmo syntetická expresivní řeč, jejíž hodnocení je naším záměrem. Pak je zde ovšem ještě faktor posluchačů, kde každý z nich může různou intenzitu či různý druh expresivity hodnotit (vnímat) různě. Naším úkolem však není hodnotit samotné posluchače, zda jsou či nejsou schopni vnímat expresivitu (což v zásadě ani nelze). Posluchačům při vyhodnocování výsledků prostě „věříme“ a pouze ověřujeme jejich míru shody. Nicméně tento faktor může naše výsledky lehce ovlivnit, a to především v takovéto úloze, kde již samotný pojem expresivity je poměrně vágním, tudíž obtížně hodnotitelným.

Dosažené výsledky vnímání expresivity pak můžeme také ještě porovnat i s předchozími průběžnými poslechovými testy, provedenými se syntetickou expresivní řečí (uvedenými v [43]), kde jsme ještě neměli přesně definovanou penalizační matici. Systém tak pracoval se základním jednoduchým nastavením[5]. Dodejme, že tento systém využíval řečový korpus složený z neutrálního korpusu a anotovaného expresivního korpusu. V [43] jsme dosáhli úspěšnosti rozpoznání expresivity u expresivně zabarvených textů[6] 45 % (u neutrálních textů pak 10 %). Při porovnání s aktuálními výsledky (rozpoznání expresivity 54 %) můžeme tedy konstatovat, že použití námi navržené penalizační matice významně zvýšilo vnímání expresivity posluchači.

Při hodnocení expresivity nám také může pomoci ještě jeden doplňující údaj. Tímto údajem je relativní četnost řečových jednotek v syntetické promluvě, jejichž dialogový akt odpovídal požadovanému pro danou promluvu, při stávajícím nastavením vah pro příznaky[7]. Naměřené hodnoty jsou uvedené

[5]Nastavení penalizace bylo binární, tj. pokud řečová jednotka byla označena požadovaným dialogovým aktem, její penalizace byla 0, pokud ne, penalizace byla 1.

[6]Expresivně zabarvenými texty rozumíme texty, jejichž obsah je expresivní.

[7]Poměr jednotek se správným dialogovým aktem lze samozřejmě zvýšit i zvýšením váhy pro příznak dialogového aktu.

v tabulce 9.6[8]. Kompletní přehled relativního zastoupení jednotek s různými dialogovými akty při syntéze expresivní řeči s konkrétním dialogovým aktem je pro syntetické věty z tohoto poslechového testu uveden v příloze F v tabulce F.19.

Tabulka 9.6: Relativní počet vybraných jednotek s požadovaným příznakem dialogového aktu ve výsledných syntetických větách.

dialogový akt	poměr jednotek
CONFIRM	74,5 %
ENCOURAGE	67,7 %
HAPPY-EMPATHY	35,2 %
SAD-EMPATHY	32,9 %
SHOW-INTEREST	39,7 %
střední hodnota	**50,0** %
NOT-SPECIFIED	3,5 %
NEUTRAL	100,0 %

Můžeme pozorovat, že pro *NOT-SPECIFIED* je poměr jednotek se správným dialogovým aktem velmi nízký (3,5 %). Při bližším zkoumání jsme zjistili, že při požadavku na jednotky s tímto dialogovým aktem, byly v drtivé většině případů vybrány jednotky *NEUTRAL* (viz tabulka F.19). To přispívá k názoru, že tento dialogový akt víceméně reprezentuje neutrální řeč, přestože v celkové penalizační matici **M** byla vzdálenost mezi *NOT-SPECIFIED* a *NEUTRAL* vypočtena jako 0,46.

Při porovnání s tabulkou 9.4, kde jsou uvedené výsledky vnímání expresivity posluchači, pak můžeme konstatovat, že výběr jednotek se správným dialogovým aktem příliš nekoresponduje s vnímáním expresivity. Například během syntézy expresivní promluvy s požadovaným dialogovým aktem *SAD-EMPATHY* bylo vybráno jen 33 % jednotek označených tímto dialogovým aktem, přesto posluchači s 63% úspěšností rozpoznali v takové syntetické řeči expresivitu. Naopak pro dialogový akt *ENCOURAGE* byl poměr těchto jednotek 68 %, expresivita však byla posluchači rozpoznána jen v 42 % případů.

9.2.2 Vyhodnocení kvality

Abychom zjistili, jestli se kvalita syntetické řeči přidáním expresivního prvku významně nezhoršila, provedli jsme vyhodnocení kvality pomocí tzv. MOS-

[8]Uvedený poměr jednotek znamená, že např. při syntéze expresivní řeči s dialogovým aktem *CONFIRM* mělo 74,5 % řečových jednotek (vybraných metodou dynamického výběru jednotek) skutečně příznak dialogového aktu *CONFIRM*.

testu (z anglického *Mean Opinion Score*). Při MOS-testu posluchači hodnotí kvalitu řeči na základě pětibodové stupnice uvedené v tabulce 9.7, kde přirozená řeč by teoreticky měla získat hodnocení 5 a velmi nekvalitní řeč 1. Test probíhal souběžně s testem vnímání expresivity, podmínky testu a posluchači byli tedy totožní.

Výsledky vyhodnocení kvality jsou souhrnně uvedené v tabulce 9.8, výsledky pro jednotlivé zkoumané dialogové akty pak v tabulce 9.9. V tabulkách uvádíme i porovnání s přirozenou řečí v procentech, které v podstatě znamená relativní hodnocení kvality syntetické řeči vztažené k hodnocení kvality přirozené řeči. Toto relativní porovnání nám umožňuje srovnávat výsledky z různých poslechových testů zaměřených na hodnocení kvality syntetické řeči pomocí MOS testů.

Tabulka 9.7: Hodnotící stupnice MOS-testu.

slovní hodnocení kvality	číselné vyjádření
vynikající	5
dobrá	4
uspokojivá	3
špatná	2
velmi špatná	1

Tabulka 9.8: Vyhodnocení kvality syntetické expresivní řeči posluchači na základě poslechového testu – souhrnné výsledky.

dialogový akt	MOS hodnocení	porovnání s přirozenou řečí
expresivní	3,5	76 %
neutrální	4,0	87 %
přirozená řeč	4,6	100 %

Při pohledu na výsledky zjistíme, že kvalita syntetické expresivní řeči je o něco horší než kvalita syntetické neutrální řeči, a to přesně o 0,5 bodu na MOS stupnici (o 11 procentních bodů). Je to přibližně stejný rozdíl jako mezi syntetickou neutrální řečí a řečí přirozenou. Toto zhoršení přikládáme větší variabilitě akustického signálu expresivní řeči, kdy při řetězení jednotek

Tabulka 9.9: Vyhodnocení kvality syntetické expresivní řeči posluchači na základě poslechového testu – výsledky jednotlivě dle dialogových aktů.

dialogový akt	MOS hodnocení	porovnání s přirozenou řečí
CONFIRM	3,9	85 %
ENCOURAGE	3,5	76 %
HAPPY-EMPATHY	3,1	67 %
SAD-EMPATHY	3,9	85 %
SHOW-INTEREST	3,3	72 %
střední hodnota	**3,5**	**76** %
NOT-SPECIFIED	3,9	83 %
NEUTRAL	4,0	87 %
přirozená řeč	4,6	100 %

může docházet k nežádoucím jevům a defektům častěji, než při syntéze klidné neutrální řeči[9]. Nicméně i tak je výsledek poměrně příznivý.

Pokud budeme opět porovnávat s předchozím systémem bez penalizační matice [43], zjistíme, že její použití se příliš nepromítlo do kvality syntetické řeči, neboť MOS hodnocení zůstalo na stejné hodnotě 3,5 (ačkoliv v poměrném porovnání s hodnocením přirozené řeči se kvalita o 3 procentní body zhoršila, z předchozích 79 % na současných 76 %, nicméně tuto změnu vnímáme spíše jako nevýznamnou).

I zde můžeme použít ještě jeden pomocný parametr pro zkoumání kvality (plynulosti) syntetické řeči. Je jím relativní četnost „hladkých“ spojů řečových jednotek. Jako hladký spoj zde označujeme spojení dvou řečových jednotek, které se v původním řečovém korpusu vyskytovaly vedle sebe, tj. jejich spojení je přirozené. Naměřené hodnoty jsou zobrazeny v tabulce 9.10.

Lze vidět, že relativní četnost hladkých spojů je přibližně stejná jak napříč všemi dialogovými akty, tak i v porovnání s neutrální syntézou. To znamená, že dochází ke stejně častému řetězení jak při syntéze expresivní řeči, tak řeči neutrální. Zde můžeme použít srovnání s dalšími průběžnými výsledky v [42], kde byla již sice použita penalizační matice ve stejném formátu jako nyní (na rozdíl od [43]), avšak její koeficienty byly rozdílné, neboť postup jejich výpočtu byl poněkud odlišný. V [42] jsme tedy dosáhli průměrné relativní četnosti hladkých spojů 72 % (zatímco nyní 79 %). V tomto ohledu jsme tedy

[9]Připomeňme, že v použitém systému ARTIC nedochází k žádným modifikacím a signálovým úpravám při řetězení jednotek, kromě jednoduchého vyhlazování spojů.

Tabulka 9.10: Relativní četnost hladkých spojů, tj. přirozených spojení dvou jednotek, které se v původním korpusu vyskytovaly za sebou.

dialogový akt	hladké spoje
CONFIRM	80 %
ENCOURAGE	76 %
HAPPY-EMPATHY	77 %
SAD-EMPATHY	80 %
SHOW-INTEREST	82 %
střední hodnota	**79** %
NOT-SPECIFIED	82 %
NEUTRAL	82 %

dosáhli zlepšení a porovnáním výsledků MOS testu dojdeme k podobnému zjištění. V [42] jsme dosáhli ohodnocení kvality 3,4 na MOS stupnici (72 % v porovnání s přirozenou řečí), nyní 3,5 (76 %). Lze tedy říci, že postupným vylepšováním penalizační matice dochází také ke zlepšování kvality syntetické řeči.

9.3 Výsledky expresivní HMM syntézy

Přestože jsme se v této práci zaměřili především na syntézu řeči metodou dynamického výběru jednotek, provedli jsme i experiment s expresivní syntézou založenou na HMM[10]. Ten je podrobněji popsán v [43]. Využili jsme metodu tzv. *stylově nezávislého modelování* popsanou na straně 34, kdy se příznak dialogového aktu stává jedním z kontextů kontextově závislých HMM modelů (mezi další kontexty patří například levý a pravý fonémový kontext, pozice v prozodickém slově nebo typ prozodému). Kvůli zvýšení robustnosti systému jsou natrénované HMM modely pomocí rozhodovacích stromů shluknuty tak, že podobné řečové jednotky sdílí stejný model[11]. Takto natrénované HMM modely jsou pak využity pro generování syntetické expresivní řeči.

Cílem vyhodnocení pak bylo zkoumat schopnost metody HMM vyjádřit v syntetické řeči expresivitu a zjistit kvalitu takto produkované řeči. Dosažené výsledky jsme také porovnali s metodou výběru jednotek. Vyhodnocení

[10] Zde použitá syntéza řeči metodou HMM je založená na systému HTS [115], adaptovaná na český jazyk [48].

[11] Algoritmus shlukování pracuje na základě podobnosti řečových jednotek v různých kontextech.

proběhlo pomocí poslechových testů, kterého se zúčastnilo 12 respondentů. Souhrnné výsledky uvádíme v tabulkách 9.11 a 9.12. Ukázky syntetické expresivní promluvy jsou dostupné v příloze G.2.

Tabulka 9.11: Vyhodnocení vnímání expresivity v syntetické řeči generované metodou HMM na základě poslechového testu – souhrnné výsledky.

dialogový akt	úspěšnost vnímání expresivity	nelze rozhodnout
expresivní	15 %	5 %
NOT-SPECIFIED	8 %	3 %

Tabulka 9.12: Vyhodnocení kvality syntetické expresivní řeči generované metodou HMM na základě poslechového testu – souhrnné výsledky.

dialogový akt	MOS hodnocení	porovnání s přirozenou řečí
expresivní + *NOT-SPECIFIED*	2.7	61 %
přirozená řeč	4.4	100 %

Lze pozorovat, že vnímání expresivity v řeči generované metodou HMM je na velmi nízké úrovni (15 %) v porovnání s metodou dynamického výběru jednotek (54 %). Podobně zaostává metoda HMM i v porovnání kvality řeči (61 % hodnoty přirozené řeči oproti 76 % u metody dynamického výběru jednotek). Přikládáme to tomu, že obecně HMM syntéza pro češtinu ještě není na takové úrovni, aby produkovala syntetickou řeč srovnatelnou s metodou výběru jednotek a v ne příliš kvalitní syntetické řeči pak posluchači těžko identifikují nějakou expresivitu.

9.4 Vnímání expresivity v dialogu

Syntéza expresivní řeči bude využita v konkrétním dialogovém systému. Cílem tedy je ověřit, že navržený systém syntézy expresivní řeči v dané úloze – dialogovém systému „rozprávění seniorů o fotografiích“ (viz část 6.1) – předčí původní systém syntézy neutrální řeči. K tomuto ověření jsme opět využili poslechového testu (tentokrát preferenčního) založeného na stejné webové aplikaci jako všechny předchozí. Připravili jsme vhodné testovací stimuly následujícím způsobem:

- Vybrali jsme 6 přibližně minutu dlouhých vhodných[12] částí z náhodně vybraných přirozených reálných dialogů člověka s počítačem (tyto části budeme dále označovat jako *minidialogy*).

- Akustický signál každého minidialogu jsme rozdělili na části, ve kterých mluví člověk, a části, ve kterých jsou reakce počítače vyjádřené prostřednictvím syntézy neutrální řeči (původní reakce avatara získané pomocí metody WoZ, viz část 6.2).

- Vhodně jsme upravili textové obsahy reakcí avatara tak, aby se při následné syntéze pouze nepřehrávaly z expresivního nebo neutrálního korpusu. Přitom jsme samozřejmě zachovali význam těchto reakcí, abychom nenarušili tok dialogu.

- Připravené texty jsme vysyntetizovali jak původním systémem syntézy neutrální řeči, tak i nově navrženým systémem syntézy expresivní řeči – hlas byl samozřejmě v obou případech stejný a před syntézou jsme určili správný (předpokládaný) dialogový akt[13].

- V částech minidialogů představujících řeč člověka jsme v některých případech provedli drobné změny, abychom zkrátili jejich celkovou délku – odstranili jsme pasáže, kde člověk mluví delší dobu nebo kde je pauza (ticho) – samozřejmě tak, abychom zachovali význam dialogu, jeho spojitost a souvislosti, a aby nebylo poznat, že k nějakým úpravám vůbec došlo.

- Části minidialogů jsme opět spojili a to tak, aby pro každý minidialog vznikly dvě verze: reakce avatara se syntetickou neutrální řečí a s expresivní řečí.

Každý ze 6 připravených minidialogů (ve dvou variantách) obsahoval průměrně 4 interakce počítače[14] vyjadřujících různé dialogové akty, nejčastěji *SHOW-INTEREST* či *ENCOURAGE*. Nicméně všechny hodnocené expresivní dialogové akty byly zastoupeny alespoň jednou. Stimuly (minidialogy) v obou variantách pak byly prostřednictvím poslechového testu postupně[15]

[12]Vhodnost jsme posuzovali z hlediska dostatečné interakce počítače (3D avatara) v patřičné části dialogu. Jistě bychom špatně hodnotili syntézu řeči v těch částech dialogu, kde avatar vůbec nereaguje, či pouze souhlasně přitakává. Hlavní úlohou avatara bylo povzbuzovat seniory k vyprávění (např. kladením různých otázek) a posouvat dialog dál.

[13]Předpokládáme, že v konečném dialogovém systému bude navržen mechanismus, který bude text určený k syntéze automaticky označovat požadovaným dialogovým aktem.

[14]Celkem bylo počítačových interakcí 23 pro každou variantu.

[15]Vždy obě varianty v jedné testové otázce.

předloženy posluchačům, jejichž úkolem bylo rozhodnout, která varianta je pro ně příjemnější, přirozenější a kterou by oni preferovali, pokud by byli na místě člověka komunikujícího s počítačem. Posluchači měli možnost vyznačit, že nejsou schopni rozhodnout. Výsledky tohoto vyhodnocení jsou uvedené v tabulce 9.13, poslechového testu ze zúčastnilo 11 posluchačů. Minidialogy jsou dostupné v příloze G.3.

Tabulka 9.13: Hodnocení syntetické expresivní řeči v dialogu v porovnání se systémem syntézy řeči neutrální pomocí preferenčního poslechového testu.

typ syntézy	preference
neutrální	8 %
expresivní	83 %
nelze rozhodnout	9 %

Z uvedených výsledků jednoznačně vyplývá, že posluchači preferovali syntézu řeči expresivní před neutrální, a to v 83 % případů. Toto je jedno z nejdůležitějších hodnocení, které znamená, že navržený systém je posluchači v této práci lépe přijímán a preferován.

Při pohledu na seznam jednotek, ze kterých byly syntetické expresivní promluvy pro tento test sestaveny, jsme si ověřili, že skutečně nedochází k pouhému přehrávání vět z expresivního korpusu, ale k jejich korektní syntéze. K tomuto ověření lze použít relativní četnost tzv. „hladkých“ spojů definovaných na straně 116. Při syntéze expresivní řeči pro účely vyhodnocení vnímání expresivity v dialogu byla naměřena průměrná relativní četnost těchto spojů 86 %. To je o něco více, než bylo naměřeno při vyhodnocení vnímání expresivity a kvality syntetické řeči v části 9.2 (viz tabulka 9.10, pro neutrální řeč 82 % a pro expresivní řeč 79 %). Výjimečný byl jeden jediný případ samostatné věty „*Aha.*“ reprezentující dialogový akt *CONFIRM*, která se ve skutečnosti v expresivním korpusu vyskytuje a byla tak v minidialogu pouze „přehrána“. Nicméně záměna jejího textového obsahu za významově podobný a v korpusu se nevyskytující je téměř nemožná a pro zachování přirozenosti v minidialogu jsme ji nemohli vypustit. Její relativní četnost „hladkých“ spojů tak celkovou průměrnou dosaženou hodnotu navyšuje.

Kapitola 10

Závěr

Přestože syntéza řeči je již na velmi kvalitní úrovni, vyjádření postojů mluvčího prostřednictvím takto syntetizované promluvy je stále velmi komplikované. Jak nastínila námi provedená akustická analýza, jejíž výsledky jsou shrnuty v části 7.3.7, vliv různých expresivních stavů na lidskou řeč je významný, co se měřitelných akustických parametrů týká. Avšak ovlivňování těchto parametrů modifikací řečového signálu během syntézy pak může mít nežádoucí vliv na kvalitu syntetizované řeči.

Uvažovat proto o obecné syntéze expresivní řeči (obzvláště pro češtinu) je za současného stavu a podmínek velmi ambiciózní. Nejprve bychom se totiž měli zaměřit na poznávání různých expresivních stavů člověka, jejich popisu a vlivu na řeč (či na lidské výrazové prostředky obecně) a hlavně na možné využití uvedených znalostí v praktických úlohách. Jednou z možných oblastí praktické aplikace systému syntézy expresivní řeči se jeví obor dialogových systémů, kde dochází ke skutečné interakci člověka s počítačem.

Má-li takový dialog probíhat přirozeně, tj. tak, aby byl i člověk přesvědčen, že hovoří s lidským protějškem, je třeba brát v úvahu dialogový systém jako celek, nejen některou z jeho částí. Pouze přirozená (i když expresivní) řeč totiž nikoho nepřesvědčí o „lidském“ chování stroje. Je to i kvalitní rozpoznání lidské řeči, nálad a pocitů, dále pak korektní porozumění a pochopení a správné reakce ve správnou dobu. V dialogovém systému, u kterého dochází k vizuálnímu kontaktu člověka s počítačem (například prostřednictvím obrazovky), je také důležitá vhodná vizualizace počítačového avatara. Při kvalitní a přirozeně znějící syntetické řeči by také tento avatar měl vypadat přirozeně. Teprve splnění těchto podmínek může stroj přiblížit lidskému chování a přesvědčit tak uživatele o tom, že nekomunikuje se strojem. V této práci se však věnujeme pouze jedné části tohoto komplexního systému, jakým takový dialogový systém je.

Naším úkolem v této práci bylo prozkoumat možnosti syntézy expresivní

řeči, která by měla napomoci přirozenějšímu vnímání řeči uměle vytvářené. Zaměřili jsme se na dialogový systém, jehož vývoj byl součástí mezinárodního projektu Companions (`http://www.companions-project.org`). Tento projekt měl za úkol stvořit umělého virtuálního společníka starším lidem, se kterým by mohli hovořit o osobních fotografiích z jejich života. Přesto se nám v rámci této práce podařilo navrhnout systém tak, aby byl bez nějakých velkých zásahů obecně použitelný v jakékoliv jiné obdobné oblasti dialogových systémů. Protože největší zkušenosti a nejlepší výsledky máme s metodou dynamického výběru jednotek, rozhodli jsme se ji použít i v případě syntézy expresivní řeči.

