Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra kybernetiky

BAKALÁŘSKÁ PRÁCE

PLZEŇ, 2015 JAN HRANIČKA

Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra kybernetiky

Bakalářská práce

Příprava dat pro hodnocení kvality syntetické řeči pomocí EEG

Autor práce:

Jan Hranička

Vedoucí práce:

Ing. Mgr. Jan Romportl, Ph. D.

Plzeň, 2015

Prohlášení

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

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

V Plzni dne .....

.....

vlastnoruční podpis

Abstrakt

Cílem této bakalářské práce je prozkoumat a navrhnout alternativní metodu hodnocení kvality syntetické řeči pomocí EEG. Součástí práce je tak i studium současných metod hodnocení řečové syntézy formou poslechových testů. Navrhovaná alternativní metoda EEG by se mohla stát prvním krokem k objektivnějšímu přístupu metod určování kvality syntézy řeči. Hlavním principem experimentální metody EEG je sledování neurofyziologického signálu člověka se snahou potvrdit hypotézu výskytu nestandardní odezvy mozkové aktivity v místech, kde se v syntetické řeči chyba nalézá. Praktickou částí práce je příprava dat k tomuto experimentu, následná realizace měření s dobrovolníky a analýza získaných signálů.

Klíčová slova: Konkatenační syntéza řeči, syntéza výběrem jednotek, hodnocení kvality, analýza chyb, EEG, evokované potenciály

Abstract

The main goal of this Bachelor's thesis is to explore and design an alternative method for speech synthesis evaluation using EEG. A part of this thesis is also studying current methods for evaluating speech synthesis using listening tests. Proposed alternative method might become the first step for a more objective approach of speech synthesis quality investigating methods. The main principle of the experimental EEG method is observation of human neurophysiological signal with an attempt to confirm our hypothesis of occurrence of non-standard brain activity response in such segments of synthetic speech where errors are occurring. The practical part of this thesis makes up the experimental data preparation followed by the realization of an experiment with participants and analysis of the obtained signals.

Keywords: Concatenative speech synthesis, unit selection, quality evaluation, error analysis, EEG, evoked potentials

Poděkování

Rád bych poděkoval vedoucímu mé práce, panu Ing. Mgr. Janu Romportlovi, Ph. D., za jeho vstřícný přístup, odborné vedení a cenné rady.

Zároveň bych rád poděkoval pracovníkům Národního ústavu duševního zdraví (NUDZ) za jejich spolupráci a propůjčení zařízení pro tuto studii, zejména MUDr. Martinovi Brunovskému, Ph. D. a MUDr. Anně Bravermanové za pomoc s realizací experimentu a s analýzou signálů naměřených na EEG.

Obsah

Kapitola 1

Úvod

Syntetická řeč, tedy proces umělého vytváření řeči, je v dnešní době oblastí už nejen pouze vědeckou a badatelskou, ale taktéž komerční a velmi rozšířenou. Vývoj počítačové syntézy řeči probíhá již řadu let, za kterou tato technologie učinila obrovský pokrok a vzniklo několik různých metod tvorby umělé řeči. V současné době existují tak kvalitní systémy produkující syntetickou řeč, že ji v mnohých situacích běžný člověk nedokáže rozeznat od řeči přirozené. Ovšem stále se v syntetické řeči objevují artefakty a chyby, které narušují přirozenost vygenerované promluvy a výrazně snižují kvalitu a příjemnost poslechu této řeči.

Právě kvalita syntetické řeči hraje u této technologie významnou roli. V důsledku s tím vzniklo několik metod hodnocení syntetické řeči, kde je snahou hledáním nedostatků zlepšit její kvalitu. Zřejmě nejrozšířenější formou prostředků pro hodnocení jsou tzv. poslechové testy, které ale nemusí být optimálním řešením kvůli své náročnosti pro hodnotícího dobrovolníka a rozsahu možných nastavení parametrů testovaného systému převodu textu na řeč. V posledních letech se objevují testy, jejichž principem je sledování neurofyziologického signálu člověka poslouchajícího syntetickou řeč a následnou analýzou získaných dat je hodnocena kvalita dané syntetické řeči. Návrhem a zkoumáním takové nové testovací metody se právě tato práce zabývá.

Důvodů, proč se zabývat hodnocením kvality syntetické řeči, je hned několik. V současné době jsou TTS systémy velmi důležitou technologií v každodenním životě mnoha lidí, ať už jde o řidiče automobilů, cestující či handicapované. V mnoha případech je zcela striktně vyžadována vysoká kvalita syntetické řeči, ať už z důvodu komfortu či bezpečnosti. Kvalitu syntetické řeči zajisté ocení automobilový průmysl, respektive výrobci GPS zařízení. Bylo by nevhodné a do jisté míry i nebezpečné, kdyby v GPS navigaci nebylo rekonstruované řeči rozumět a řidič by tak byl nucen navigaci věnovat bližší pozornost, a tím pádem ztratit pozornost od řízení. Dalším příkladem může být hlášení na vlako-

vých nádražích, kde špatně srozumitelná rekonstruovaná řeč vede ke zkreslení informací a dezorientaci cestujících. Syntetická řeč se ale objevuje i na mnoha dalších místech, jako například v chytrých telefonech, na internetu, v telekomunikačních službách,[1] kde je taktéž vyžadována vysoká úroveň srozumitelnosti. Dalším důležitým faktorem, kterému je věnována velká pozornost, je přirozenost řeči. Ta má velký vliv zejména na komfort jejího poslechu.

1.1 Cíl práce

Hlavním cílem této práce je navrhnout a experimentálně prozkoumat alternativní metodu hodnocení kvality syntetické řeči sledováním neurofyziologických signálů člověka pomocí EEG. Průběh a rozbor těchto signálů by mohl pomoci detekovat výskyt chyb v syntetické řeči a přispět ke zvyšování kvality TTS systémů. V rámci práce je nutné ověřit naší hypotézu přítomnosti neobvyklé odezvy mozkové aktivity v místech výskytu chyb při poslechu porušené resyntetizované promluvy.

Abychom mohli tuto hypotézu ověřit, je třeba připravit zvuková data ve formě přirozené řeči, do kterých budou manuálně vkládány specifické chyby, které bude nutné taktéž definovat. O těchto chybách pak budeme mít veškeré potřebné údaje pro analýzu naměřených aktivit na EEG. V poslední řadě bude třeba na základě výsledků experimentálního měření navrhnout pokračování výzkumu, případně změnu postupu v experimentu.

1.2 Obsah práce

Pro lepší orientaci je v tomto oddílu krátce popsána struktura práce a stručný obsah jednotlivých kapitol.

V kapitole 2, **Syntetická řeč a systémy TTS**, je stručně shrnuto, co je syntetická řeč. Jsou zde popsány základní metody její tvorby s důrazem na konkatenační metodu výběrem jednotek, která byla využita v praktické části úlohy.

V kapitole 3, **Analýza chyb v syntetické řeči s výběrem jednotek**, jsou podrobněji analyzovány možné typy chyb v konkatenační syntéze řeči. Je zde vysvětlována příčina

[1] Jako příklad využití syntézy v telekomunikačních službách si můžeme představit velmi rozšířený pojem **zákaznická linka**, kde může být také použita syntéza. V mnoha případech se nejedná jen o TTS systém, ale o hlasový dialogový systém pro případnou komunikaci s volajícím.

vzniku těchto chyb, míra jejich vlivu na přirozenost a srozumitelnost poslechu výsledné řeči.

Ve 4. kapitole, **Hodnocení kvality syntetické řeči**, jsou představeny základní metody pro určování kvality formou poslechových testů. Jsou zde rozebírány testy srozumitelnosti a přirozenosti, které kvalitu dané řeči testují na segmentální, respektive suprasegmentální úrovni. Ve druhé části této kapitoly je náhled a přiblížení experimentální metody EEG, jež se stala vodítkem pro návrh našeho experimentu. Nakonec je prezentován návrh experimentálního protokolu a přiblížení podstaty navrhovaného experimentu.