Shrnutí navrženého systému je již podrobněji popsáno v části 8.4. Zde tedy již jen stručně v bodech zopakujeme postup vývoje syntézy expresivní řeči pro dialogový systém, který by měl sloužit ke komunikaci mezi člověkem a počítačem na dané téma. Dále pak zhodnotíme dosažené výsledky a uvedeme návrhy na možná budoucí vylepšení tohoto systému.

10.1 Stručný souhrn

Seznámení se s plánovaným dialogovým systémem. Byly definovány požadavky kladené na dialogový systém týkající se zejména expresivního vyjadřování prostřednictvím řeči.

Získání reálných přirozených dat. Abychom si ujasnili tematickou oblast, ve které bude dialogový systém působit, rozhodli jsme se začít získáním přirozených dat. Tento sběr dat proběhl metodou *Wizard of Oz*, tj. nahráváním reálných dialogů člověka s počítačem, kdy počítač byl skrytě ovládán lidskými operátory.

Definice expresivity. Jestliže mluvíme o syntéze expresivní řeči, je třeba nadefinovat a popsat, co to vlastně taková expresivita, potažmo expresivní řeč, vlastně je. Samozřejmě nejsme schopni získat nějaký kompletní výčet všech možných stavů, se kterými se může systém v budoucnu potýkat. Na základě studia reálných dialogů a podobných popisných schémat navržených v minulosti jsme však nadefinovali množinu několika expresivních kategorií, které jsme označili jako dialogové akty. Předpokládáme, že tyto dialogové akty budou dostatečně popisovat expresivitu, která se v takových dialozích může vyskytnout.

Nahrávání a zpracování expresivního korpusu pro syntézu. Zvolili jsme metodu předem připraveného scénáře a vybrali jsme vhodného řečníka pro nahrávání expresivního korpusu. Metodou automatické segmentace byl tento korpus segmentován na jednotlivé řečové jed-

notky. Prostřednictvím poslechového testu byl také anotován nezávisle několika posluchači pomocí navržených dialogových aktů.

Akustická analýza expresivního korpusu. Nahraný expresivní korpus jsme poté podrobili rozsáhlé analýze akustického signálu. Ta měla odhalit odlišnosti různých akustických parametrů a jejich statistických charakteristik napříč definovanými dialogovými akty.

Modifikace stávajícího systému. Na základě anotací a akustické analýzy jsme navrhli a provedli modifikace hodnotící funkce (ceny cíle), která ovlivňuje výběr řečových jednotek z korpusu. Cílem bylo, abychom z neutrálního i expresivního korpusu mohli vybírat ty jednotky, které jsou během syntézy expresivní řeči nejvhodnější.

Tímto postupem jsme získali systém syntézy expresivní řeči pro konkrétní dialogový systém. Jeho úspěšnost a praktické výsledky jsme následně ověřili pomocí poslechových testů.

10.2 Zhodnocení výsledků

Podrobné výsledky jsou uvedené v kapitole 9, nicméně si zde dovolíme prezentovat jejich krátký souhrn. Ověření jsme provedli pouze pro vybrané expresivní kategorie, jejichž reprezentanti se v expresivním korpusu vyskytovaly v dostatečném počtu. Chtěli jsme se tak vyhnout zkreslení výsledků v důsledku řídkých dat.

Pomocí poslechových testů jsme ověřovali úspěšnost identifikace expresivity v syntetické řeči. Bylo zjištěno, že posluchači byli schopni detekovat v syntetické řeči expresivitu s přibližně stejnou úspěšností (dokonce o něco vyšší) jako u přirozených nahrávek z expresivního korpusu.

Pro vyhodnocení kvality syntetické řeči jsme opět využili poslechového testu. Zjistili jsme, že použitím jednotek z expresivního korpusu se kvalita syntetické řeči mírně zhoršila v porovnání s původní neutrální syntetickou řečí. To se ovšem dalo předpokládat, neboť expresivní řeč bývá obvykle mnohem dynamičtější a variabilnější než klidná neutrální promluva. Nicméně dosažené výsledky jsou z hlediska kvality stále poměrně příznivé.

Protože syntéza expresivní řeči byla navržena především pro konkrétní aplikaci v dialogovém systému, vyhodnotili jsme její úspěšnost i z pohledu použití v dialogu. To jsme provedli pomocí preferenčního poslechového testu, kdy se posluchači měli rozhodovat mezi původní syntézou neutrální řeči a nově navrženým systémem při poslechu dialogu mezi člověkem a počítačem. Výsledkem byla výrazná preference námi navrženého systému. Tento výsle-

dek považujeme za jeden z nejdůležitějších, neboť právě ten reprezentuje přínos syntézy expresivní řeči v dialogu.

10.3 Návrhy pro další vývoj

Vzhledem ke komplexnosti řešené úlohy jsme v některých fázích byli nuceni udělat určité ústupky. V této části bychom alespoň v několika bodech rádi nastínili, jakým možným dalším směrem by se mohl ubírat vývoj syntézy expresivní řeči metodou dynamického výběru jednotek či jakým způsobem modifikovat stávající postup, aby se kvalita syntetické expresivní řeči ještě zlepšila.

- Rozdělit penalizační matici (především její akustickou část) pro jednotlivé fonémy či skupiny fonémů, neboť každý takový foném či skupina může být v expresivní řeči ovlivněna odlišným způsobem. Vznikla by tak fonémově závislá penalizační matice. Zde se ale může projevit problém řídkosti dat, neboť pro některé fonémy či skupiny fonémů jich nemáme pro určité dialogové akty dostatečné množství.

- Prozkoumat vliv expresivity na různé části určité promluvy, tj. zda se expresivita projevuje třeba jen v přízvučných slabikách, či pouze na začátku/konci promluvy, a případně podle toho opět vytvořit více penalizačních matic.

- Zapracovat výsledky formantové analýzy či analýzy dalších akustických parametrů (spektrální sklon, tranzienty) do penalizační matice. To do jisté míry souvisí s vytvořením fonémově závislé penalizační matice, neboť některé akustické parametry jsou dostupné jen pro určitou skupinu fonémů, nebo je dokonce potřeba jejich hodnoty pro různé fonémy odlišovat (jako například právě hodnoty formantových frekvencí).

- Zvážit použití jiné reprezentace akustických parametrů, například na základě dalších (či jiných) statistických charakteristik. To zahrnuje bližší zkoumání pravděpodobnostního rozdělení hodnot jednotlivých akustických parametrů.

- Upravit množinu dialogových aktů, například jejich sloučením na základě podobnosti[1]. Tím bychom mohli dosáhnout menšího počtu dialogových aktů a částečně tak vyřešit problém řídkosti dat pro fonémově závislou penalizační matici.

[1]Podobnost by mohla být založena třeba na výsledcích akustické analýzy, resp. akustické penalizační matici.

- Revidovat postup návrhu textů určených k nahrávání expresivního korpusu. Jedním ze způsobů je například lepší vyvážení zastoupení (četnosti) jednotlivých dialogových aktů. Korpus je sice anotován pomocí dialogových aktů až po nahrání, ale bylo by možné provést nějaký odhad výskytu dialogových aktů v textech a podle toho pak soubor textů upravit. Další možností je také fonémové vyvážení v rámci jednotlivých dialogových aktů, aby byl každý foném označený určitým dialogovým aktem reprezentován dostatečným množstvím řečových jednotek.
- Zvážit možnost modifikace řečového signálu, který vzniká řetězením řečových jednotek. Ta by mohla odstranit jeho „rozkolísanost“, která se objevuje při použití jednotek z expresivního korpusu. Bylo by však nutné vzít v potaz, že taková modifikace může degradovat řečový signál.

Příloha A

Metoda maximální věrohodnosti [1]

Metoda maximální věrohodnosti je statistickou technikou bodového odhadu parametrů vhodného pravděpodobnostního modelu dobře vysvětlujícího pozorovaná data. Pokud $y_1, y_2, \ldots, y_n$ je náhodný výběr z diskrétního pravděpodobnostního rozdělení popsaného pravděpodobnostní funkcí f_Θ závislou na (vektorovém) parametru Θ, pak maximálně věrohodný odhad je taková hodnota $\hat{\Theta}$, pří níž za daných hodnot $y_1, y_2, \ldots, y_n$ nabývá maxima věrohodností funkce

$$\mathcal{L}(\Theta, y_1, y_2, \ldots, y_n) = f_\Theta(y_1, y_2, \ldots, y_n), \tag{A.1}$$

neboli

$$\hat{\Theta}(y_1, y_2, \ldots, y_n) = \arg\max_\Theta \mathcal{L}(\Theta, y_1, y_2, \ldots, y_n). \tag{A.2}$$

Pro naši aplikaci na určení objektivní anotace[2] z m subjektivních anotací (pro m posluchačů) tedy pro každý dialogový akt D zvlášť uvažujme množinu n vektorů $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n$, kde n je počet zkoumaných vět v korpusu. Vektor $\mathbf{o}_t$ (délky m) obsahuje binární odpovědi[3] posluchačů $1 \ldots m$ pro konkrétní dialogový akt D a pro větu t ($t \in T$, kde $T = \{1, 2, \ldots, n\}$), tedy

$$\mathbf{o}_t = [O_t^{(1)}, O_t^{(2)}, \ldots, O_t^{(m)}]^T, \tag{A.3}$$

kde $O^{(j)}$ je náhodný proces

$$O^{(j)} = \{O_t^{(j)} : t \in T\}, \tag{A.4}$$

[1] Částečně převzato z [97].

[2] Význam výrazu objektivní anotace v naší práci byl uveden na straně 63.

[3] Binární odpovědí rozumíme fakt, že věta představuje/nepředstavuje konkrétní dialogový akt D.

a $O_t^{(j)}$ jsou náhodné proměnné, pro něž platí

$$\forall j : O_t^{(j)} = \begin{cases} 1 & \text{věta } t \text{ podle posluchače } j \text{ představuje dialogový akt } D \\ 0 & \text{věta } t \text{ podle posluchače } j \text{ nepředstavuje dialogový akt } D \end{cases}$$

Nechť X je náhodný proces definovaný

$$X = \{X_t : t \in T\}, \tag{A.5}$$

kde X_t jsou náhodné proměnné, pro něž platí $X_t = 1$ pokud věta t představuje dialogový akt D a $X_t = 0$ v opačném případě.

Dále předpokládejme, že X je stacionární proces s alternativním pravděpodobnostním rozdělením, tedy

$$X \sim A(p_X), \tag{A.6}$$

kde $\forall h, i \in T : p_h = p_i = p_X$ a $P(X_t = 0) = 1 - P(X_t = 1) = p_X$. Tento předpoklad znamená, že u všech vět se stejnou pravděpodobností očekáváme, že daný dialogový akt D představují.

Předpokládejme, že vektory $\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_n$ jsou náhodným výběrem z vyšetřovaného pravděpodobnostního rozdělení. Nechť je toto pravděpodobnostní rozdělení závislé na $2m + 1$ parametrech:

$$\Theta = [p_X, r_X^{(1)}, r_X^{(2)}, \ldots, r_X^{(m)}, f_X^{(1)}, f_X^{(2)}, \ldots, f_X^{(m)}]^T, \tag{A.7}$$

kde p_X je parametr alternativního rozdělení ze vztahu A.6 a $r_X^{(j)}, f_X^{(j)}$ stojí v popisu kauzálních vztahů mezi procesem X a procesy $O^{(j)}$ (tedy mezi skutečným výskytem dialogového aktu D a anotací posluchačů) vyjádřených prostřednictvím pravděpodobností:

$$P(O_t^{(j)} = 1 | X_t = 1) = r_X^{(j)}, \tag{A.8}$$

$$P(O_t^{(j)} = 0 | X_t = 1) = 1 - r_X^{(j)}, \tag{A.9}$$

$$P(O_t^{(j)} = 0 | X_t = 0) = f_X^{(j)}, \tag{A.10}$$

$$P(O_t^{(j)} = 1 | X_t = 0) = 1 - f_X^{(j)}. \tag{A.11}$$

Vztah A.8 vyjadřuje pravděpodobnost, že j-tý posluchač označí větu t jako větu představující dialogový akt D, když tato věta skutečně tento dialogový akt představuje. Oproti tomu vztah A.9 znamená pravděpodobnost,

že j-tý posluchač udělá „chybu“ a dialogový akt D u věty t neoznačí, přestože tato věta dialogový akt D reprezentuje. Analogicky vztah A.10 znamená pravděpodobnost, že posluchač j neoznačí dialogový akt D u věty t a tato věta tento dialogový akt skutečně nepředstavuje a vztah A.11 reprezentuje pravděpodobnost toho, že posluchač j „chybuje“ a označí dialogový akt D u věty, která D ve skutečnosti nepředstavuje.

Úkolem tedy je nalézt takový odhad parametrů Θ, který maximalizuje věrohodnostní funkci danou jako pravděpodobnost pozorování za podmínky Θ. Vztah A.1 lze tedy přepsat do tvaru

$$\mathcal{L}(\Theta, o_1, o_2, \ldots, o_n) = P(o_1, o_2, \ldots, o_n|\Theta). \tag{A.12}$$

Získáme-li z A.12 podle A.2 odhad $\hat{\Theta}$, pak dosazením do vztahů A.8 – A.11 můžeme určit nejpravděpodobnější trajektorii procesu X podle rozhodovacího kritéria

$$X_t = 1 \Leftrightarrow P(X_t = 1|o_t) > P(X_t = 0|o_t), \tag{A.13}$$

kde

$$P(X_t = 0|o_t) = 1 - P(X_t = 1|o_t), \tag{A.14}$$

přičemž

$$P(X_t = 1|o_t) = \frac{\prod_{j=1}^{m} P(O_t^{(j)}|X_t = 1) \cdot P(X_t = 1)}{P(o_t)}. \tag{A.15}$$

Pravděpodobnost $P(o_t)$ je sice možné vypočítat, nicméně pro dané t je konstantní a lze je tedy pro účely rozhodovacího kritéria A.13 vynechat. Navíc po dosazení do A.12 bude analytické řešení velice obtížné, proto je vhodné využít iterativní numerické výpočty založené na algoritmu EM, který je popsaný v příloze B.

Příloha B

Algoritmus EM [1]

Algoritmus EM (z anglického „Expectation Maximization algorithm“) je iterativní metoda pro nalezení maximálně věrohodného odhadu parametrů Θ pravděpodobnostního rozdělení, kdy tyto parametry závisí na skrytých náhodných proměnných. Algoritmus je založen na vhodném stanovení počátečních hodnot parametrů rozdělení a v následném opakování těchto dvou kroků:

1. Odhad očekávaných („expected“) hodnot skrytých proměnných na základě aktuálních hodnot parametrů rozdělení.

2. Přepočítání parametrů rozdělení tak, aby byla maximalizována věrohodnost pozorovaných proměnných na základě aktuálního odhadu skrytých proměnných.

Algoritmus v těchto krocích probíhá, dokud hodnota věrohodnostní funkce L (významně) roste (definice věrohodnostní funkce a další zde použité termíny a výrazy jsou uvedeny v příloze A).

V naší úloze se tedy pro odhad trajektorie procesu X nejprve stanoví počáteční hodnota parametru $\hat{\Theta}_0$, z níž se pro každou větu vypočte pravděpodobnost, že daná věta představuje dialogový akt D. Ze znalosti hodnot těchto pravděpodobností se určí nový odhad $\hat{\Theta}_1$ tak, aby vzrostla hodnota věrohodnostní funkce, tedy platí

$$\mathcal{L}(\hat{\Theta}_1, o_1, o_2, \ldots, o_n) > \mathcal{L}(\hat{\Theta}_0, o_1, o_2, \ldots, o_n). \tag{B.1}$$

Tento proces se opakuje, dokud není splněna ukončovací podmínka

$$\mathcal{L}(\hat{\Theta}_i, o_1, o_2, \ldots, o_n) - \mathcal{L}(\hat{\Theta}_{i-1}, o_1, o_2, \ldots, o_n) < \epsilon, \tag{B.2}$$

[1] Částečně převzato z [97].

kde ϵ je předem stanovený kladný práh.

K odhadu trajektorie X ve smyslu anotace věty dialogovým aktem D bylo použito následujícího algoritmu[2] (označme $J = \{1, 2, \ldots, m\}$ jako množinu všech posluchačů):

1. Zvol počáteční odhad parametru $\hat{\Theta}_0$.

2. Proveď pravděpodobnostní odhad trajektorie $\hat{X}_i$ na základě parametru $\hat{\Theta}_i$:

   (a) Pro každé t (tj. pro všechny věty) vypočti hodnotu

   $$p_{t1} = \prod_{j \in J} P(O_t^{(j)} | X_t = 1, \hat{\Theta}_i) \cdot P(X_t = 1, \hat{\Theta}_i) \tag{B.3}$$

   kde ve členu $P(O_t^{(j)} | X_t = 1, \hat{\Theta}_i)$ budou využívány aktuální hodnoty parametrů A.8 a A.9 podle toho, zda hodnota $O_t^{(j)}$ pro konkrétní dvojici j, t je rovna 1 či 0, a $P(X_t = 1, \hat{\Theta}_i)$ je rovno aktuální hodnotě $1 - p_X$.

   (b) Analogicky pro každé t vypočti hodnotu

   $$p_{t0} = \prod_{j \in J} P(O_t^{(j)} | X_t = 0, \hat{\Theta}_i) \cdot P(X_t = 0, \hat{\Theta}_i). \tag{B.4}$$

   (c) Pro každé t vypočti pravděpodobnost, že tato věta představuje dialogový akt D, dle vztahu normalizujícího hodnotu p_{t1}, aby součet pravděpodobností obou disjunktních jevů byl 1 (čímž odpadne nutnost výpočtu $p(o_t)$ ze vztahu A.15):

   $$P(X_t = 1 | o_t, \hat{\Theta}_i) = \frac{p_{t1}}{p_{t0} + p_{t1}}. \tag{B.5}$$

3. Z pravděpodobnostního odhadu $\hat{X}_i$ (tj. z hodnot $P(X_t = 1 | o_t, \hat{\Theta}_i)$) aktuálně vypočtených v kroku (2c) vypočti nový odhad parametru $\hat{\Theta}_{i+1}$:

   (a) Nový odhad parametru alternativního rozdělení $A(p_X)$ urči jako

   $$p_x^{\hat{\Theta}_{i+1}} = 1 - P(X_t = 1 | \hat{\Theta}_{i+1}) = 1 - E_t \left\{ P(X_t = 1 | o_t, \hat{\Theta}_i) \right\}, \tag{B.6}$$

[2] Algoritmus lze považovat za EM ačkoliv neprovádí explicitní výpočet věrohodnostní funkce.

neboť střední hodnota aposteriorních pravděpodobností za předpokladu konstantosti $P(o_t)$ pro všechna t je rovna apriorní pravděpodobnosti. Nestranným odhadem střední hodnoty je průměr, tedy:

$$p_x^{\hat{\Theta}_{i+1}} = 1 - \frac{1}{n}\sum_{t=1}^{n} P(X_t = 1|o_t, \hat{\Theta}_i). \tag{B.7}$$

(b) Pro každé j (tj. pro všechny posluchače) vypočti nový odhad parametru $r_x^{(j)}$ (pravděpodobnost A.8) jako poměr součtu hodnot $P(X_t = 1|o_t, \hat{\Theta}_i)$ pro ty věty, u nichž posluchač j zvolil hodnotu 1 (věta představuje dialogový akt D), vůči součtu hodnot $P(X_t = 1|o_t, \hat{\Theta}_i)$ pro všechny věty:

$$\begin{aligned} r_X^{(j)\hat{\Theta}_{i+1}} &= P(O_t^{(j)} = 1|X_t = 1, \hat{\Theta}_{i+1}) \\ &= \frac{\sum_{t=1}^{n} O_t^{(j)} \cdot P(X_t = 1|o_t, \hat{\Theta}_i)}{\sum_{t=1}^{n} P(X_t = 1|o_t, \hat{\Theta}_i)}. \end{aligned} \tag{B.8}$$

Jde v podstatě o zobecnění relativní četnosti případů, kdy j-tý posluchač správně označil větu dialogovým aktem D. Kdyby skutečná trajektorie X byla známá, pak by šlo o speciální případ, v němž by pravděpodobnost vypočtená dle B.8 byla rovna relativní četnosti dané jako počet případů, v nichž j-tý posluchač označil větu dialogovým aktem D, když tato věta dialogový akt D skutečně představovala (tj. počet případů, kdy $O_t^{(j)} = X_t = 1$), dělený celkovým počtem případů, kdy věta skutečně představuje dialogový akt D (tj. $\sum_{t=1}^{n} X_t$).