Kapitola 5, **Příprava experimentu**, detailněji popisuje praktickou část, tedy přípravu dat pro experimentální hodnocení kvality syntetické řeči pomocí EEG. V této kapitole se dočtete, jaká data byla vybrána pro experiment a jakým způsobem byla získávána, jakými prostředky byly do nahrávek zaváděny chyby, jejichž rozdělení je popsáno v kapitole 4, a do jaké formy bylo nutné nahrávky upravit, aby byly použitelné. Dále je zde také popsáno aplikační prostředí Presentation, které slouží k řízení neurobehaviorálních experimentů, jakým způsobem a proč je toto prostředí používáno pro naše experimenty a jak se v něm dají experimenty řídit.

V kapitole 6, **Průběh a vyhodnocení EEG experimentu**, se dočtete o průběhu experimentálního měření a o analýze získaných signálů z měření na EEG. Taktéž se dozvíte více z pozadí experimentu.

Na závěr, v kapitole 7, je tato práce shrnuta, rekapitulovány cíle a výsledky, kterých bylo dosaženo. Dále je vedena diskuse nad úspěšností experimentu a návrh na další pokračování či změny zkoumané problematiky.

Kapitola 2

Syntetická řeč a systémy TTS

Doba, kdy byly pojmy syntetická řeč a počítačová syntéza řeči tématem jen v okruhu výzkumných pracovníků, jsou již dávno pryč. Ačkoliv si to mnozí lidé neuvědomují, v dnešní době moderních technologií máme možnost na syntetickou řeč narazit v každodenním životě, například cestou vlakem do práce nebo ve vašem chytrém telefonu. Slouží tak především ke zlepšení a zpříjemnění komunikace mezi člověkem a strojem (počítačem). Jelikož je řeč nejčastější dorozumívací prostředek mezi lidmi, je přirozené, že se lidé snaží tuto vlastnost převést i na stroje. Syntetická řeč tak nevědomky vstoupila do našich životů. Toto odvětví moderních technologií má svou pozici nejen pro usnadnění práce milionům lidí, ale také té hrstce pacientů, kteří přišli o zrak či hlasivky a nejsou tak schopni číst nebo komunikovat s okolím jinak, než psaním na papír nebo znakovou řečí, které ale obyčejný člověk těžko porozumí. Ovšem dostupnost technologie syntetické řeči není vším. Velmi důležitým faktorem je právě kvalita řeči generované počítačem, neboť právě kvalita této umělé řeči ji více a více posouvá k hranici řeči přirozené.

Abychom se mohli bavit o kvalitě syntetické řeči a o metodách hodnocení její kvality, je více než žádoucí stručně si shrnout, co je přirozená a syntetická řeč, jaký je mezi nimi rozdíl a jakými prostředky a metodami se umělá řeč tvoří. Nebudeme se dopodrobna zabývat, jakými prostředky je třeba řeč analyzovat, abychom mohli syntetickou řeč vytvářet. O analýze řečového signálu, o rozboru jazyka a o detailnějším popisu metod tvorby umělé řeči se můžete více dočíst v [3], [7] a [8].

2.1 Přirozená řeč a její vlastnosti

Ještě než se začneme zabývat syntetickou řečí, někdy označované jako umělou, bude dobré získat nějaké znalosti o řeči mluvené, resp. přirozené. Lidská řeč je již od nepaměti nejběžnějším a nejrozšířenějším komunikačním prostředkem mezi lidmi. Na rozdíl od psaní a

čtení je mluvená forma jazyka přirozenější a méně náročná, neboť při psaní či čtení nemůže člověk dělat více věcí najednou (řídit, dívat se na televizi apod.). Vhodným příkladem je fakt, že mluvenou řeč jsou děti schopné se naučit již v raném dětství, zatímco číst a psát neumí ani v dospělosti stále mnoho lidí po celém světě, zejména v rozvojových zemích.

2.1.1 Tvorba mluvené řeči

Mluvená řeč je vytvářena hlasovým traktem, který se skládá z dechového, hlasového a artikulačního ústrojí. Každé z těchto ústrojí plní svou důležitou roli, přičemž výsledkem je řeč (resp. akustické vlnění). Jednotlivé části hlasového traktu plní specifickou funkci a jsou to:

- **Dechové ústrojí**, které funguje nejen jako zprostředkovatel nezbytného kyslíku pro tělo, ale také jako zdroj energie pro vytváření řeči. Při mluvení (popř. výdechu) proudí vzduch přes další části hlasového traktu, kde je přeměňován na akustické vlny a modifikován do své finální podoby. Síla výdechu a proudění vzduchu ovlivňuje nejen intenzitu, ale také výšku promluvy, proto kupříkladu při křiku spotřebujete mnohem více kyslíku, než při šepotu.