(c) Analogicky vypočti nový odhad parametru $f_X^{(j)}$ (pravděpodobnost A.10) jako poměr součtu hodnot $P(X_t = 0|o_t, \hat{\Theta}_i)$ pro ty věty, u nichž posluchač j zvolil hodnotu 0 (tedy neoznačil dialogový akt D), vůči součtu hodnot $P(X_t = 0|o_t, \hat{\Theta}_i)$ pro všechny věty:

$$\begin{aligned} f_X^{(j)\hat{\Theta}_{i+1}} &= P(O_t^{(j)} = 0|X_t = 0, \hat{\Theta}_{i+1}) \\ &= \frac{\sum_{t=1}^{n} (1 - O_t^{(j)}) \cdot P(X_t = 1|o_t, \hat{\Theta}_i)}{\sum_{t=1}^{n} (1 - P(X_t = 1|o_t, \hat{\Theta}_i))}. \end{aligned} \tag{B.9}$$

4. Pokud je rozdíl $\hat{\Theta}_{i+1} - \hat{\Theta}_i$ větší než předem stanovený práh, vrať se ke kroku (2). V opačném případě pokračuj dále.

5. Podle rozhodovacího kritéria A.13 urči z posledního získaného odhadu parametru $\hat{\Theta}$ nejpravděpodobnější trajektorii procesu X a algoritmus ukonči.

Příloha C

Statistické charakteristiky

Statistické charakteristiky (charakteristiky náhodné veličiny) jsou vhodně vybrané číselné údaje, které shrnují základní informace o rozdělení pravděpodobnosti náhodné veličiny. Charakteristiky nám o náhodné veličině poskytují pouze základní a hrubou představu, neboť (obvykle) nepostačují k jednoznačnému popisu rozdělení pravděpodobnosti. Naproti tomu rozdělení pravděpodobnosti sice poskytuje jednoznačný popis náhodné veličiny, není však dostatečně přehledné.

Uvažujme náhodnou veličinu X (soubor dat), kde x_i je jedno pozorování takovéto veličiny (jeden prvek v souboru dat), $i = 1 \dots N$, kde N je počet pozorování (počet prvků v souboru dat). Potom mezi nejdůležitější charakteristiky takové veličiny (souboru dat) patří:

střední hodnota – $E(X)$ je parametr rozdělení náhodné veličiny, který je pro diskrétní rozdělení definován jako

$$E(X) = \sum_{i=1}^{N} x_i \cdot P(x_i). \tag{C.1}$$

Pro odhad střední hodnoty využíváme aritmetický průměr (označovaný jako μ nebo také $\bar{x}$), který lze spočítat jako

$$\mu = \frac{1}{N} \sum_{i=1}^{N} x_i. \tag{C.2}$$

rozptyl – $var(X)$ nebo $\sigma^2(X)$ je střední hodnota kvadrátů odchylek od střední hodnoty (jde o druhý centrální moment)

$$var(X) = E[(X - \mu)^2], \tag{C.3}$$

nebo také

$$var(X) = E(X^2) - [E(X)]^2. \tag{C.4}$$

směrodatná odchylka – $\sigma = \sqrt{varX}$ vyjadřuje kvadratický průměr odchylek hodnot od jejich střední hodnoty

$$\sigma = \sqrt{E(X^2) - [E(X)]^2}, \tag{C.5}$$

nebo také

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{C.6}$$

kde $\bar{x}$ je aritmetický průměr.

výběrová směrodatná odchylka

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}, \tag{C.7}$$

kde $\bar{x}$ je aritmetický průměr.

koeficient šikmosti – γ_1 popisuje nesymetrii rozdělení náhodné veličiny

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{E[X - E(X)]^3}{\sqrt{(var(X))^3}}, \tag{C.8}$$

kde μ_3 je třetí centrální moment, σ je směrodatná odchylka, $E(X)$ je střední hodnota a $var(X)$ je rozptyl. Koeficient šikmosti pro normální rozdělení $\gamma_1 = 0$.

koeficient špičatosti – γ_2 porovnává rozdělení náhodné veličiny s normálním rozdělením pravděpodobnosti

$$\gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{E[X - E(X)]^4}{(var(X))^4}, \tag{C.9}$$

kde μ_4 je čtvrtý centrální moment, σ je směrodatná odchylka, $E(X)$ je střední hodnota a $var(X)$ je rozptyl. Koeficient špičatosti pro normální rozdělení $\gamma_2 = 3$, někdy se také udává koeficient špičatosti zmenšený o 3, tj. pro normální rozdělení pak platí $\gamma_2 = 0$. Pokud v této práci budeme zmiňovat koeficient špičatosti, bude to právě tato upravená varianta.

kvantil – Q_p dělí seřazený statistický soubor na několik stejně velkých částí.

percentil – Dělí statistický soubor na setiny, jako k-tý percentil označujeme $Q_{k/100}$.

kvartil – Tři kvartily dělí seřazený statistický soubor na čtyři stejně velké části, tj. 25 % prvků má hodnoty menší než dolní kvartil $Q_{0,25}$ a 75 % prvků hodnoty menší než horní kvartil $Q_{0,75}$. Někdy se také označují jako Q_1 a Q_3.

medián – Kvantil rozdělující seřazený statistický soubor na dvě stejně početné množiny, tj. jedná se o kvantil $Q_{0,5}$.

mezikvartilové rozpětí – Pomocí horního a dolního kvartilu lze zavést mezikvartilové rozpětí, které definujeme jako hodnotu $Q_{0,75} - Q_{0,25}$.

Příloha D

Modifikovaná Thompson Tau metoda

Odlehlé prvky (outliery, z anglického *outliers*) v množině dat jsou definované jako takové prvky, které nejsou statisticky konzistentní s ostatními prvky. Při práci s takovými prvky v datech musíme být opatrní, neboť je velmi komplikované určit, které prvky mají být považovány za tzv. outliery. Pokud outliery v datech identifikujeme, musíme se rozhodnout, jak se s nimi vypořádat: zda je odstranit, zachovat, nebo nahradit nějakou definovanou hodnotou. Protože v naší úloze je možné nalezené outliery odstranit[1], na následujících řádcích popíšeme metodu, která slouží právě k identifikaci a zároveň i k odstranění outlierů.

Uvažujme soubor n měření veličiny x, tj. mějme prvky x_1, x_2, ..., x_n. Pokud seřadíme jednotlivé prvky veličiny x podle hodnoty, jako *podezřelé prvky* můžeme obvykle označit buď několik prvních nebo několik posledních hodnot takto seřazeného souboru dat – outliery jsou většinou největší nebo naopak nejmenší hodnoty dané veličiny.

Metoda *Thompson* τ [110] je statistická metoda odhadující, které podezřelé prvky jsou outliery (a měly by být ze souboru hodnot odstraněny). Postup pro modifikovanou Thompson τ metodu[2] [1] je následující:

1. Vypočítáme střední hodnotu $\bar{x}$ a výběrovou směrodatnou odchylku s běžnými metodami.

[1] Veškeré hodnoty naměřené během akustické analýzy v části 7.3 (u kterých provádíme detekci outlierů) jsou výstupem metod automatického zpracování velkého množství dat. Proto předpokládáme, že většina outlierů představuje chyby měření, případně chyby dalšího zpracování těchto měření.

[2] Modifikace spočívá v rozdílném výpočtu směrodatné odchylky.

Tabulka D.1: Tabulka vybraných hodnot pro modifikované Thompsonovo τ.

N	τ		N	τ		N	τ
3	1.1511		21	1.8891		40	1.9240
4	1.4250		22	1.8926		42	1.9257
5	1.5712		23	1.8957		44	1.9273
6	1.6563		24	1.8985		46	1.9288
7	1.7110		25	1.9011		48	1.9301
8	1.7491		26	1.9035		50	1.9314
9	1.7770		27	1.9057		55	1.9340
10	1.7984		28	1.9078		60	1.9362
12	1.8290		30	1.9114		100	1.9459
14	1.8498		32	1.9146		200	1.9530
16	1.8649		34	1.9174		1000	1.9586
18	1.8764		36	1.9198		5000	1.9597
20	1.8853		38	1.9220		$(\to \infty)$	1.9600

2. Pro každý prvek ze souboru hodnot vypočítáme absolutní hodnotu odchylky od střední hodnoty jako $\delta_i = |d_i| = |x_i - \bar{x}|$.

3. Prvek, jehož odchylka δ_i je největší se stává potenciálním outlierem (tedy podezřelým prvkem).

4. Hodnota modifikovaného Thompsonova τ je vypočítána z kritických hodnot pravděpodobnostní hustotní funkce studentova t rozdělení a je funkcí počtu prvků N množiny dat:

$$\tau = \frac{t_{\alpha/2} \cdot (N-1)}{\sqrt{N}\sqrt{N-2+t_{\alpha/2}^2}},$$

kde

- N je počet prvků v množině dat;
- $t_{\alpha/2}$ je kritická hodnota studentova t rozdělení (hodnoty jsou většinou tabelizovány), za předpokladu, že hladina významnosti $\alpha = 0{,}05$ a počet stupňů volnosti $df = N - 2$.

Vybrané hodnoty pro modifikované Thompsonovo τ jsou pro ilustraci uvedené v tabulce D.1.

5. Použitím jednoduchého pravidla určíme, zda označit podezřelý prvek jako outlier:
    - Pokud $\delta_i > \tau \cdot s$, vyřadíme prvek jako outlier.
    - Pokud $\delta_i \leq \tau \cdot s$, zachováme prvek, není to outlier.

Při využití modifikované metody Thompsonova τ v každém cyklu algoritmu posuzujeme vždy jen jeden prvek, ten s největší odchylkou od střední hodnoty δ_i. Pokud prvek identifikujeme jako outlier, odstraníme ho z množiny dat a pokračujeme znovu od začátku – vypočítáme novou střední hodnotu $\bar{x}$ a směrodatnou odchylku s a prověřujeme další podezřelý prvek. Celý tento proces opakujeme, dokud se v datech objevují outliery.

Příloha E

Wilksova metoda

Definice odlehlých prvků (outlierů) a důvody pro potřebu jejich identifikaci a případné odstranění jsou popsány v příloze D. Mimo to ještě musíme uvést, že pro data s vícerozměrnými prvky máme v zásadě dvě možnosti jak při identifikaci outlierů postupovat:

- Najít outliery pomocí běžných metod[1] v každé dimenzi zvlášť a pak odstranit všechny prvky, jenž byly alespoň v některé dimenzi označeny za outliery.
- Použít metodu pro identifikaci outlierů ve vícerozměrném prostoru, například dále popsaný postup využívající Wilksovu metodu [124].

V naší úloze jsme pro zpracování výsledků akustické analýzy a jejich následné využití při tvorbě akustické části penalizační matice[2] využili právě metodu identifikace a odstranění outlierů ve vícerozměrném prostoru. Jako dimenze tohoto prostoru byly vybrány tři akustické parametry: $F0$, doba trvání a hodnoty RMS, vždy pro ty segmenty řečového signálu, které reprezentují všechny znělé fonémy.

Předpokládejme, že máme vstupní soubor dat X ve formě matice $\mathbf{X}$, jejíž rozměry jsou $n \times p$, kde n je počet řádků (počet prvků ve vstupních datech) a p je počet sloupců (dimenze prvků):

$$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p), \tag{E.1}$$

kde $\mathbf{x}_i$ je sloupcový vektor délky n představující soubor hodnot pro dimenzi i a množina $I = \{i : i = 1 \ldots n\}$ je množina indexů všech sloupcových vektorů tvořících matici $\mathbf{X}$.

[1] Jako příklad uveďme metodu modifikovaného Thompsonova τ popsanou v příloze D.

[2] Postup návrhu akustické penalizační matice je uvedený v části 8.1.2.

Wilksova metoda

Algoritmus využívající Wilksovu metodu k detekci a odstranění outlierů [116] je následující:

1. Nastavíme hladinu významnosti $\alpha = 0{,}05$.
2. Vypočítáme matici $\mathbf{D}$ jako

$$\mathbf{D} = (\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_p), \tag{E.2}$$

   kde $\mathbf{d}_i$ je sloupcový vektor délky n a

$$\forall i : \mathbf{d}_i = \mathbf{x}_i - \bar{x}_i, \tag{E.3}$$

   přičemž rozdílem $\mathbf{x}_i - \bar{x}_i$ vyjadřujeme, že od každého prvku vektoru $\mathbf{x}_i$ odečteme aritmetický průměr hodnot vektoru $\mathbf{x}_i$.
3. Dále pro $\mathbf{X}$ vypočítáme kovarianční matici $\mathbf{S}$.
4. Vypočítáme matici $\mathbf{E}$ jako:

$$\mathbf{E} = \mathbf{D} \cdot \mathbf{S}^{-1} \cdot \mathbf{D}^T \tag{E.4}$$

   a určíme vektor $\mathbf{e}$ obsahující prvky na hlavní diagonále matice $\mathbf{E}$.
5. Stanovíme práh ϵ podle rovnice

$$\epsilon = \frac{p \cdot (n-1)^2 \cdot F_c}{n \cdot (n-p-1) + (n \cdot p \cdot F_c)}, \tag{E.5}$$

   kde F_c je kritická hodnota Fisherovo-Snedecorova F-rozdělení s p a $n-p-1$ stupni volnosti pro hladinu významnosti α.
6. Určíme množinu J jako podmnožinu množiny I:

$$J = \{j \in I : e_j \geq \epsilon\}. \tag{E.6}$$

   Množina J tedy tvoří množinu indexů sloupcových vektorů $\mathbf{x}_j$, které byly identifikované jako outliery.
7. Odstraň ze souboru dat X prvky $\mathbf{x}_j$ pro která platí $j \in J$.
8. Algoritmus ukonči.

Příloha F

Výsledky

Tabulka F.1: Část přepisu reálného rozhovoru mezi seniorem a avatarem. Ve špičatých závorkách uvádíme i zaznamenané neřečové události, znak ∼ pak znamená nedokončený úsek či slovo.

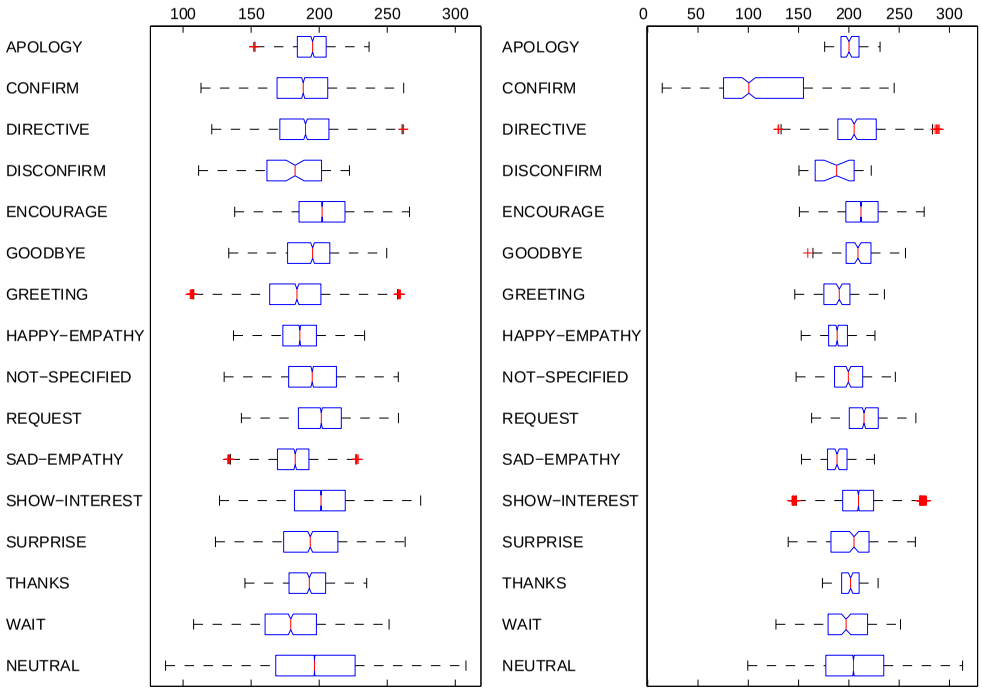
řečník	přepis
	...
AVATAR:	Byli jste letos na houbách?
SENIOR:	No manžel, manžel na houbách byl a nosil obrovský množství hub, to je takovej jeho koníček.
AVATAR:	Manžel chodí sám?
SENIOR:	<EE-HESITATION> buď sám anebo s dětmi, já moc na houby nechodim, já mě <MM-HESITATION> to neza∼ nebaví sbírat houby, ale manžel chodí hodně a někdy chodí s nim ještě buď teď už vnuci taky třeba a hlavně tedy syn, ten nejmladší, ten s nim chodil hodně.
AVATAR:	Řeknete mi ještě něco o téhle fotce?
SENIOR:	No možná jedině ještě to, že ta borovice, pod kterou jsou tam skrčeni, tak ta tenkrát měla tak dva metry, no a teď už jsou to borovice možná deseti no tak deseti, osmi metrové.
SENIOR:	Jak čas ubíhá, tak nejen děti, ale i borovice rostou.
	...

Tabulka F.2: Příklady anotací expresivního korpusu s využitím dialogových aktů metodami prosté většiny a maximální věrohodnosti. U metody prosté většiny uvádíme procento posluchačů, kteří se na daném dialogovém aktu shodli. U metody maximální věrohodnosti uvádíme přesnost pravděpodobnostního odhadu (leží v intervalu $\langle 0, 1 \rangle$). Dále také uvádíme, zda by anotace dané věty byla považována za věrohodnou a mohla by být dále využita.

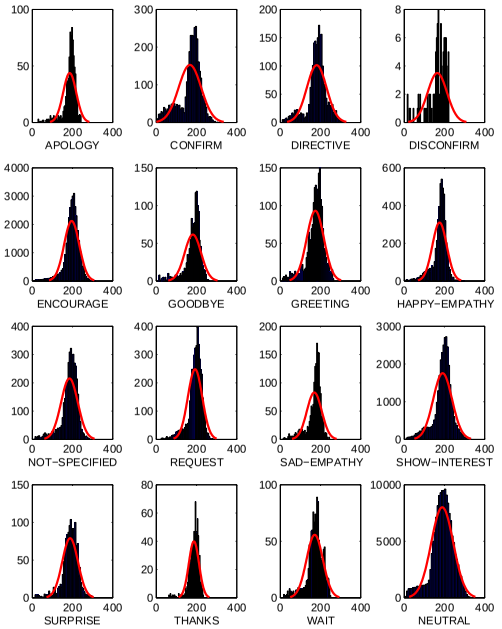
text věty	**prostá většina**	**shoda posluchačů**	**využití**	**maximální věrohodnost**	**přesnost modelu**	**využití**
Co je na ní vidět?	SHOW-INTEREST	42 %	NE	SHOW-INTEREST	0,9999	ANO
Jak se jmenují rodiče?	SHOW-INTEREST	58 %	ANO	SHOW-INTEREST	1,0000	ANO
Bylo tam tenkrát hezké počasí?	ENCOURAGE	67 %	ANO	ENCOURAGE	1,0000	ANO
Hmm, dobře.	CONFIRM	100 %	ANO	CONFIRM	1,0000	ANO
Jste na té fotce vy?	SHOW-INTEREST	58 %	ANO	ENCOURAGE	0,9990	ANO
Vy jste ho znal tedy osobně?	SURPRISE	42 %	NE	SURPRISE	1,0000	ANO
Přejdeme na další fotku.	DIRECTIVE	100 %	ANO	DIRECTIVE	1,0000	ANO
Copak to máme tady?	SHOW-INTEREST	33 %	NE	SHOW-INTEREST	0,3483	NE
Tak podíváme se na další fotku?	REQUEST	75 %	ANO	REQUEST	1,0000	ANO
To jsou ale krásná štěňátka.	HAPPY-EMPATHY	92 %	ANO	HAPPY-EMPATHY	1,0000	ANO
Kdo je kdo?	ENCOURAGE	42 %	NE	SHOW-INTEREST	0,1830	NE
Pěšky?	SURPRISE	42 %	NE	SURPRISE	1,0000	ANO
Tak se podíváme dál.	DIRECTIVE	83 %	ANO	DIRECTIVE	1,0000	ANO
Děkujeme vám za váš čas.	THANKS	92 %	ANO	THANKS	1,0000	ANO
Dobře.	CONFIRM	92 %	ANO	CONFIRM	1,0000	ANO
To nevadí.	CONFIRM	25 %	NE	HAPPY-EMPATHY	0,3041	NE
Znáte někoho dalšího na té fotce?	ENCOURAGE	75 %	ANO	ENCOURAGE	1,0000	ANO
Tak to je fajn.	CONFIRM	58 %	ANO	HAPPY-EMPATHY	1,0000	ANO
To je nějaký zvyk?	SHOW-INTEREST	58 %	ANO	ENCOURAGE	0,9995	ANO
A to je v Belgii zvykem, nechat děti pokřtít?	SHOW-INTEREST	58 %	ANO	ENCOURAGE	1,0000	ANO
Tak se podíváme na další.	REQUEST	83 %	ANO	REQUEST	1,0000	ANO
Na shledanou	GOODBYE	100 %	ANO	GOODBYE	1,0000	ANO
Všichni?	SURPRISE	33 %	NE	SURPRISE	1,0000	ANO
Chudá je tedy hubená.	NOT-SPECIFIED	58 %	ANO	NOT-SPECIFIED	1,0000	ANO
Ale kde?	ENCOURAGE	33 %	NE	NOT-SPECIFIED	0,8091	ANO
Omlouvám se, ale budeme už muset končit.	APOLOGY	100 %	ANO	APOLOGY	1,0000	ANO

Tabulka F.3: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty $F0$ znělých fonémů a fonému /e/, metoda odstranění outlierů WILKS.

	Všechny znělé fonémy					**Foném /e/**				
Dialogový akt	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O
APOLOGY	186 ± 32	- 1,88	4,41	892	5%	198 ± 23	- 1,44	3,57	131	2%
CONFIRM	167 ± 56	- 0,93	0,37	3916	12%	114 ± 53	0,44	- 0,76	355	1%
DIRECTIVE	181 ± 49	- 0,89	1,50	2208	4%	195 ± 54	- 0,99	1,06	475	0%
DISCONFIRM	164 ± 48	- 1,39	1,53	111	5%	175 ± 41	- 1,63	2,62	28	4%
ENCOURAGE	196 ± 38	- 1,37	3,68	33 807	1%	208 ± 32	- 0,86	1,81	6227	0%
GOODBYE	182 ± 40	- 1,71	3,68	1368	6%	209 ± 21	- 0,10	- 0,03	158	1%
GREETING	174 ± 43	- 1,13	2,09	2107	5%	186 ± 30	- 0,54	1,67	230	1%
HAPPY-EMPATHY	176 ± 35	- 1,35	2,68	4969	2%	182 ± 32	- 1,46	3,25	904	0%
NOT-SPECIFIED	184 ± 42	- 1,39	2,70	4010	2%	195 ± 35	- 1,18	2,93	375	1%
REQUEST	193 ± 36	- 1,36	2,78	4415	2%	212 ± 25	- 0,76	1,31	880	1%
SAD-EMPATHY	169 ± 38	- 1,55	2,58	1712	1%	186 ± 21	- 1,32	4,28	232	2%
SHOW-INTEREST	191 ± 47	- 0,99	1,64	32 765	2%	205 ± 45	- 0,77	1,65	5093	0%
SURPRISE	188 ± 38	- 1,01	2,31	1426	2%	201 ± 31	- 0,14	0,58	185	1%
THANKS	187 ± 25	- 1,63	4,67	771	3%	201 ± 12	- 0,17	0,11	144	0%
WAIT	171 ± 41	- 1,11	1,83	1368	9%	192 ± 36	- 0,99	1,05	190	1%
NEUTRAL	189 ± 55	- 0,70	0,88	160 988	3%	201 ± 50	- 0,60	1,06	19 002	0%



Obrázek F.1: Boxplot pro hodnoty $F0$ všech znělých fonémů (vlevo) a všech fonémů /e/ (vpravo), metoda odstranění outlierů TT. Hodnoty uvedené v [Hz].



Obrázek F.2: Histogramy hodnot $F0$ všech znělých fonémů při použití metody WILKS pro odstranění outlierů. Hodnoty jsou uvedené v [Hz].

Tabulka F.4: Statistické charakteristiky, celkový počet vět v korpusu a poměr odstraněných outlierů pro průměrné hodnoty $F0$ celých vět pro znělé fonémy a foném /e/, metoda odstranění outlierů TT.

	Všechny znělé fonémy					**Foném /e/**				
Dialogový akt	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O
APOLOGY	185 ± 12	- 0,28	- 0,48	35	0%	202 ± 10	- 0,05	- 0,63	35	11%
CONFIRM	164 ± 28	- 0,22	- 0,51	698	0%	111 ± 51	0,49	- 0,66	332	0%
DIRECTIVE	181 ± 13	0,19	- 0,41	163	1%	196 ± 16	0,15	0,11	163	2%
DISCONFIRM	151 ± 42	- 1,11	- 0,16	13	0%	181 ± 24	- 0,10	- 1,07	13	15%
ENCOURAGE	195 ± 13	- 0,08	- 0,42	1981	2%	208 ± 18	- 0,08	- 0,22	1895	5%
GOODBYE	183 ± 13	- 0,21	- 0,77	99	2%	214 ± 20	0,18	- 0,63	99	0%
GREETING	175 ± 11	0,32	- 0,01	98	5%	190 ± 16	- 0,14	- 0,27	98	3%
HAPPY-EMPATHY	174 ± 13	0,01	- 0,29	482	3%	186 ± 13	- 0,20	- 0,15	459	12%
NOT-SPECIFIED	183 ± 15	0,14	- 0,55	263	2%	196 ± 17	0,07	- 0,34	189	12%
REQUEST	192 ± 10	- 0,04	- 0,17	278	2%	214 ± 18	- 0,06	- 0,31	277	1%
SAD-EMPATHY	166 ± 11	0,05	- 0,18	168	5%	187 ± 11	0,41	- 0,18	160	8%
SHOW-INTEREST	191 ± 13	- 0,02	- 0,35	2439	3%	205 ± 21	- 0,11	- 0,35	2242	8%
SURPRISE	185 ± 18	- 0,45	- 0,11	104	2%	199 ± 24	0,19	- 0,81	94	4%
THANKS	184 ± 11	- 0,38	- 0,33	52	0%	201 ± 11	- 0,08	- 0,56	52	2%
WAIT	171 ± 8	0,52	- 0,05	51	0%	188 ± 18	- 0,26	0,02	51	4%
NEUTRAL	189 ± 11	0,04	- 0,37	3964	3%	202 ± 23	- 0,06	- 0,29	3922	4%

Tabulka F.5: Statistické charakteristiky, celkový počet vět v korpusu a poměr odstraněných outlierů pro maximální hodnoty $F0$ celých vět pro znělé fonémy a foném /e/, metoda odstranění outlierů TT.

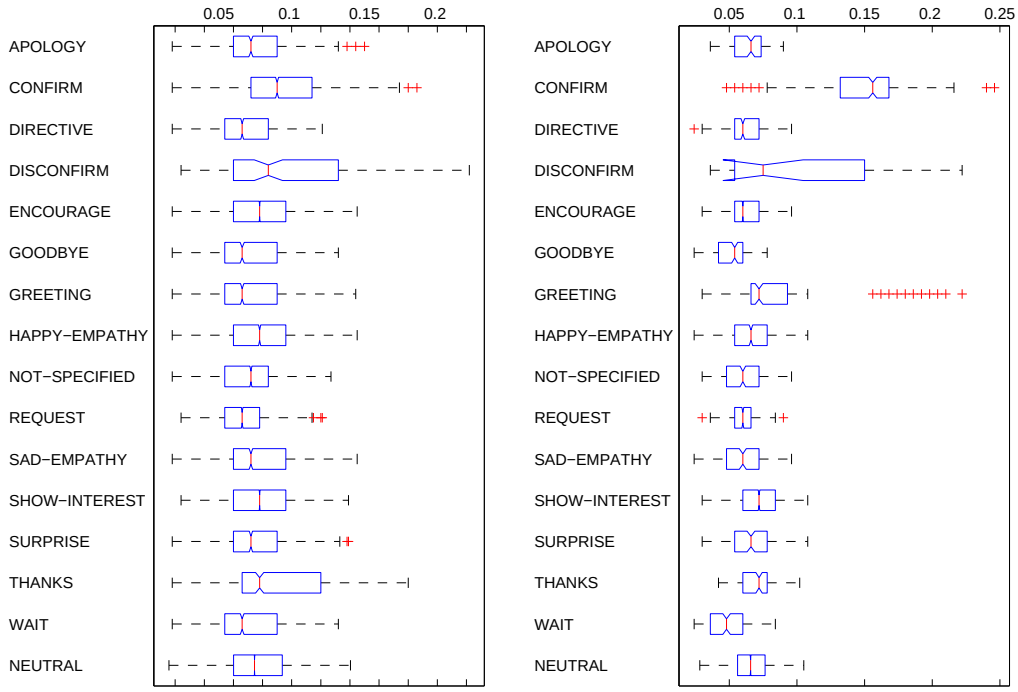
	Všechny znělé fonémy					**Foném /e/**				
Dialogový akt	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O
APOLOGY	221 ± 14	- 0,06	- 0,89	35	3%	217 ± 14	0,02	- 0,47	35	6%
CONFIRM	214 ± 26	0,36	- 0,58	698	1%	112 ± 52	0,51	- 0,67	332	0%
DIRECTIVE	234 ± 30	0,61	- 0,79	163	0%	230 ± 33	0,54	- 0,90	163	0%
DISCONFIRM	208 ± 9	0,12	- 1,09	13	23%	201 ± 14	0,27	- 1,53	13	23%
ENCOURAGE	237 ± 21	0,07	- 0,45	1981	1%	227 ± 23	- 0,03	- 0,51	1895	2%
GOODBYE	221 ± 15	0,20	- 0,81	99	1%	217 ± 16	0,27	- 0,66	99	1%
GREETING	236 ± 18	- 0,35	- 0,80	98	3%	199 ± 11	0,07	0,26	98	11%
HAPPY-EMPATHY	208 ± 16	0,42	- 0,34	482	3%	194 ± 15	0,32	- 0,15	459	7%
NOT-SPECIFIED	226 ± 22	0,16	- 0,08	263	1%	201 ± 21	0,19	- 0,46	189	8%
REQUEST	232 ± 18	- 0,09	- 0,45	278	1%	226 ± 19	- 0,11	- 0,38	277	1%
SAD-EMPATHY	197 ± 11	0,29	- 0,19	168	4%	189 ± 11	0,28	- 0,54	160	8%
SHOW-INTEREST	250 ± 27	0,14	- 0,63	2439	0%	230 ± 33	0,46	- 0,48	2242	2%
SURPRISE	229 ± 27	0,14	- 0,08	104	0%	209 ± 28	0,03	- 0,94	94	3%
THANKS	211 ± 11	- 0,21	- 0,15	52	0%	206 ± 12	0,06	- 0,60	52	2%
WAIT	228 ± 12	0,07	- 0,78	51	4%	225 ± 14	- 0,31	- 0,60	51	8%
NEUTRAL	286 ± 17	- 0,05	- 0,31	3964	2%	252 ± 32	- 0,14	- 0,56	3922	2%

Tabulka F.6: Statistické charakteristiky, celkový počet vět v korpusu a poměr odstraněných outlierů pro minimální hodnoty $F0$ celých vět pro znělé fonémy a foném /e/, metoda odstranění outlierů TT.

	Všechny znělé fonémy					**Foném /e/**				
Dialogový akt	μ $\pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O	μ $\pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O
APOLOGY	71 ± 41	- 0,01	- 1,31	35	0%	183 ± 22	- 1,07	0,08	35	9%
CONFIRM	92 ± 51	0,28	- 0,94	698	0%	109 ± 50	0,53	- 0,55	332	0%
DIRECTIVE	69 ± 30	0,06	- 0,60	163	2%	148 ± 56	- 0,63	- 1,18	163	0%
DISCONFIRM	69 ± 44	0,57	- 0,56	13	0%	143 ± 60	- 0,77	- 0,56	13	0%
ENCOURAGE	108 ± 44	- 0,35	- 0,78	1981	0%	185 ± 31	- 0,50	- 0,06	1895	4%
GOODBYE	86 ± 46	0,08	- 1,14	99	0%	210 ± 24	- 0,24	- 0,48	99	1%
GREETING	70 ± 42	0,56	- 0,78	98	0%	181 ± 22	- 0,62	- 0,12	98	9%
HAPPY-EMPATHY	102 ± 32	- 0,35	- 0,18	482	2%	184 ± 12	- 0,05	- 0,17	459	25%
NOT-SPECIFIED	89 ± 42	0,01	- 0,92	263	0%	192 ± 15	0,02	- 0,37	189	21%
REQUEST	103 ± 40	- 0,51	- 0,52	278	0%	203 ± 23	- 0,39	- 0,20	277	7%
SAD-EMPATHY	87 ± 34	- 0,11	- 0,44	168	0%	185 ± 13	- 0,08	- 0,02	160	10%
SHOW-INTEREST	94 ± 42	0,13	- 0,57	2439	0%	183 ± 46	- 0,54	0,24	2242	2%
SURPRISE	105 ± 42	- 0,46	- 0,53	104	0%	190 ± 28	- 0,16	- 0,30	94	5%
THANKS	112 ± 48	- 0,43	- 1,06	52	0%	195 ± 12	- 0,50	- 0,05	52	0%
WAIT	56 ± 26	0,55	- 0,66	51	2%	151 ± 35	- 0,39	- 0,08	51	2%
NEUTRAL	41 ± 23	0,64	- 0,52	3964	5%	149 ± 50	- 0,50	- 0,17	3922	1%

Tabulka F.7: Statistické charakteristiky, celkový počet vět v korpusu a poměr odstraněných outlierů pro rozsah hodnot $F0$ celých vět pro znělé fonémy a foném /e/, metoda odstranění outlierů TT.

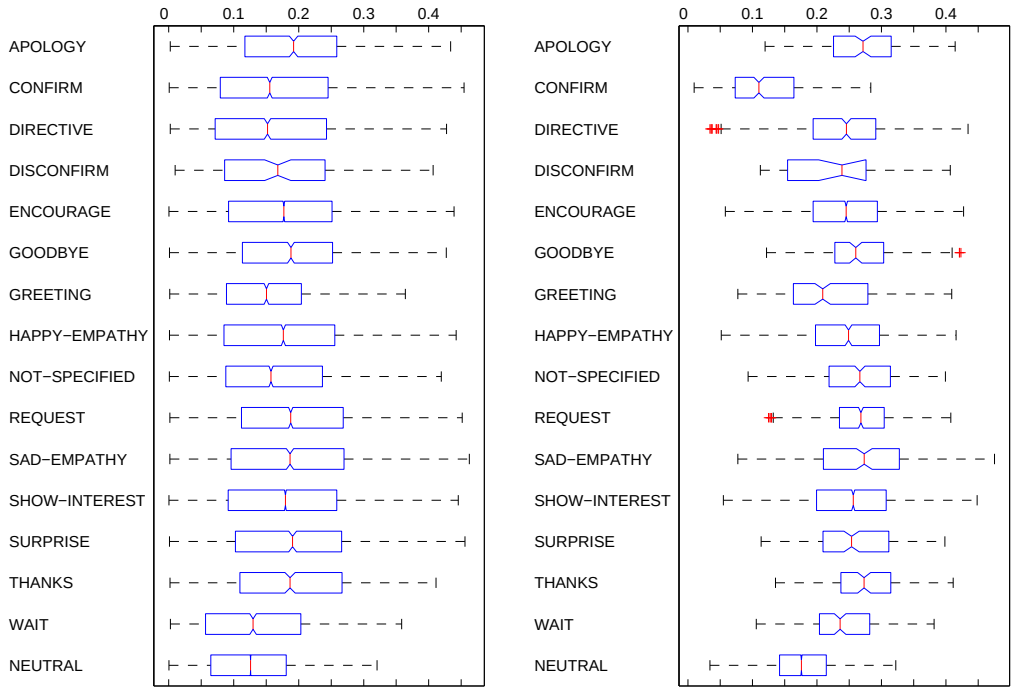
	Všechny znělé fonémy					**Foném /e/**				
Dialogový akt	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O	$\mu \pm\sigma$ [Hz]	γ_1	γ_2	N_P	P_O
APOLOGY	149 ± 43	0,06	- 1,14	35	0%	35 ± 25	0,81	- 0,44	35	9%
CONFIRM	123 ± 55	- 0,07	- 0,75	698	0%	71 ± 38	0,72	- 0,84	332	97%
DIRECTIVE	163 ± 45	- 0,04	- 0,74	163	0%	81 ± 81	0,77	- 0,97	163	0%
DISCONFIRM	117 ± 60	- 0,01	- 0,86	13	0%	38 ± 25	- 0,02	- 1,58	13	46%
ENCOURAGE	130 ± 50	0,24	- 0,71	1981	0%	48 ± 34	0,66	- 0,42	1895	18%
GOODBYE	135 ± 45	- 0,04	- 1,06	99	0%	14 ± 13	0,78	- 0,63	99	52%
GREETING	166 ± 46	- 0,32	- 0,71	98	2%	52 ± 25	0,22	- 0,49	98	48%
HAPPY-EMPATHY	108 ± 34	0,44	- 0,51	482	1%	37 ± 36	0,93	- 0,49	459	37%
NOT-SPECIFIED	137 ± 44	0,06	- 0,55	263	1%	38 ± 37	1,14	0,21	189	52%
REQUEST	129 ± 43	0,32	- 0,52	278	0%	16 ± 10	0,46	- 0,53	277	19%
SAD-EMPATHY	113 ± 34	0,20	- 0,55	168	1%	15 ± 11	0,42	- 0,79	160	75%
SHOW-INTEREST	156 ± 55	- 0,13	- 0,57	2439	0%	66 ± 58	0,85	- 0,41	2242	31%
SURPRISE	124 ± 50	- 0,05	- 0,47	104	0%	24 ± 14	0,13	- 0,72	94	56%
THANKS	99 ± 50	0,47	- 0,95	52	0%	9 ± 6	0,59	- 0,62	52	17%
WAIT	173 ± 29	- 0,20	- 0,57	51	4%	70 ± 33	0,21	- 0,47	51	6%
NEUTRAL	244 ± 30	- 0,32	- 0,45	3964	3%	104 ± 55	0,35	- 0,51	3922	8%



Obrázek F.3: Boxplot pro dobu trvání všech fonémů (vlevo) a fonému */e/* (vpravo), metoda odstranění outlierů TT. Hodnoty jsou uvedené v [s].

Tabulka F.8: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro dobu trvání všech fonémů a fonému /e/, metoda odstranění outlierů WILKS.

	Všechny znělé fonémy					**Všechny fonémy *e***				
Dialogový akt	$\mu \pm\sigma$ [ms]	γ_1	γ_2	N_P	P_O	$\mu \pm\sigma$ [ms]	γ_1	γ_2	N_P	P_O
APOLOGY	87 ± 41	1,80	4,20	1189	1%	66 ± 16	1,58	6,97	131	2%
CONFIRM	118 ± 65	1,74	3,78	4635	1%	144 ± 43	- 0,55	0,51	355	0%
DIRECTIVE	88 ± 50	1,97	4,64	3003	1%	73 ± 37	1,81	2,95	475	0%
DISCONFIRM	108 ± 66	1,77	3,96	145	2%	111 ± 75	1,46	2,01	28	4%
ENCOURAGE	90 ± 42	1,72	3,77	47 224	0%	79 ± 43	2,19	4,41	6227	0%
GOODBYE	80 ± 40	2,31	7,56	1743	1%	51 ± 12	- 0,11	- 0,94	158	1%
GREETING	86 ± 45	1,63	2,77	2732	1%	94 ± 53	1,25	0,02	230	0%
HAPPY-EMPATHY	94 ± 50	1,87	4,00	6891	0%	76 ± 36	1,84	3,24	904	0%
NOT-SPECIFIED	85 ± 44	2,11	5,39	5465	1%	70 ± 35	2,42	6,13	375	1%
REQUEST	87 ± 44	1,81	4,07	6049	1%	66 ± 24	2,62	8,97	880	1%
SAD-EMPATHY	96 ± 55	1,87	3,72	2408	1%	65 ± 25	2,18	7,77	232	2%
SHOW-INTEREST	91 ± 42	1,76	3,96	44 394	0%	84 ± 42	2,00	3,84	5093	0%
SURPRISE	89 ± 43	1,77	3,65	1870	1%	73 ± 32	2,46	7,33	185	2%
THANKS	99 ± 47	1,45	2,46	989	0%	70 ± 13	0,12	- 0,01	144	0%
WAIT	74 ± 32	1,50	4,32	2033	1%	51 ± 19	1,38	3,09	190	1%
NEUTRAL	89 ± 42	1,91	4,94	218 700	0%	75 ± 31	2,36	6,75	19 002	0%



Obrázek F.4: Boxplot pro hodnoty RMS všech fonémů (vlevo) a fonému /*e*/ (vpravo), metoda odstranění outlierů TT.

Tabulka F.9: Statistické charakteristiky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty RMS všech fonémů a fonému */e/*, metoda odstranění outlierů WILKS.

	Všechny fonémy					**Foném /e/**				
Dialogový akt	$\mu \pm \sigma$	γ_1	γ_2	N_P	P_O	$\mu \pm \sigma$	γ_1	γ_2	N_P	P_O
APOLOGY	0,19 ±0,10	0,06	- 0,57	1189	0%	0,27 ±0,07	- 0,06	- 0,37	131	0%
CONFIRM	0,18 ±0,12	0,61	- 0,27	4878	0%	0,14 ±0,08	1,02	0,42	355	0%
DIRECTIVE	0,17 ±0,11	0,49	- 0,37	3006	0%	0,24 ±0,09	- 0,34	- 0,07	475	0%
DISCONFIRM	0,17 ±0,10	0,23	- 0,56	147	0%	0,23 ±0,08	0,30	- 0,58	28	0%
ENCOURAGE	0,18 ±0,11	0,21	- 0,63	47 292	0%	0,24 ±0,07	- 0,08	- 0,34	6227	0%
GOODBYE	0,18 ±0,10	- 0,02	- 0,64	1744	0%	0,27 ±0,07	0,46	- 0,19	158	0%
GREETING	0,15 ±0,09	0,42	0,18	2734	0%	0,23 ±0,07	0,52	- 0,71	230	0%
HAPPY-EMPATHY	0,17 ±0,10	0,13	- 0,95	6896	0%	0,24 ±0,08	- 0,39	- 0,31	904	0%
NOT-SPECIFIED	0,17 ±0,10	0,31	- 0,57	5468	0%	0,26 ±0,07	- 0,52	0,06	375	0%
REQUEST	0,19 ±0,10	0,06	- 0,81	6058	0%	0,27 ±0,06	- 0,28	0,14	880	0%
SAD-EMPATHY	0,19 ±0,11	0,23	- 0,78	2408	0%	0,27 ±0,08	0,06	- 0,44	232	0%
SHOW-INTEREST	0,18 ±0,11	0,19	- 0,74	44 481	0%	0,25 ±0,08	- 0,28	- 0,46	5093	0%
SURPRISE	0,19 ±0,11	0,13	- 0,69	1874	0%	0,25 ±0,07	- 0,22	- 0,39	185	0%
THANKS	0,19 ±0,09	0,08	- 1,00	989	0%	0,28 ±0,06	0,05	- 0,54	144	0%
WAIT	0,14 ±0,09	0,33	- 0,74	2037	0%	0,24 ±0,06	0,30	- 0,13	190	0%
NEUTRAL	0,13 ±0,08	0,42	- 0,21	218 700	0%	0,18 ±0,06	0,28	0,29	19 002	0%

Tabulka F.10: Střední hodnoty a směrodatné odchylky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty formantů fonému /*a*/, metoda odstranění outlierů TT.

a	**F1**		**F2**		**F3**		**Všechny formanty**	
Dialogový akt	μ	σ	μ	σ	μ	σ	N_P	P_O
APOLOGY	724	207	1448	206	2766	118	52	21,2%
CONFIRM	618	257	1393	227	2419	506	180	3,9%
DIRECTIVE	664	123	1643	222	2869	434	279	12,2%
DISCONFIRM	758	187	1599	198	2906	168	6	16,7%
ENCOURAGE	668	199	1473	263	2748	618	3354	1,5%
GOODBYE	730	104	1731	74	2952	327	159	27,7%
GREETING	603	259	1415	314	2566	679	170	0,6%
HAPPY-EMPATHY	693	96	1555	154	2813	170	418	29,2%
NOT-SPECIFIED	644	174	1525	274	2804	556	550	3,1%
REQUEST	696	73	1692	91	3023	417	477	22,4%
SAD-EMPATHY	547	234	1422	258	2532	495	135	3,7%
SHOW-INTEREST	631	210	1489	333	2719	678	3558	1,2%
SURPRISE	613	228	1387	296	2613	677	166	1,8%
THANKS	673	182	1495	162	2720	467	97	9,3%
WAIT	659	88	1695	205	2508	249	3	0,0%
NEUTRAL	419	183	1654	354	2747	348	13 056	17,6%

Tabulka F.11: Střední hodnoty a směrodatné odchylky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty formantů fonému /*e*/, metoda odstranění outlierů TT.

e	**F1**		**F2**		**F3**		**Všechny formanty**	
Dialogový akt	μ	σ	μ	σ	μ	σ	N_P	P_O
APOLOGY	487	112	1873	230	2808	170	131	19,1%
CONFIRM	353	102	1913	195	2810	137	355	20,8%
DIRECTIVE	468	131	1824	431	2813	252	475	16,2%
DISCONFIRM	516	104	1878	185	2893	147	28	21,4%
ENCOURAGE	463	113	2077	288	2995	151	6225	29,5%
GOODBYE	552	134	1861	63	2941	69	158	32,9%
GREETING	522	95	2047	175	3153	410	230	17,4%
HAPPY-EMPATHY	490	86	2020	262	3068	469	904	5,8%
NOT-SPECIFIED	476	108	2031	263	3102	417	374	11,8%
REQUEST	461	131	1863	479	2994	535	879	0,8%
SAD-EMPATHY	428	119	2054	311	2902	112	232	28,0%
SHOW-INTEREST	485	104	2056	216	2971	117	5092	31,2%
SURPRISE	489	100	2004	236	2947	113	185	29,2%
THANKS	498	122	2262	262	3406	541	144	4,2%
WAIT	413	79	2370	180	3377	549	190	4,7%
NEUTRAL	408	180	1645	359	2736	358	18 991	18,3%

Tabulka F.12: Střední hodnoty a směrodatné odchylky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty formantů fonému */i/*, metoda odstranění outlierů TT.

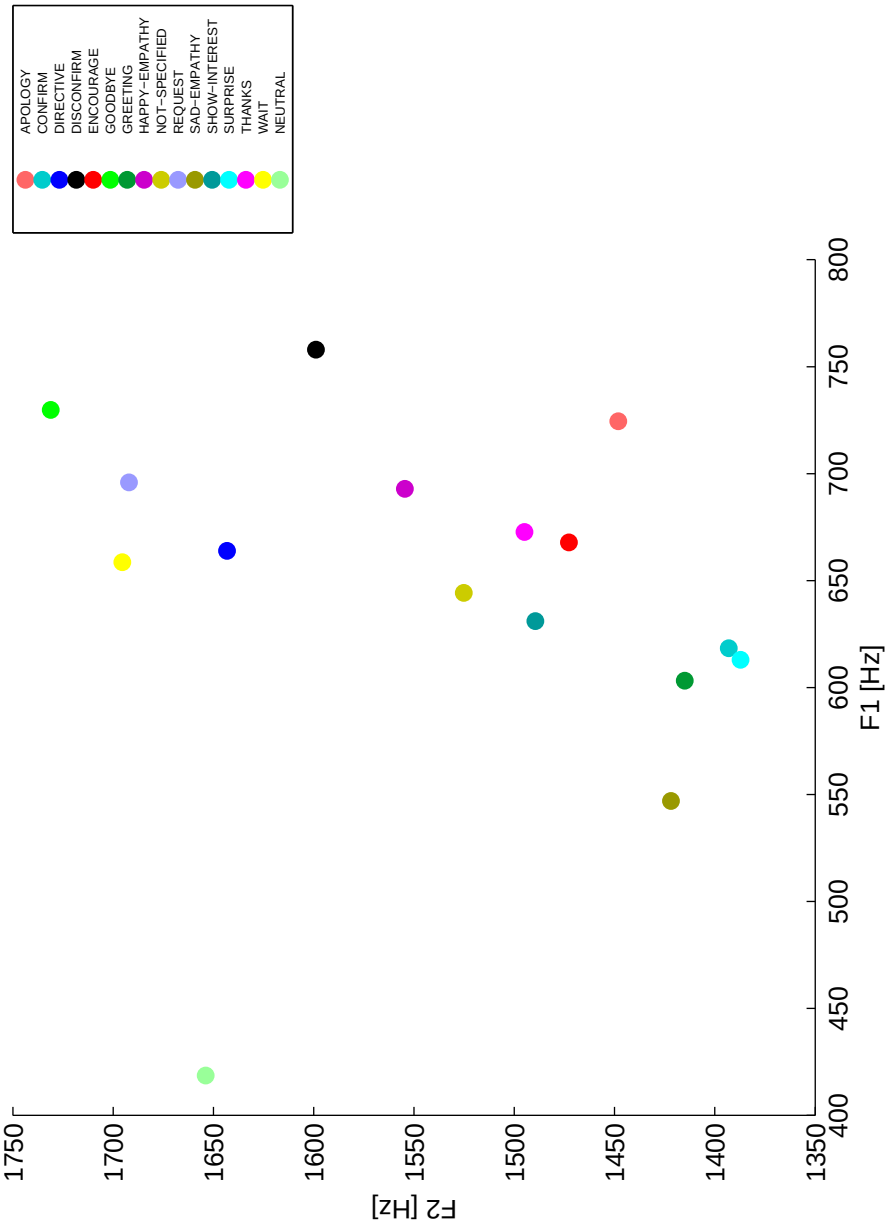
i	**F1**		**F2**		**F3**		**Všechny formanty**	
Dialogový akt	μ	σ	μ	σ	μ	σ	N_P	P_O
APOLOGY	320	45	2347	136	3224	505	56	16,1%
CONFIRM	311	30	2365	149	2989	122	14	14,3%
DIRECTIVE	331	53	2466	183	3292	440	25	12,0%
DISCONFIRM	358	50	2349	217	3492	628	7	0,0%
ENCOURAGE	344	56	2408	201	3246	440	2681	14,2%
GOODBYE	320	48	2375	103	2970	84	136	27,9%
GREETING	330	52	2382	194	3085	162	215	24,2%
HAPPY-EMPATHY	359	55	2293	271	3214	512	117	4,3%
NOT-SPECIFIED	339	49	2382	168	3197	443	360	14,2%
REQUEST	321	38	2410	69	3192	396	67	20,9%
SAD-EMPATHY	362	70	2191	268	3242	546	57	10,5%
SHOW-INTEREST	347	64	2327	208	3004	146	2525	26,6%
SURPRISE	343	60	2325	278	3039	209	108	13,0%
THANKS	331	73	2456	183	3107	328	8	12,5%
WAIT	329	44	2309	330	3308	531	220	9,5%
NEUTRAL	409	184	1652	363	2743	364	13 342	17,6%

Tabulka F.13: Střední hodnoty a směrodatné odchylky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty formantů fonému */o/*, metoda odstranění outlierů TT.

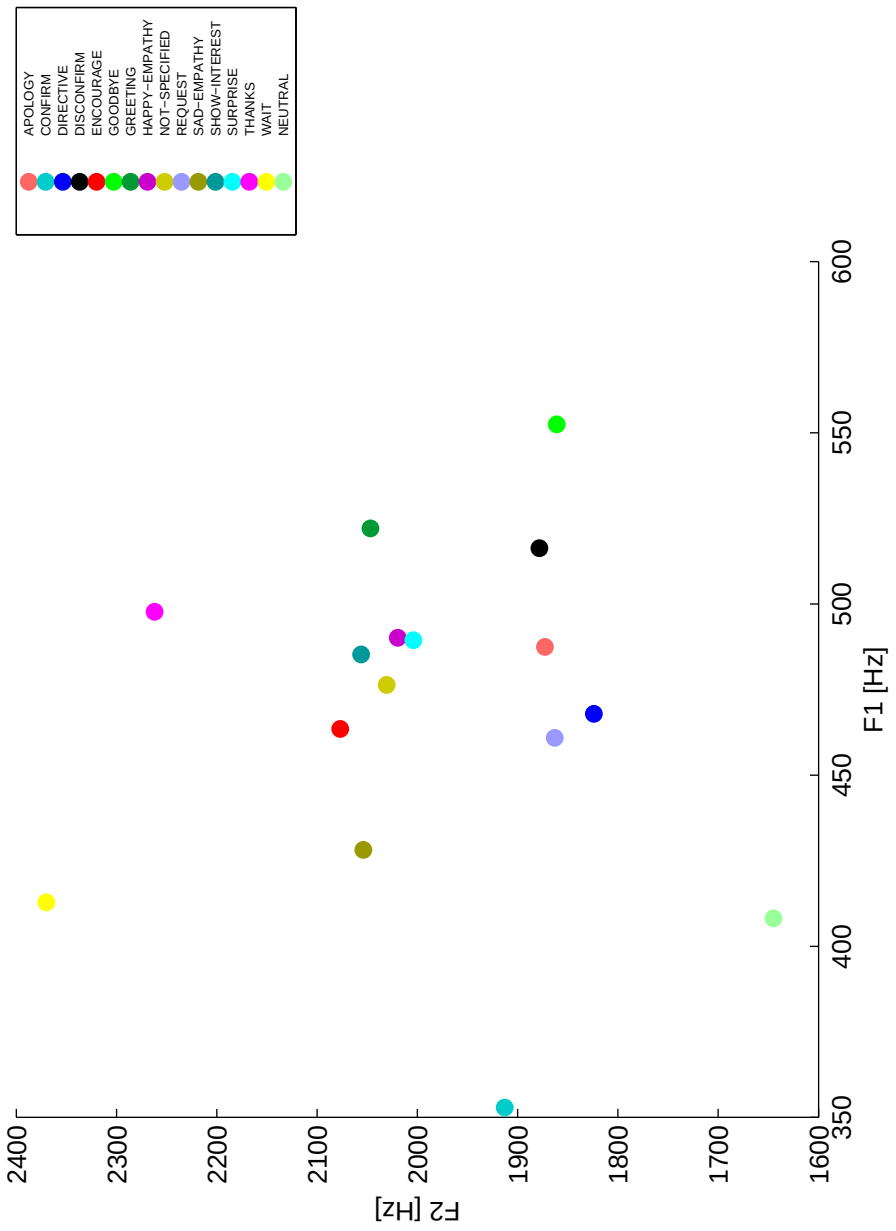
o	**F1**		**F2**		**F3**		**Všechny formanty**	
Dialogový akt	μ	σ	μ	σ	μ	σ	N_P	P_O
APOLOGY	396	99	1153	236	2853	441	111	4,5%
CONFIRM	436	76	1405	147	2702	135	663	13,7%
DIRECTIVE	358	62	1083	73	2724	202	129	23,3%
DISCONFIRM	467	163	1268	211	2797	112	10	10,0%
ENCOURAGE	417	76	1194	193	2861	155	3338	18,5%
GOODBYE	402	126	1006	75	2741	226	145	33,1%
GREETING	415	58	1295	348	2830	144	106	10,4%
HAPPY-EMPATHY	455	71	1371	182	2742	144	745	17,2%
NOT-SPECIFIED	420	79	1146	226	2861	153	486	22,0%
REQUEST	376	51	1075	72	2808	170	394	33,0%
SAD-EMPATHY	414	95	1289	159	2781	147	271	18,5%
SHOW-INTEREST	419	74	1282	225	2860	174	3820	18,3%
SURPRISE	440	77	1228	194	2832	158	183	16,9%
THANKS	395	100	1021	88	3143	508	17	11,8%
WAIT	426	50	1038	52	2800	96	139	18,0%
NEUTRAL	408	180	1654	371	2752	366	13 843	18,0%

Tabulka F.14: Střední hodnoty a směrodatné odchylky, celkový počet segmentů v korpusu a poměr odstraněných outlierů pro hodnoty formantů fonému /*u*/, metoda odstranění outlierů TT.

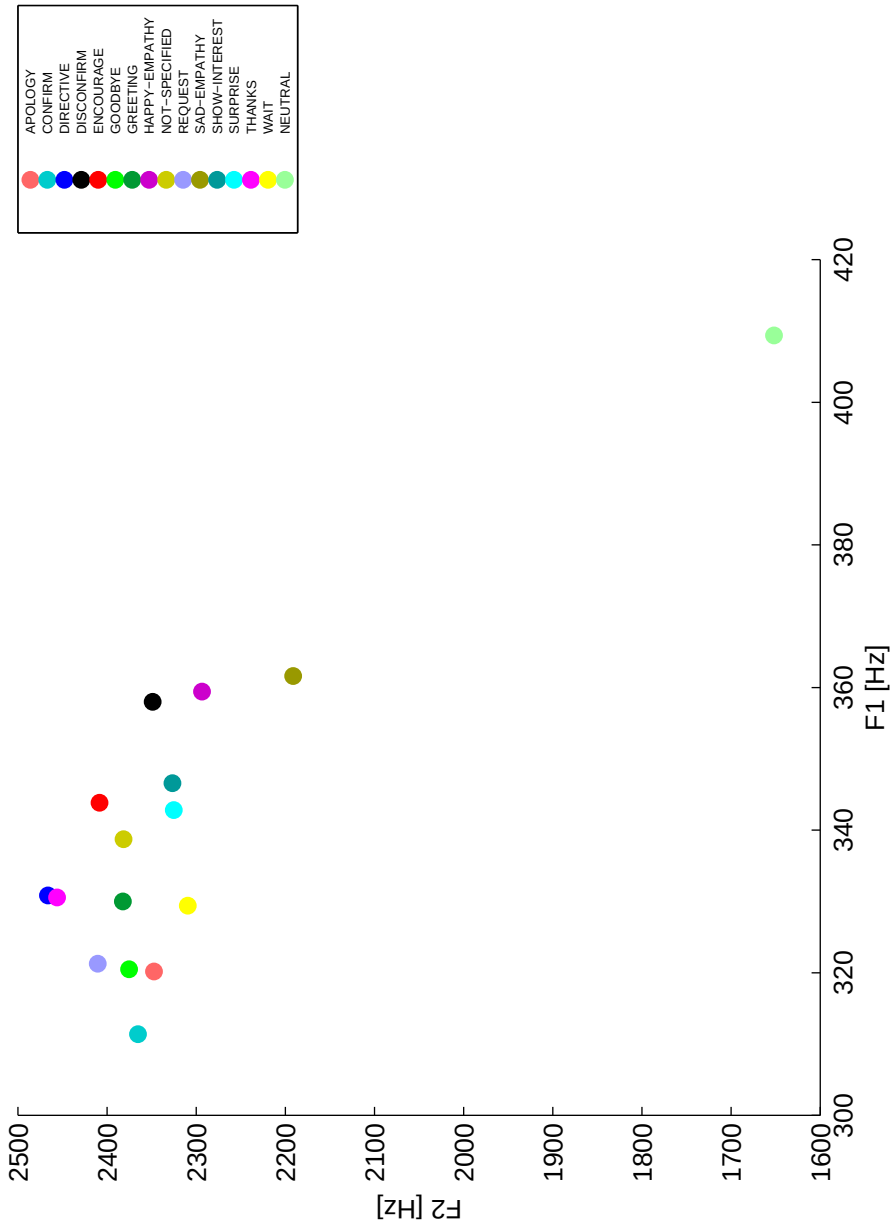
u	**F1**		**F2**		**F3**		**Všechny formanty**	
Dialogový akt	μ	σ	μ	σ	μ	σ	N_P	P_O
APOLOGY	254	30	1108	93	2508	179	64	25,0%
CONFIRM	266	32	1110	105	2830	505	322	13,4%
DIRECTIVE	213	29	909	139	2750	282	32	21,9%
DISCONFIRM	288	0	1198	0	2269	0	1	0,0%
ENCOURAGE	279	41	1037	278	2721	278	751	20,8%
GREETING	277	32	1409	343	2743	239	211	17,1%
HAPPY-EMPATHY	265	36	1104	201	2566	114	116	25,0%
NOT-SPECIFIED	277	41	1024	248	2587	227	61	18,0%
REQUEST	260	27	794	135	2914	337	149	14,8%
SAD-EMPATHY	257	35	1088	168	2478	164	71	21,1%
SHOW-INTEREST	266	39	1103	260	2738	359	670	16,9%
SURPRISE	259	32	1121	352	2725	526	28	0,0%
THANKS	311	33	1115	115	3005	584	55	7,3%
WAIT	308	40	1023	238	2827	180	94	18,1%
NEUTRAL	418	189	1667	367	2760	354	5958	17,0%



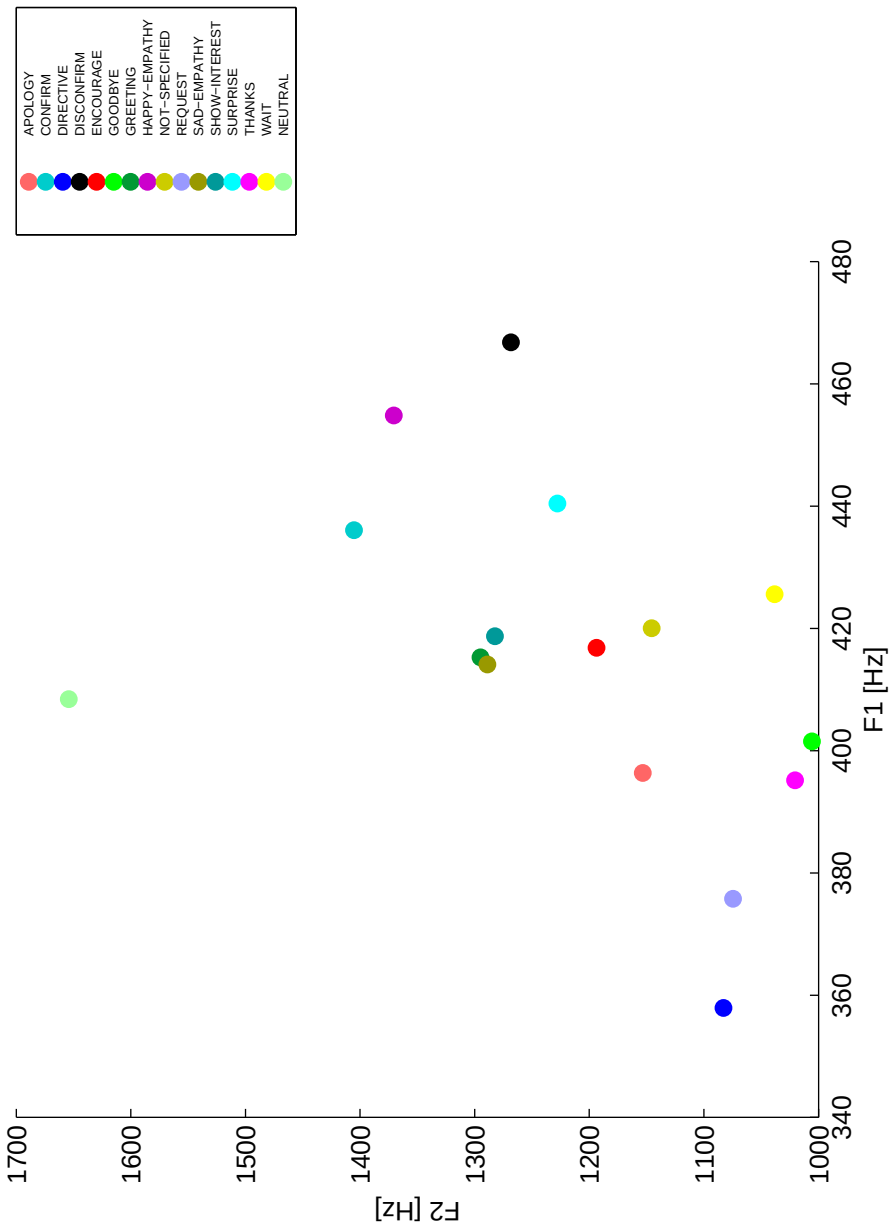
Obrázek F.5: Hodnoty formantových frekvencí v rovině $F1 \times F2$ pro samohlásku /a/.



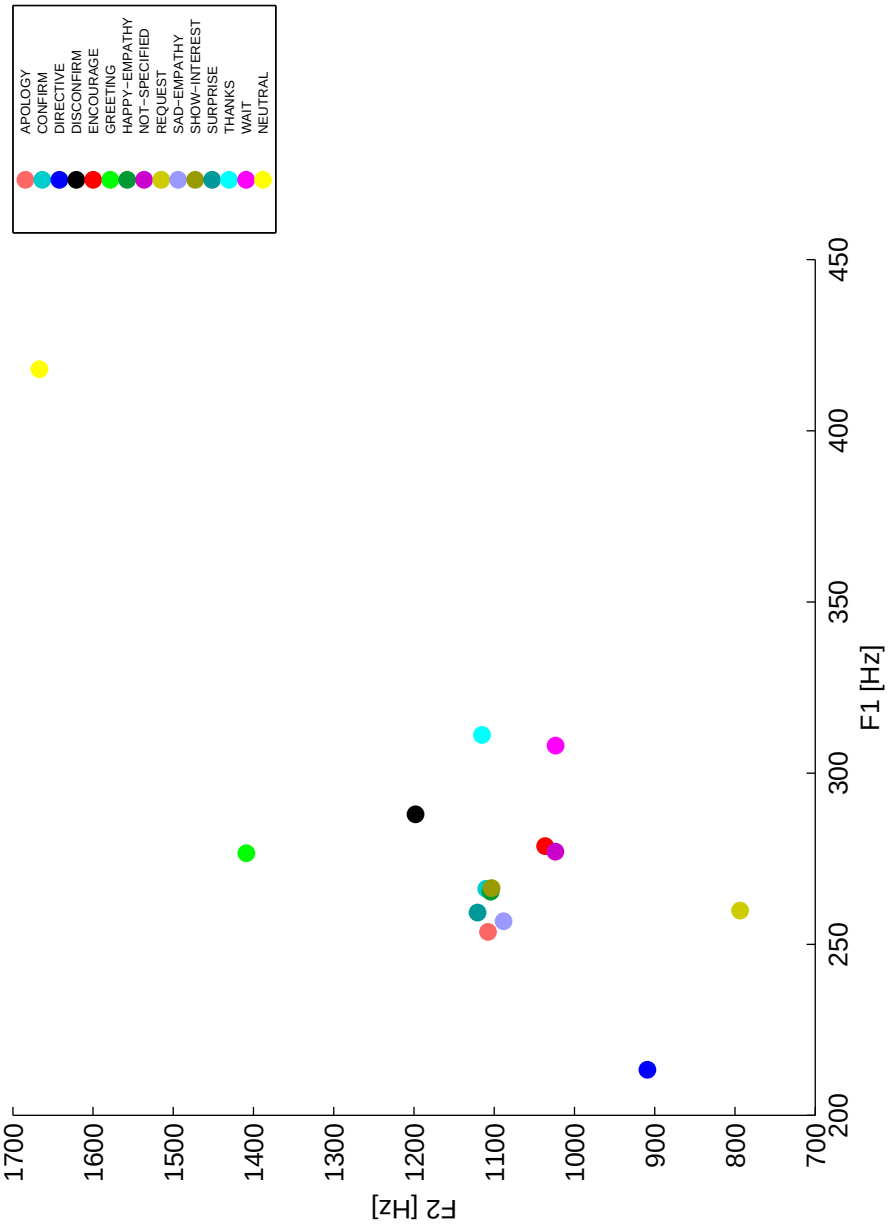
Obrázek F.6: Hodnoty formantových frekvencí v rovině $F1 \times F2$ pro samohlásku /e/.



Obrázek F.7: Hodnoty formantových frekvencí v rovině $F1 \times F2$ pro samohlásku /*i*/.



Obrázek F.8: Hodnoty formantových frekvencí v rovině $F1 \times F2$ pro samohlásku */o/*.



Obrázek F.9: Hodnoty formantových frekvencí v rovině $F1 \times F2$ pro samohlásku /u/.

Tabulka F.15: Porovnání p-hodnot pro $F0$ všech znělých fonémů dosažené testem ANOVA mezi jednotlivými dvojicemi dialogových aktů. Tučně jsou označené hodnoty větší než 0,05, které znamenají statistickou nevýznamnost rozdílu středních hodnot.

dialogový akt	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	NEUTRAL
APOLOGY	**1,00**	<0,01	**0,07**	<0,01	<0,01	**0,55**	<0,01	<0,01	**0,29**	<0,01	<0,01	<0,01	**0,94**	<0,01	<0,01	<0,01
CONFIRM	<0,01	**1,00**	0,03	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**0,42**	<0,01	<0,01
DIRECTIVE	**0,07**	0,03	**1,00**	<0,01	<0,01	0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	0,05	**0,40**	<0,01	<0,01
DISCONFIRM	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	**0,48**	0,04	<0,01	<0,01	**0,90**	<0,01	<0,01	<0,01	**0,46**	<0,01
ENCOURAGE	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	0,00	<0,01	<0,01	<0,01	**0,91**	<0,01	<0,01	<0,01	<0,01
GOODBYE	**0,55**	<0,01	0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	**0,58**	<0,01	<0,01	<0,01	**0,54**	<0,01	<0,01	<0,01
GREETING	<0,01	<0,01	<0,01	**0,48**	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	0,02	<0,01	<0,01	<0,01	<0,01	<0,01
HAPPY-EMPATHY	<0,01	<0,01	<0,01	0,04	0,00	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	0,00	<0,01	<0,01	<0,01	<0,01
NOT-SPECIFIED	**0,29**	<0,01	<0,01	<0,01	<0,01	**0,58**	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	**0,19**	<0,01	<0,01	<0,01
OTHER	<0,01	0,03	**0,08**	<0,01	**0,50**	0,02	<0,01	<0,01	**0,07**	**0,29**	<0,01	**0,56**	**0,08**	<0,01	<0,01	**0,46**
REQUEST	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01
SAD-EMPATHY	<0,01	<0,01	<0,01	**0,90**	<0,01	<0,01	0,02	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01
SHOW-INTEREST	<0,01	<0,01	<0,01	<0,01	**0,91**	<0,01	<0,01	0,00	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01
SURPRISE	**0,94**	<0,01	0,05	<0,01	<0,01	**0,54**	<0,01	<0,01	**0,19**	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01
THANKS	<0,01	**0,42**	**0,40**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01
WAIT	<0,01	<0,01	<0,01	**0,46**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01
NEUTRAL	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**

Tabulka F.16: Porovnání p-hodnot pro dobu trvání všech fonémů dosažené testem ANOVA mezi jednotlivými dvojicemi dialogových aktů. Tučně jsou označené hodnoty větší než 0,05, které znamenají statistickou nevýznamnost rozdílu středních hodnot.

dialogový akt	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	NEUTRAL
APOLOGY	**1,00**	<0,01	<0,01	**0,33**	0,04	<0,01	<0,01	<0,01	<0,01	<0,01	**0,88**	**0,34**	<0,01	**0,22**	<0,01	<0,01
CONFIRM	<0,01	**1,00**	<0,01	<0,01	0,00	<0,01	<0,01	<0,01	0,00	0,00	<0,01	0,00	<0,01	<0,01	<0,01	0,00
DIRECTIVE	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01
DISCONFIRM	**0,33**	<0,01	<0,01	**1,00**	**0,64**	<0,01	<0,01	**0,98**	<0,01	<0,01	**0,34**	**0,34**	**0,64**	0,04	<0,01	0,03
ENCOURAGE	0,04	0,00	<0,01	**0,64**	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	0,01	<0,01	<0,01	<0,01	<0,01	<0,01
GOODBYE	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**0,14**	<0,01
GREETING	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01
HAPPY-EMPATHY	<0,01	<0,01	<0,01	**0,98**	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	**0,08**	<0,01	<0,01	<0,01
NOT-SPECIFIED	<0,01	0,00	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01
OTHER	**0,57**	<0,01	**0,61**	**0,37**	**0,37**	0,04	**0,13**	**0,29**	**0,90**	**0,14**	**0,54**	**0,41**	**0,19**	**0,59**	**0,15**	**0,71**
REQUEST	<0,01	0,00	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01
SAD-EMPATHY	**0,88**	<0,01	<0,01	**0,34**	0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	**0,33**	<0,01	**0,14**	<0,01	<0,01
SHOW-INTEREST	**0,34**	0,00	<0,01	**0,34**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**0,33**	**1,00**	<0,01	0,01	<0,01	<0,01
SURPRISE	<0,01	<0,01	<0,01	**0,64**	<0,01	<0,01	<0,01	**0,08**	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01
THANKS	**0,22**	<0,01	<0,01	0,04	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**0,14**	0,01	<0,01	**1,00**	<0,01	**0,54**
WAIT	<0,01	<0,01	<0,01	<0,01	<0,01	**0,14**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01
NEUTRAL	<0,01	0,00	<0,01	0,03	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**0,54**	<0,01	**1,00**

Tabulka F.17: Porovnání p-hodnot pro RMS všech fonémů dosažené testem ANOVA mezi jednotlivými dvojicemi dialogových aktů. Tučně jsou označené hodnoty větší než 0,05, které znamenají statistickou nevýznamnost rozdílu středních hodnot.

dialogový akt	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	NEUTRAL
APOLOGY	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**0,11**	<0,01	<0,01	**0,55**	**0,12**	**0,90**	**0,22**	<0,01	0,00
CONFIRM	<0,01	**1,00**	<0,01	<0,01	0,00	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	0,00	<0,01	<0,01	<0,01	<0,01
DIRECTIVE	<0,01	<0,01	**1,00**	**0,15**	<0,01	<0,01	<0,01	<0,01	**0,59**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	0,00
DISCONFIRM	<0,01	<0,01	**0,15**	**1,00**	**0,13**	**0,33**	<0,01	**0,10**	**0,13**	<0,01	0,02	**0,08**	0,03	**0,10**	<0,01	<0,01
ENCOURAGE	<0,01	0,00	<0,01	**0,13**	**1,00**	**0,09**	<0,01	0,02	<0,01	<0,01	<0,01	<0,01	<0,01	**0,24**	<0,01	0,00
GOODBYE	<0,01	<0,01	<0,01	**0,33**	**0,09**	**1,00**	<0,01	0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**0,06**	<0,01	0,00
GREETING	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01
HAPPY-EMPATHY	**0,11**	<0,01	<0,01	**0,10**	0,02	0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	**0,69**	**0,07**	**0,91**	<0,01	0,00
NOT-SPECIFIED	<0,01	<0,01	**0,59**	**0,13**	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	0,00
OTHER	<0,01	**0,66**	**0,32**	**0,07**	0,02	0,03	**0,76**	0,02	**0,23**	<0,01	0,01	0,02	0,01	0,02	**0,91**	**0,05**
REQUEST	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	0,01	<0,01	<0,01	<0,01	<0,01	0,00
SAD-EMPATHY	**0,55**	<0,01	<0,01	0,02	<0,01	<0,01	<0,01	<0,01	<0,01	0,01	**1,00**	<0,01	**0,42**	**0,07**	<0,01	0,00
SHOW-INTEREST	**0,12**	0,00	<0,01	**0,08**	<0,01	<0,01	<0,01	**0,69**	<0,01	<0,01	<0,01	**1,00**	**0,07**	**0,97**	<0,01	0,00
SURPRISE	**0,90**	<0,01	<0,01	0,03	<0,01	<0,01	<0,01	**0,07**	<0,01	<0,01	**0,42**	**0,07**	**1,00**	**0,25**	<0,01	0,00
THANKS	**0,22**	<0,01	<0,01	**0,10**	**0,24**	**0,06**	<0,01	**0,91**	<0,01	<0,01	**0,07**	**0,97**	**0,25**	**1,00**	<0,01	<0,01
WAIT	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	<0,01	**1,00**	<0,01
NEUTRAL	0,00	<0,01	0,00	<0,01	0,00	0,00	<0,01	0,00	0,00	0,00	0,00	0,00	0,00	<0,01	<0,01	**1,00**

Tabulka F.18: Prvotní akustická penalizační matice $\mathbf{A}'$.

	APOLOGY	CONFIRM	DIRECTIVE	DISCONFIRM	ENCOURAGE	GOODBYE	GREETING	HAPPY-EMPATHY	NOT-SPECIFIED	OTHER	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	THANKS	WAIT	**str. hod.**
APOLOGY	0,00	1,97	1,43	1,38	0,80	1,05	1,40	1,05	1,13	1,91	0,85	1,20	1,21	1,08	0,77	1,33	**1,16**
CONFIRM	1,97	0,00	1,16	0,94	1,56	1,85	1,62	1,43	1,68	2,36	1,73	1,27	1,31	1,43	1,84	1,69	**1,49**
DIRECTIVE	1,43	1,16	0,00	0,87	0,81	1,04	1,01	0,77	0,69	2,01	0,88	0,85	0,46	0,57	1,61	1,38	**0,97**
DISCONFIRM	1,38	0,94	0,87	0,00	1,05	1,15	1,06	0,81	1,09	2,10	1,19	0,75	0,93	0,96	1,42	1,30	**1,06**
ENCOURAGE	0,80	1,56	0,81	1,05	0,00	0,75	1,11	0,64	0,63	1,83	0,44	0,85	0,58	0,41	1,07	1,37	**0,87**
GOODBYE	1,05	1,85	1,04	1,15	0,75	0,00	1,39	0,54	0,51	2,25	0,66	0,76	0,96	0,70	1,41	1,70	**1,05**
GREETING	1,40	1,62	1,01	1,06	1,11	1,39	0,00	1,28	1,09	2,05	1,32	1,38	1,13	1,17	1,72	1,22	**1,25**
HAPPY-EMPATHY	1,05	1,43	0,77	0,81	0,64	0,54	1,28	0,00	0,60	1,99	0,60	0,41	0,70	0,47	1,19	1,44	**0,87**
NOT-SPECIFIED	1,13	1,68	0,69	1,09	0,63	0,51	1,09	0,60	0,00	2,10	0,58	0,84	0,68	0,51	1,51	1,53	**0,95**
OTHER	1,91	2,36	2,01	2,10	1,83	2,25	2,05	1,99	2,10	0,00	1,80	2,23	1,75	1,93	1,63	1,31	**1,83**
REQUEST	0,85	1,73	0,88	1,19	0,44	0,66	1,32	0,60	0,58	1,80	0,00	0,84	0,58	0,43	1,18	1,42	**0,91**
SAD-EMPATHY	1,20	1,27	0,85	0,75	0,85	0,76	1,38	0,41	0,84	2,23	0,84	0,00	0,87	0,67	1,33	1,60	**0,99**
SHOW-INTEREST	1,21	1,31	0,46	0,93	0,58	0,96	1,13	0,70	0,68	1,75	0,58	0,87	0,00	0,39	1,38	1,29	**0,89**
SURPRISE	1,08	1,43	0,57	0,96	0,41	0,70	1,17	0,47	0,51	1,93	0,43	0,67	0,39	0,00	1,31	1,46	**0,84**
THANKS	0,77	1,84	1,61	1,42	1,07	1,41	1,72	1,19	1,51	1,63	1,18	1,33	1,38	1,31	0,00	1,26	**1,29**
WAIT	1,33	1,69	1,38	1,30	1,37	1,70	1,22	1,44	1,53	1,31	1,42	1,60	1,29	1,46	1,26	0,00	**1,33**
NEUTRAL	1,85	1,71	0,97	1,29	1,29	1,44	0,99	1,36	1,10	1,99	1,41	1,59	1,06	1,22	2,03	1,49	**1,42**

Tabulka F.19: Relativní počet vybraných jednotek s příznakem jednotlivých dialogových aktů ve výsledných syntetických větách. Tučně jsou zvýrazněny ty hodnoty, které reprezentují relativní počet jednotek s takovým dialogovým aktem, který byl při syntéze požadován.

požadovaný DA \ vybraný DA	CONFIRM	DIRECTIVE	ENCOURAGE	HAPPY-EMPATHY	REQUEST	SAD-EMPATHY	SHOW-INTEREST	SURPRISE	NOT-SPECIFIED	NEUTRAL
CONFIRM	**74,5 %**	0,0 %	3,6 %	0,0 %	0,0 %	3,6 %	5,1 %	0,0 %	0,0 %	14,3 %
ENCOURAGE	0,0 %	2,5 %	**67,7 %**	0,0 %	1,2 %	0,0 %	15,5 %	0,0 %	1,9 %	11,2 %
HAPPY-EMPATHY	0,6 %	0,0 %	30,3 %	**35,2 %**	0,0 %	0,0 %	3,9 %	1,9 %	1,2 %	27,8 %
SAD-EMPATHY	0,0 %	0,0 %	13,9 %	3,2 %	0,0 %	**32,9 %**	25,3 %	0,0 %	5,6 %	19,6 %
SHOW-INTEREST	0,0 %	0,0 %	19,9 %	6,9 %	0,0 %	0,0 %	**39,7 %**	0,0 %	17,1 %	16,4 %
NOT-SPECIFIED	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	**3,5 %**	96,5 %
NEUTRAL	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	0,0 %	**100,0 %**

Příloha G

Ukázky

Na přiloženém CD lze nalézt některé ukázky syntézy expresivní řeči, které byly použity v poslechových testech při vyhodnocování výsledků této práce. Ukázky jsou uloženy ve formě zvukových souborů ve formátu MP3.

G.1 Syntéza expresivní řeči metodou výběru jednotek

Ukázky pro syntézu expresivní řeči metodou výběru jednotek se nachází v adresáři `synteza_USEL`. Název každého souboru obsahuje označení dialogového aktu, pro který byla daná ukázka syntetizována. Na CD se nachází 4 zvukové soubory pro každý z následujících dialogových aktů: *CONFIRM*, *ENCOURAGE*, *HAPPY-EMPATHY*, *SAD-EMPATHY*, *SHOW-INTEREST*, *NOT-SPECIFIED* a *NEUTRAL*[1].

G.2 Syntéza expresivní řeči metodou HMM

Ukázky pro syntézu expresivní řeči metodou HMM se nachází v adresáři `synteza_HMM`. Název každého souboru obsahuje označení dialogového aktu, pro který byla daná ukázka syntetizována. Na CD se nachází 3 zvukové soubory pro každý z následujících dialogových aktů: *CONFIRM*, *ENCOURAGE*, *HAPPY-EMPATHY*, *SAD-EMPATHY*, *SHOW-INTEREST* a *NOT-SPECIFIED*.

[1]Dialogový akt *NEUTRAL* v podstatě představuje původní syntézu neutrální řeči, neboť z tabulky F.19 plyne, že při syntéze řeči pro tento dialogový akt byly použity výhradně jednotky označené dialogovým aktem *NEUTRAL*.

G.3 Syntéza expresivní řeči metodou výběru jednotek v dialogu

Ukázky použití syntézy expresivní řeči v dialogu mezi seniory a počítačem se nachází v adresáři `synteza_v_dialogu`. Ten obsahuje 6 krátkých úryvků z dialogů (tzv. minidialogů) ve dvou verzích. Minidialogy jsou sestavené částečně z reálné promluvy seniorů a částečně ze syntetizovaných promluv avatara tak, aby výsledný minidialog byl souvislý a smysluplný. V jedné verzi minidialogů je použita původní syntéza neutrální řeči, ve druhé pak syntéza expresivní řeči metodou výběru jednotek tak, jak byla představena v této práci.

Resumé (česky)

Tato disertační práce se zabývá syntézou expresivní řeči v dialogu, kdy pro popis expresivity jsou použity dialogové akty – diskrétní expresivní kategorie. Cílem práce je vytvořit postup pro vývoj syntézy expresivní řeči metodou dynamického výběru jednotek pro dialogový systém v oblasti rozhovorů mezi člověkem (seniorem) a počítačem na dané téma osobních fotografií ze života. Tohoto cíle je dosaženo modifikací současných algoritmů používaných pro syntézu neutrální řeči. Základem je vytvoření expresivního řečového korpusu anotovaného pomocí nadefinovaných dialogových aktů. Zkoumání anotací tohoto korpusu a analýza expresivních řečových dat z hlediska různých akustických parametrů potom poskytují informace, jak od sebe odlišit řečové jednotky označené různými dialogovými akty. Toho je následně využito právě při výběru řečových jednotek z inventáře v průběhu syntézy řeči.

Ačkoliv je práce zaměřena na konkrétní oblast dialogového systému, klade si za cíl popsat postup vývoje syntézy expresivní řeči pro dialog obecněji. Popsaný postup by tak mohl být využit i v podobných systémech zaměřených na jiná témata, kde by byly definovány jiné expresivní kategorie, případně by byl použit i jiný popis expresivity. V takovém případě by však zřejmě musel být postup přizpůsoben zvolenému popisu.

Vyhodnocení dosažených výsledků je pak realizováno prostřednictvím poslechových testů, kdy posluchači hodnotí dva základní aspekty syntetické expresivní řeči: kvalitu a schopnost vyjádřit expresivitu. Vyhodnocení je provedeno jak pro izolované promluvy, tak v rámci dialogu. Z výsledků vyplývá, že syntetická expresivní řeč je hodnocena kladně, přestože její kvalita je ve srovnání se syntézou neutrální řeči o něco horší. Dokáže však na posluchače přenášet expresivitu a zvýšit tak přirozenost syntetické řeči, což bylo jedním z hlavních cílů této práce.

Resumé (anglicky)

This dissertation deals with expressive speech synthesis in a dialogue. Dialogue acts – discrete expressive categories – are used for expressivity description. The aim of this work is to create a procedure for development of expressive speech synthesis using unit selection method for a dialogue system in a limited domain. The domain is limited to dialogues between a human and a computer on a given topic of reminiscing about personal photographs. The main goal of this work is achieved by modification of current algorithms that are used for neutral speech synthesis. The basic task when solving this issue is to create an expressive speech corpus and its annotation using predefined set of dialogue acts. On the basis of both annotations and acoustic analysis of the speech data in terms of various acoustic parameters, we can enumerate differences between various dialogue acts. These numerical differences are then used in the process of selecting speech units from a unit inventory during speech synthesis.

Although this work is focused on a specific dialogue system with limited domain, the goal is also to describe the procedure of development of expressive speech synthesis in general. Thus, the described procedure could be also used in similar systems that are focused on other topics with differently defined expressive categories or with different expressivity description. In this case, the procedure would need to be adapted to such a description.

An evaluation of achieved results is performed using listening tests. The listeners asses two basic aspects of synthetic expressive speech: speech quality and expressivity perception. The evaluation is performed for isolated utterances as well as for utterances in a dialogue. We can conclude that the synthetic expressive speech is rated positively even though it is of worse quality when comparing with neutral speech synthesis. However, synthetic expressive speech is able to transmit expressivity on listeners and to improve the naturalness of the synthetic speech.

Literatura

[1] ASME PTC 19.1-2005: Test uncertainty. 2006, [standard].

[2] The MBROLA Project. 2012, [online; citováno 2012-11-05].
URL `http://tcts.fpms.ac.be/synthesis/`

[3] Alexandersson, J.; Buschbeck-Wolf, B.; Fujinami, T.; aj.: Dialogue Acts in VERBMOBIL-2 - Second Edition. Technická zpráva, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany, 1998.

[4] Allen, J.; Core, M.: Draft of DAMSL: Dialog Act Markup in Several Layers. WWW page, 1997, [online; citováno 2012-08-15].
URL `http://www.cs.rochester.edu/research/cisd/resources/damsl/RevisedManual/RevisedManual.html`

[5] Altun, H.; Yalcinoz, T.; Curtis, K. M.: Accurate Parameter Estimation for an Articulatory Speech Synthesizer with an Improved Neural Network Mapping. *Turkish Journal of Electrical Engineering & Computer Sciences*, ročník 9, č. 2, 2001: s. 147–160.

[6] Baer, T.; Gore, J. C.; Gracco, L. C.; aj.: Analysis of Vocal Tract Shape and Dimensions Using Magnetic Resonance Imaging: Vowels. *The Journal of the Acoustical Society of America*, ročník 90, č. 2, 1991: s. 799–828, ISSN 0001-4966.

[7] Batliner, A.; Fischer, K.; Huber, R.; aj.: Desperately Seeking Emotions or: Actors, Wizards, and Human Beings. In *ISCA Workshop on Speech and Emotion*, Newcastle, UK, 2000, s. 195–200.

[8] Black, A. W.; Zen, H.; Tokuda, K.: Statistical Parametric Speech Synthesis. In *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP*, ročník 4, Honolulu, HI, USA, 15–20 April 2007, ISBN 1-4244-0728-1, ISSN 1520-6149, s. 1229–1232.

[9] Boersma, P.; Weenink, D.: Praat: doing phonetics by computer. 2009, [software,verze 5.1.05, citováno 2012-10-29].
URL `http://www.praat.org/`

[10] Bulut, M.; Narayanan, S. S.; Syrdal, A. K.: Expressive Speech Synthesis Using a Concatenative Synthesiser. In *Proceedings of the 7th International Conference on Spoken Language Processing – ICSLP*, Denver, CO, USA, 2002, s. 1265–1268.

[11] Burkhardt, F.: Emofilt: the Simulation of Emotional Speech by Prosody-Transformation. In *Proceedings of Interspeech*, Lisbon, Portugal, 2005, s. 509–512.

[12] Burkhardt, F.; Paeschke, A.; Rolfes, M.; aj.: A Database of German Emotional Speech. In *Proceedings of Interspeech*, Lisbon, Portugal, 2005, s. 1517–1520.

[13] Burkhardt, F.; Sendlmeier, W. F.: Verification of Acoustical Correlates of Emotional Speech Using Formant-Synthesis. In *ISCA Workshop on Speech and Emotion*, Newcastle, UK, 2000, s. 151–156.

[14] Cahn, J. E.: The Generation of Affect in Synthesized Speech. *Journal of the American Voice I/O Society*, ročník 8, 1990: s. 1–19.

[15] Campbell, N.: Databases of Emotional Speech. In *ISCA Workshop on Speech and Emotion*, Newcastle, UK, 2000, s. 34–38.

[16] Campbell, N.: The Recording of Emotional Speech; JST/CREST Database Research. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation – LREC*, ročník 6, Las Palmas, Spain, 2002, s. 2029–2032.

[17] Campbell, N.: Towards Synthesising Expressive Speech; Designing and Collecting Expressive Speech Data. In *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, s. 1637–1640.

[18] Campbell, N.; Mokhiari, P.: Using a non-spontaneous speech synthesizer as a driver for a spontaneous speech synthesizer. In *ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition*, Tokyo, Japan, 2003, s. 239–242.

[19] Carletta, J.: Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics*, ročník 22, č. 2, 1996: s. 249–254, ISSN 0891-2017.

[20] Carlson, R.; Sigvardson, T.; Sjolander, A.: Data–Driven Formant Synthesis. Technická Zpráva 44, KTH, Stockholm, Sweden, 2002.

[21] Cauldwell, R.: Where did the anger go? The role of context in interpreting emotions in speech. In *ISCA Workshop on Speech and Emotion*, Newcastle, UK, 2000, s. 127–131.

[22] Cohen, J. A.: A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, ročník 20, č. 1, 1960: s. 37–46.

[23] Core, M. G.; Allen, J. F.: Coding Dialogs with the DAMSL Annotation Scheme. In *Working Notes of the AAAI Fall Symposium on Communicative Action in Humans and Machines*, Cambridge, MA, USA, November 1997, s. 28–35.

[24] Cornelius, R. R.: *The science of emotion: Research and tradition in the psychology of emotions*. NJ, USA: Prentice-Hall, Englewood Cliffs, 1996, 260 s.

[25] Cowie, R.: Describing the emotional states expressed in speech. In *ISCA Workshop on Speech and Emotion*, Newcastle, uk, 2000, s. 11–18.

[26] Cowie, R.; Douglas-Cowie, E.: Automatic statistical analysis of the signal and prosodic signs of emotion in speech. In *Proc. Conf. Fourth Int Spoken Language ICSLP 96*, ročník 3, 1996, s. 1989–1992.

[27] Dempster, A. P.; Laird, N. M.; Rubin, D. B.: Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, ročník 39, č. 1, 1977: s. 1–38, ISSN 0035-9246, with discussion.

[28] Douglas-Cowie, E.; Campbell, N.; Cowie, R.; aj.: Emotional Speech: Towards a New Generation of Databases. *Speech Communication*, ročník 40, č. 1-2, 2003: s. 33–60.

[29] Edgington, M.: Investigating the Limitations of Concatenative Synthesis. In *Proceedings of Eurospeech*, Rhodes/Athens, Greece, 1997, s. 593–596.

[30] Eide, E. M.; Aaron, A.; Bakis, R.; aj.: A Corpus-Based Approach to <AHEM/> Expressive Speech Synthesis. In *Proceedings of the 5th ISCA Speech Synthesis Workshop – SSW5*, Pittsburgh, PA, USA, 2004, s. 79–84.

[31] Eide, E. M.; Fernandez, R.: Database Mining for Flexible Concatenative Text-to-Speech. In *Proceedings of ICASSP*, ročník 4, 2007, s. 697–700.

[32] Engwall, O.: A 3D Vocal Tract Model for Articulatory and Visual Speech Synthesis. In *Fonetik–98*, editace P. Branderud; H. Traunmüller, Stockholm, Sweden: Stockholm University, 1998, s. 196–199.

[33] Erickson, D.: Expressive speech: Production, perception and application to speech synthesis. *Acoustical Science and Technology*, ročník 26, č. 4, July 2005: s. 317–325.

[34] F. Alias, X. L.: Evolutionary Weight Tuning for Unit Selection Based on Diphone Pairs. In *Proceedings of Eurospeech*, ročník 2, Geneve, Switzerland, 2003, s. 1333–1336.

[35] Fernandez, R.; Ramabhadran, B.: Automatic Exploration of Corpus-Specific Properties for Expressive Text-to-Speech: A Case Study in Emphasis. In *Proceedings of the 5th ISCA Speech Synthesis Workshop – SSW6*, Bonn, Germany, 2007, s. 34–39.

[36] Fleiss, J. L.: Measuring nominal scale agreement among many raters. *Psychological Bulletin*, ročník 76, č. 5, 1971: s. 378–382, doi:10.1037/h0031619.

[37] Frolov, M. V.; Milovanova, G. B.; Lazarev, N. V.; aj.: Speech as an indicator of the mental status of operators and depressed patients. *Human Physiology*, ročník 25, č. 1, 1999: s. 42–47.

[38] Gobl, C.; Chasaide, A. N.: The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, ročník 40, č. 1–2, 2003: s. 189–212, ISSN 0167-6393, doi:10.1016/S0167-6393(02)00082-1.

[39] Grill, P.; Tučková, J.: FORANA. In *Proceedings of International Conference on Technical Computing Prague*, Prague, Czech Republic, 2009, ISBN 978-80-7080-733-0, s. 32–39.

[40] Grimm, M.; Kroschel, K.; Narayanan, S.: The Vera am Mittag German audio-visual emotional speech database. In *IEEE International Conference on Multimedia and Expo*, April 26 2008, s. 865–868, doi:10.1109/ICME.2008.4607572.

[41] Grůber, M.: Syntéza expresivní řeči. Odborná práce ke státní doktorské zkoušce. Západočeská univerzita v Plzni, 2008.

[42] Grůber, M.: Enumerating Differences Between Various Communicative Functions for Purposes of Czech Expressive Speech Synthesis in Limited Domain. In *Proceedings of Interspeech*, Portland, Oregon, USA, 2012.

[43] Grůber, M.; Hanzlíček, Z.: Czech Expressive Speech Synthesis in Limited Domain: Comparison of Unit Selection and HMM-Based Approaches. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 7499, Berlin-Heidelberg, Germany: Springer, 2012, s. 656–664, doi:10.1007/978-3-642-15760-8_36.

[44] Grůber, M.; Legát, M.: Single Speaker Acoustic Analysis of Czech Speech for Purposes of Emotional Speech Synthesis. In *Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, ročník 2, Aberdeen, UK: The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2008, ISBN 1-902956-61-3, s. 84–87.

[45] Grůber, M.; Tihelka, D.; Matoušek, J.: Evaluation of Various Unit Types in the Unit Selection Approach for the Czech Language Using the Festival System. In *Proceedings of the 5th ISCA Speech Synthesis Workshop – SSW6*, Bonn, Germany: Rheinische Friedrich-Wilhelms-Universität, 2007, s. 276–281.

[46] Hajič, J.; Böhmová, A.; Hajičová, E.; aj.: *Treebanks: Building and Using Parsed Corpora*, kapitola The Prague Dependency Treebank: A Three-Level Annotation Scenario. Netherlands, Amsterdam: Kluwer, 2000, s. 103–127.

[47] Hamza, W.; Bakis, R.; Eide, E. M.; aj.: The IBM Expressive Speech Synthesis System. In *Proceedings of the 8th International Conference on Spoken Language Processing – ISCLP*, Jeju, Korea, 2004, s. 2577–2580.

[48] Hanzlíček, Z.: Czech HMM-Based Speech Synthesis. In *Text, Speech and Dialogue, proceedings of the 13th International Conference TSD*, *Lecture Notes in Computer Science*, ročník 6231, Berlin-Heidelberg, Germany: Springer, 2010, ISBN 978-3-642-15759-2, s. 291–298.

[49] Hanzlíček, Z.: Czech HMM-based Speech Synthesis: Experiments with Model Adaptation. In *Text, Speech and Dialogue, Proceedings of the 14th International Conference TSD 2011*, *Lecture Notes in Artificial*

Intelligence, ročník 6836, Berlin-Heidelberg, Germany: Springer, 2011, s. 107–114.

[50] Hanzlíček, Z.; Matoušek, J.: First Steps Towards New Czech Voice Conversion System. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 4188, Berlin-Heidelberg, Germany: Springer, 2006, ISBN 978-3-540-39090-9, s. 383–390, doi:10.1007/11846406_48.

[51] Hanzlíček, Z.; Matoušek, J.: Voice Conversion Based on Probabilistic Parameter Transformation and Extended Inter-speaker Residual Prediction. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 4629, Berlin-Heidelberg, Germany: Springer, 2007, ISBN 978-3-540-74627-0, s. 480–487, 10.1007/978-3-540-74628-7_62.

[52] Helfrich, H.; Standke, R.; Scherer, K. R.: Vocal indicators of psychoactive drug effects. *Speech Communication*, ročník 3, č. 3, 1984: s. 245–252, ISSN 0167-6393, doi:10.1016/0167-6393(84)90019-0.

[53] Hirschová, M.: Řečový akt, řečové jednánía komunikačnífunkce výpovědi. *Slovo a slovesnost*, ročník 65, č. 3, August 2004: s. 163–174.

[54] Hofer, G.: *Emotional Speech Synthesis*. Diplomová práce, University of Edinburgh, UK, 2004.

[55] Hofer, G.; Richmond, K.; Clark, R.: Informed Blending of Databases for Emotional Speech. In *Proceedings of Interspeech*, Lisbon, Portugal: ISCA, 2005, s. 501–504.

[56] Hunt, A. J.; Black, A. W.: Unit selection in a concatenative speech synthesis system using a large speech database. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, ročník 1, 1996, ISSN 1520-6149, s. 373–376, doi:10.1109/ICASSP.1996.541110.

[57] Iida, A.; Campbell, N.; Higuchi, F.; aj.: A Corpus-based Speech Synthesis System with Emotion. *Speech Communication*, ročník 40, č. 1-2, 2003: s. 161–187.

[58] Iida, A.; Campbell, N.; Iga, S.; aj.: A speech synthesis system with emotion for assisting communication. In *ISCA Workshop on Speech and Emotion*, 2000, s. 167–172.

[59] Jekat, S.; Klein, A.; Maier, E.; aj.: Dialogue Acts in VERBMOBIL. Technická zpráva, German Research Center for Artificial Intelligence (DFKI), Saarbrücken, Germany, 1995.

[60] Johnson, W. L.; Narayanan, S. S.; Whitney, R.; aj.: Limited domain synthesis of expressive military speech for animated characters. In *Proceedings of the 7th International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002, s. 163–166.

[61] Jurafsky, D.; Shrilberg, L.; Biasca, D.: Switchboard-DAMSL Labeling Project Coder's Manual. Technická Zpráva 97–02, University of Colorado, Institute of Cognitive Science, Boulder, Colorado, USA, 1997.

[62] Kain, A.; Macon, M. W.: Spectral voice conversion for text-to-speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, ročník 1, 1998, s. 285–288.

[63] Kawanami, H.; Iwami, Y.; Toda, T.; aj.: GMM-based Voice Conversion Applied to Emotional Speech Synthesis. *IEEE Tranactions on Speech and Audio Processing*, ročník 7, 1999: s. 2401–2404.

[64] Kienast, M.; ; Sendlmeier, W. F.: Acoustical analysis of spectral and temporal changes in emotional speech. In *ISCA Workshop on Speech and Emotion*, Newcastle, UK, 2000, s. 92–97.

[65] Klein, M.: An Overview of the State of the Art of Coding Schemes for Dialogue Act Annotation. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 1692, Berlin-Heidelberg, Germany: Springer, 1999, ISBN 978-3-540-66494-9, s. 274–279, doi: 10.1007/3-540-48239-3_50.

[66] Krstulovic, S.; Hunecke, A.; Schroder, M.: An HMM-Based Speech Synthesis System Applied to German and its Adaptation to a Limited Set of Expressive Football Announcements. In *Proceedings of Interspeech*, Antwerp, Belgium, 2007, s. 1897–1900.

[67] Latacz, L.; Mattheyses, W.; Verhelst, W.: Joint Target and Join Cost Weight Training for Unit Selection Synthesis. In *Proceedings of Interspeech*, Florence, Italy: ISCA, August 2011, s. 321–324.

[68] Legát, M.; Matoušek, J.; Tihelka, D.: A Robust Multi-phase Pitchmark Detection Algorithm. In *Proceedings of Interspeech*, Antwerp, Belgium, 2007, s. 1641–1644.

[69] Legát, M.; Matoušek, J.; Tihelka, D.: On the detection of pitch marks using a robust multi-phase algorithm. *Speech Communication*, ročník 53, č. 4, 2011: s. 552–566, ISSN 0167-6393, doi:10.1016/j.specom.2011.01.008.

[70] Matoušek, J.; Romportl, J.: On Building Phonetically and Prosodically Rich Speech Corpus for Text-to-speech Synthesis. In *Proceedings of the second IASTED international conference on Computational intelligence*, San Francisco: ACTA Press, 2006, ISBN 0-88986-602-3, s. 442–447.

[71] Matoušek, J.; Romportl, J.; Tihelka, D.; aj.: Recent Improvements on ARTIC: Czech Text-to-Speech System. In *Proceedings of Interspeech, 8th International Conference on Spoken Language Processing – ICSLP*, ročník 3, Jeju, Korea: Sunjin Printing Co., 2004, s. 1933–1936.

[72] Matoušek, J.; Tihelka, D.; Psutka, J.: Automatic Segmentation for Czech Concatenative Speech Synthesis Using Statistical Approach with Boundary-specific Correction. In *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, s. 301–304.

[73] Matoušek, J.; Tihelka, D.; Psutka, J.: Experiments with automatic segmentation for Czech speech synthesis. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 2807, Berlin-Heidelberg, Germany: Springer, 2003, ISBN 3-540-20024-X, s. 287–294.

[74] Matoušek, J.; Tihelka, D.; Romportl, J.: Current State of Czech Text-to-Speech System ARTIC. In *Text, Speech and Dialogue, proceedings of the 9th International Conference TSD*, *Lecture Notes in Computer Science*, ročník 4188, Berlin-Heidelberg, Germany: Springer, 2006, ISBN 3-540-39090-1, ISSN 0302-9743, s. 439–446.

[75] McIntyre, G.; Gocke, R.: Researching Emotions in Speech. In *Proceedings of th 11th International Conference on Speech Science & Technology*, editace P. Warren; C. I. Watson, New Zealand: University of Auckland, 2006, s. 264–269.

[76] Mehrabian, A.: Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in Temperament. *Current Psychology*, ročník 14, 1996: s. 261–292, ISSN 1046-1310, 10.1007/BF02686918.

[77] Montero, J. M.; Gutiérrez-Ariola, J.; Palazuelos, S.; aj.: Emotional Speech Synthesis: From Speech Database to TTS. In *Proceedings of the 5th International Conference on Spoken Language Processing – ICSLP*, ročník 3, Sydney, Australia, 1998, s. 923–926.

[78] Narayanan, S. S.; Alwan, A.; Haker, K.: An Articulatory Study of Fricative Consonants Using Magnetic Resonance Imaging. *The Journal of the Acoustical Society of America*, ročník 98, č. 3, 1995: s. 1325–1347.

[79] Oudeyer, P.-Y.: The production and recognition of emotions in speech: features and algorithms. *International Journal of Human-Computer Studies*, ročník 59, č. 1—2, 2003: s. 157–183, ISSN 1071-5819, doi: 10.1016/S1071-5819(02)00141-6.

[80] Pantazis, Y.; Stylianou, Y.: *Progress in Nonlinear Speech Processing*, *Lecture Notes in Computer Science*, ročník 4391, kapitola On the Detection of Discontinuities in Concatenative Speech Synthesis. Berlin-Heidelberg, Germany: Springer, 2007, ISBN 978-3-540-71503-0, s. 89–100.

[81] Pechac, M.: Plutchikova teorie emocí. 2012, [Online; accessed 15-August-2012].
URL `http://ei.czechian.net/webs/emoce/plutchik.php`

[82] Pitrelli, J. F.; Bakis, R.; Eide, E. M.; aj.: The IBM expressive text-to-speech synthesis system for American English. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 14, č. 4, 2006: s. 1099–1108.

[83] Plutchik, R.; Kellerman, H.: *Theories of emotion*. Emotion, theory, research, and experience, New York, USA: Academic Press, 1980, ISBN 9780125587013.

[84] Polzin, T. S.; Waibel, A.: Emotion-Sensitive Human-Computer Interfaces. In *ISCA Workshop on Speech and Emotion*, Newcastle, UK, 2000, s. 201–206.

[85] Přibil, J.; Přibilová, A.: Statistical Analysis of Spectral Properties and Prosodic Parameters of Emotional Speech. *Measurement Science Review*, ročník 9, 2009: s. 95–104, doi:10.2478/v10048-009-0016-4.

[86] Přibil, J.; Přibilová, A.: Statistical Analysis of Complementary Spectral Features of Emotional Speech in Czech and Slovak. In *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, ročník 6836, editace I. Habernal; V. Matoušek, Berlin-Heidelberg, Germany: Springer, 2011, ISBN 978-3-642-23537-5, s. 299–306, doi:10.1007/978-3-642-23538-2_38.

[87] Přibil, J.; Přibilová, A.; Frollo, I.: Analysis of spectral properties of acoustic noise produced during magnetic resonance imaging. *Applied Acoustics*, ročník 73, č. 8, 2012: s. 687–697, ISSN 0003-682X, doi:10.1016/j.apacoust.2012.01.007.

[88] Přibilová, A.; Přibil, J.: Non-linear frequency scale mapping for voice conversion in text-to-speech system with cepstral description. *Speech Communication*, ročník 48, č. 12, 2006: s. 1691–1703, ISSN 0167-6393, doi:10.1016/j.specom.2006.08.001.

[89] Quin, L.; Ling, Z. H.; Wu, Y. J.; aj.: *Chinese Spoken Language Processing*, *Lecture Notes in Computer Science*, ročník 4274, kapitola HMM-Based Emotional Speech Synthesis Using Average Emotion Model. Berlin-Heidelberg, Germany: Springer, 2006, s. 233–240.

[90] R. Artstein, M. P.: Inter-Coder Agreement for Computational Linguistics. *Computational Li*, ročník 34, č. 4, 2008: s. 555–596, doi: 10.1162/coli.07-034-R2.

[91] Rank, E.; Pirker, H.: Generating Emotional Speech with a Concatenative Synthesizer. In *Proceedings of the 5th International Conference on Spoken Language Processing – ICSLP*, ročník 3, Sydney, Australia, 1998, s. 671–674.

[92] Reichl, J.; Všetička, M.: Multimediální Encyklopedie Fyziky. 2012, [Online; accessed 15-August-2012].
URL `http://fyzika.jreichl.com/main.article/view/210-weber-fechneruv-psychofyzikalni-zakon`

[93] Reithinger, N.; Maier, E.: Utilizing statistical dialogue act processing in VERBMOBIL. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, Stroudsburg, PA, USA: Association for Computational Linguistics, 1995, s. 116–121, doi: 10.3115/981658.981674.

[94] Rencher, A. C.: *Methods of Multivariate Analysis*. Wiley Series in Probability and Statistics, John Wiley & Sons, Inc., 2002, 727 s., doi: 10.1002/0471271357.

[95] Roach, P.; Stibbard, R.; Osborne, J.; aj.: Transcription of Prosodic and Paralinguistic Features of Emotional Speech. *Journal of the International Phonetic Association*, ročník 28, 1998: s. 83–94.

[96] Romportl, J.: Prosodic Phrases and Semantic Accents in Speech Corpus for Czech TTS Synthesis. In *Text, Speech and Dialogue, proceedings of the 11th International Conference TSD*, *Lecture Notes in Artificial Intelligence*, ročník 5246, Berlin–Heidelberg, Germany: Springer, 2008, ISBN 978-3-540-87390-7, ISSN 0302-9743, s. 493–500.

[97] Romportl, J.: *Zvyšování přirozenosti strojově vytvářené řeči v oblasti suprasegmentálních zvukových jevů*. Dizertační práce, Západočeská Univerzita v Plzni, 2008.

[98] Russell, J. A.: A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, ročník 39, 1980: s. 1161–1178.

[99] Scherer, K. R.: Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, ročník 40, č. 1-2, 2003: s. 227–256.

[100] Schröder, M.: *Speech and Emotion Research*. Dizertační práce, Universität des Saarlandes, Germany, 2003.

[101] Schroder, M.: Emotional Speech Synthesis: A Review. In *Proceedings of Eurospeech*, Aalborg, Denmark, 2001, s. 561–564.

[102] Sondhi, M.; Schroeter, J.: A hybrid time-frequency domain articulatory speech synthesizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, ročník 35, č. 7, 1987: s. 955–967, ISSN 0096-3518, doi: 10.1109/TASSP.1987.1165240.

[103] Sridhar, V. K. R.; Syrdal, A. K.; Conkie, A.; aj.: Enriching text-to-speech synthesis using automatic dialog acts tags. In *Proceedings of Interspeech*, Florence, Italy: ISCA, August 2011, s. 317–320.

[104] Stolcke, A.; Ries, K.; Coccaro, N.; aj.: Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistic*, ročník 26, č. 3, 2000: s. 339–373.

[105] Styger, T.; Keller, E.: *Fundamentals of Speech Synthesis and Speech Recognition: Basic Concepts, State of the Art, and Future Challenges*, kapitola Formant Synthesis. John Wiley and Sons Ltd., Chichester, UK, 1995, s. 109–128.

[106] Syrdal, A. K.; Conkie, A.; Kim, Y.-J.; aj.: Speech acts and dialog TTS. In *Proceedings of the 7th ISCA Speech Synthesis Workshop – SSW7*, Kyoto, Japan, 2010, s. 179–183.

[107] Syrdal, A. K.; Kim, Y.-J.: Dialog Speech Acts and Prosody: Considerations for TTS. In *Proceedings of Speech Prosody*, Campinas, Brazil, May 2008, s. 661–665.

[108] Tachibana, M.; Yamagishi, J.; Onishi, K.; aj.: HMM-Based Speech Synthesis with Various Speaking Styles Using Model Interpolation. In *Proceedings of Speech Prosody*, Nara, Japan, 2004, s. 413–416.

[109] Taylor, P.: *Text-to-Speech Syntesis*. Cambridge University Press, 2009, ISBN 978-0-521-89927-7, 626 s.

[110] Thompson, W. R.: On a Criterion for the Rejection of Observations and the Distribution of the Ratio of Deviation to Sample Standard Deviation. *The Annals of Mathematical Statistics*, ročník 6, č. 4, 1935: s. 214–219, ISSN 00034851.

[111] Tihelka, D.: *Metody on-line výběru jednotek pro konkatenační metodu syntézi řeči*. Odborná práce ke státní doktorské zkoušce, Západočeská univerzita v Plzni, Czech Republic, 2003.

[112] Toda, T.; Black, A. W.; Tokuda, K.: Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 15, č. 8, 2007: s. 2222–2235.

[113] Tokuda, K.; Masuko, T.; Miyazaki, N.; aj.: Multi-Space Probability Distribution HMM. *IEICE Transactions on Information and Systems*, ročník E85-D, č. 3, 2002: s. 455–464, ISSN 0916-8532.

[114] Tokuda, K.; Zen, H.; Black, A. W.: An HMM-Based Speech Synthesis System Applied to English. In *IEEE Speech Synthesis Workshop*, Santa Monica, CA, USA, 2002, ISBN 0-7803-7395-2, s. 227–230.

[115] Tokuda, K.; Zen, H.; Yamagishi, J.; aj.: The HMM-Based Speech Synthesis System (HTS). [online; citováno 2012-10-29].
URL `http://hts.ics.nitech.ac.jp/`

[116] Trujillo-Ortiz, A.; Hernandez-Walls, R.; Castro-Perez, A.; aj.: MOUTLIER1: Detection of Outlier in Multivariate Samples Test. A MATLAB file, 2006, [online; citováno 2012-10-29].
URL `http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=12252`

[117] Tsuzuki, R.; Zen, H.; Tokuda, K.; aj.: Constructing Emotional Speech Synthesizers with Limited Speech Database. In *Proceedings of Interspeech, Proceedings of the 8th International Conference on Spoken Language Processing – ICSLP*, Jeju, Korea: Sunjin Printing Co., 2004, s. 1185–1188.

[118] Turk, O.; Schroder, M.: Evaluation of Expressive Speech Synthesis With Voice Conversion and Copy Resynthesis Techniques. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 18, č. 5, 2010: s. 965–973.

[119] Vepa, J.; King, S.: *Text to Speech Synthesis: New Paradigms and Advances*, kapitola Join Cost for Unit Selection Speech Synthesis. NJ, USA: Prentice-Hall, Englewood Cliffs, 2004, s. 35–62.

[120] Železný., M.; Krňoul, Z.; Císař, P.; aj.: Design, Implementation and Evaluation of the Czech Realistic Audio-visual Speech Synthesis. *Signal Processing*, ročník 12, 2006: s. 3657–3673, ISSN 0165-1684.

[121] Whissell, C. M.: *Emotion: Theory, Research and Experience*, kapitola The Dictionary of Affect in Language. NY, USA: Academic Press, New York, 1989, s. 113–131.

[122] Whittaker, S.; Walker, M.; Moore, J.: Fish or Fowl: A Wizard of Oz Evaluation of Dialogue Strategies in the Restaurant Domain. In *Language Resources and Evaluation Conference*, Gran Canaria, Spain, 2002.

[123] Wikipedia: Weber–Fechner law — Wikipedia, The Free Encyclopedia. 2012, [online; citováno 2012-08-15].
URL `http://en.wikipedia.org/w/index.php?title=Weber-Fechner_law&oldid=501019055`

[124] Wilks, S. S.: Multivariate statistical outlier. *The Indian Journal of Statistics*, ročník 25, č. 4, 1963: s. 407–426.

[125] Williams, C. E.; Stevens, K.: On determining the emotional states of pilots during flights: An exploratory study. *Aerospace Medicine*, ročník 40, 1969: s. 1369–1372.

[126] Wu, C.-H.; Hsia, C.-C.; Liu, T.-H.; aj.: Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis. *IEEE Transactions on Audio, Speech, and Language Processing*, ročník 14, č. 4, 2006: s. 1109–1116.

[127] Yamagishi, J.; Onishi, K.; Masuko, T.; aj.: Modeling of Various Speaking Styles and Emotions for HMM-Based Speech Synthesis. In *Proceedings of Eurospeech*, Geneva, Switzerland, 2003, s. 2461–2464.

[128] Yamagishi, J.; Onishi, K.; Masuko, T.; aj.: Acoustic Modeling of Speaking Styles and Emotional Expressions in HMM-Based Speech Synthesis. *IEICE Transactions on Information and Systems*, ročník E88-D, č. 3, 2005: s. 502–509, ISSN 0916-8532, doi:10.1093/ietisy/e88-d.3.502.

[129] Yang, H.; Meng, H.; Cai, L.: Modeling the acoustic correlates of dialog act for expressive Chinese TTS synthesis. *IET Conference Publications*, ročník 2008, č. CP544, 2008: s. 49–53, doi:10.1049/cp:20080758. URL `http://link.aip.org/link/abstract/IEECPS/v2008/iCP544/p49/s1`

[130] Yoshimura, T.; Tokuda, K.; Masuko, T.; aj.: Simultaneous Modeling of Spectrum, Pitch and Duration in HMM-Based Speech Synthesis. In *Proceedings of Eurospeech*, Budapest, Hungary, 1999, s. 2347–2350.

[131] Young, S.; Evermann, G.; Gales, M.; aj.: *The HTK Book (for HTK Version 3.4)*. Cambridge, U.K.: Cambridge University, 2006.

[132] Zen, H.; Tokuda, K.; Black, A. W.: Statistical parametric speech synthesis. *Speech Communication*, ročník 51, 2009: s. 1039–1064.

[133] Zovato, E.; Pacchiotti, A.; Quazza, S.; aj.: Towards Emotional Speech Synthesis: A Rule Based Approach. In *Proceedings of the 5th ISCA Speech Synthesis Workshop – SSW5*, Pittsburgh, PA, USA, 2004, s. 219–220.

[134] Zovato, E.; Sandri, S.; Quazza, S.; aj.: Prosodic Analysis of a Multistyle Corpus in the Perspective of Emotional Speech Synthesis. In *Proceedings of Interspeech, 8th International Conference on Spoken Language Processing – ICSLP*, ročník 3, Jeju, Korea: Sunjin Printing Co., 2004, s. 1897–1900.

Seznam publikovaných prací

Publikace v angličtině

1. Grůber, M. : **Enumerating Differences Between Various Communicative Functions for Purposes of Czech Expressive Speech Synthesis in Limited Domain**. Proceedings of International Conference Interspeech, Portland, Oregon, USA, 2012.

2. Grůber, M., Hanzlíček Z. : **Czech Expressive Speech Synthesis in Limited Domain: Comparison of Unit Selection and HMM-based Approaches**. Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 7499, p. 656–664, Springer, Berlin-Heidelberg, Germany, 2012.

3. Grůber, M. : **Acoustic Analysis of Czech Expressive Recordings from a Single Speaker in Terms of Various Communicative Functions**. Proceedings of the 11thIEEE International Symposium on Signal Processing and Information Technology, p. 267–272, IEEE, 345 E 47TH ST, NEW YORK, NY 10017, USA, 2011.

4. Matoušek, J., Hanzlíček, Z., Campr, M., Krňoul, Z., Campr, P., Grůber, M. : **Web-Based System for Automatic Reading of Technical Documents for Vision Impaired Students**. Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 6836, p. 364–371, Springer, Berlin-Heidelberg, Germany, 2011.

5. Grůber, M., Legát, M., Ircing, P., Romportl, J., Psutka, J. : **Czech Senior COMPANION: Wizard of Oz Data Collection and Expressive Speech Corpus Recording and Annotation**. Human Language Technology. Challenges for Computer Science and Linguistics, Lecture Notes in Computer Science, vol. 6562, p. 280–290, editace Z. Vetulani, Springer, Berlin-Heidelberg, Germany, 2011.

6. Matoušek, J., Campr, M., Hanzlíček, Z., Grůber, M. : **Automatic Reading of Educational Texts for Vision Impaired Students**. Proceedings of the Conference Universal Learning Design, p. 169–180, Masaryk University, Brno, 2011.

7. Grůber, M., Tihelka, D. : **Expressive Speech Synthesis for Czech Limited Domain Dialogue System - Basic Experiments**. 2010 IEEE 10th International Conference on Signal Processing Proceedings, vol. 1, p. 561–564, Institute of Electrical and Electronics Engineers, Inc., Beijing, China, 2010.

8. Grůber, M., Matoušek, J. : **Listening-test-based annotation of communicative functions for expressive speech synthesis**. Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 6231, p. 283–290, Springer, Berlin-Heidelberg, Germany, 2010.

9. Matoušek, J., Tihelka, D., Grůber, M. : **On Building a New Slovak Voice for the Czech Unit-Selection TTS System ARTIC**. Speech Processing, vol. 2010, p. 140–146, Institute of Photonics and Electronics AS CR, Prague, 2010.

10. Grůber, M., Legát, M. : **Development of Expressive Speech Synthesis for Czech Limited Domain Dialogue System**. Speech Processing, vol. 2009, p. 100–106, Institute of Photonics and Electronics AS CR, Prague, Czech Republic, 2009.

11. Grůber, M., Legát, M., Ircing, P., Romportl, J., Psutka, J. : **Czech Senior COMPANION: Wizard of Oz Data Collection and Expressive Speech Corpus Recording**. Human Language Technologies as a Challenge for Computer Science and Linguistics, p. 266–269, Wydawnictvo Poznanskie Sp. z o.o., Poznan, Poland, 2009.

12. Legát, M., Grůber, M., Ircing, P. : **Wizard of Oz Data Collection for the Czech Senior Companion Dialogue System**. Fourth International Workshop on Human-Computer Conversation, p. 1–4, University of Sheffield, 2008.

13. Grůber, M., Legát, M. : **Single Speaker Acoustic Analysis of Czech Speech for Purposes of Emotional Speech Synthesis**. Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine, vol. 2, p. 84–87, The Society for the Study of Artificial Intelligence and Simulation of Behaviour, 2008.

14. Grůber, M., Legát, M., Tihelka, D. : **Corpus Recording and Checking on the Recorded Data**. The 1st Young Researchers Conference on Applied Sciences, p. 174–179, Západočeská univerzita, Plzeň, 2007.

15. Legát, M., Grůber, M., Matoušek, J. : **The Issue of Checking the Volume Consistence of Speech Corpus During Recording**. The 1st Young Researchers Conference on Applied Sciences, p. 206–211, Západočeská univerzita, Plzeň, 2007.

16. Grůber, M., Tihelka, D., Matoušek, J. : **Evaluation of various unit types in the unit selection approach for the Czech language using the Festival system**. Sixth ISCA Workshop on Speech Synthesis, p. 276–281, Rheinische Friedrich-Wilhelms-Universität, Bonn, 2007.

17. Grůber, M., Legát, M. : **Classification and regression in R**. Technical report. Department of Information Systems, University of Minho, Portugal, 2005.

Publikace v češtině

1. Grůber, M. : **Syntéza expresivní řeči. Odborná práce ke státní doktorské zkoušce**. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Plzeň, 2008.

2. Grůber, M. : **Použití různých typů jednotek v přístupu dynamického výběru jednotek**. Diplomová práce. Západočeská univerzita v Plzni, Fakulta aplikovaných věd, Plzeň, 2006.