- **Hlasové ústrojí** představuje primární systém tvorby řeči. V hrtanu se nacházejí hlasivky, které představují pružný orgán, který se jako hlavní podílí na tvorbě řeči. Vzduch z dechového ústrojí proudí hlasivkami, které svým kmitáním vytvářejí zvukové vlny. Tento proud zvukových vln tvoří základní hlasivkový tón, respektive frekvenci základního hlasivkového tónu označovanou jako F_0.[1]

- **Artikulační ústrojí** je poslední částí hlasového traktu a stará se o modifikaci zvukových vln vytvářených hlasivkami. Do tohoto ústrojí patří jednak nadhrtanové dutiny (hrdelní, ústní, nosní), které ovlivňují signál pasivně, neboť jsou nepohyblivé, a artikulační orgány (jazyk, rty, měkké patro), které se pohybují, čímž jsou schopné vytvářet velké množství různých modifikací zvukového signálu. Samotný hrtan může svým pohybem měnit délku celého hlasového traktu.

[1] Frekvence F_0 se ve většině případů pohybuje v rozmezí 60-400Hz. Navíc je rozdílná u mužů, kteří mají hlubší hlas, a žen, které mají hlas vyšší. Nejvyšší frekvence pak bývá u malých dětí.

2.1.2 Poruchy mluvené řeči

Ačkoliv je hlasový trakt velice důmyslně propracovaný a složitý systém (důkazem je obtížnost matematicko-fyzikálního popisu, modelování a simulace tohoto systému, o který se snaží artikulační syntéza), i v případě přirozené řeči můžeme mluvit o situacích, kdy je řeč nějakým způsobem porušená nebo „nekvalitní“.

V mnoha případech je porucha řeči způsobena anomálií v artikulačním ústrojí, které se stará právě o modifikaci zvukového signálu. Příkladem může být špatné postavení zubů či následek obrny. Většina běžných poruch řeči se však dají odstranit na specializovaných pracovištích, ať už formou speciálních lekcí učení řeči nebo chirurgickým zákrokem. Existuje rovněž řada testovacích metod, které pomáhají odhalit skryté vady řeči a odstranit je již v dětství.

V horších případech může dojít až k úplné ztrátě řeči, neboli k němotě, která může být způsobena jednak zdravotními problémy,[2] ale také psychickými. Při ztrátě řeči hrají hlavní roli právě moderní technologie tvorby řeči umělé, které jsou dnes již poměrně dostupné. Pomocí nich mají postižení pacienti možnost opět komunikovat s okolním světem formou řeči. Právě kvalita a forma umělé řeči pak hraje nesmírně důležitou roli v životě těchto lidí. Je tedy nezbytné se kvalitou umělé řeči zabývat.

2.2 Syntetická řeč a její vlastnosti

Syntetická řeč a její tvorba je velmi rozsáhlá problematika, o které se můžete více dočíst v [3], odkud tato práce mnohé čerpá. Samotná syntéza řeči je jednou z inženýrských disciplín zpracování řečového signálu.

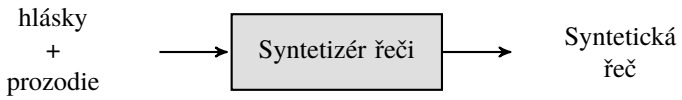
Hranice mezi syntetickou a přirozenou řečí ovšem není snadno definovatelná. Jak se píše v [7], řeč vytvořená strojem je umělá ve smyslu absence člověka jako přirozeného původce promluv lidské řeči. Na jednu stranu je umělost řeči vnímána lidským posluchačem tehdy, pokud zní řeč generovaná strojem nepřirozeně či „plechově“, na druhou stranu je-li tato promluva svou kvalitou stejná, jako promluva lidská, bude se považovat za přirozenou. Syntetická řeč tedy nemusí být nutně nositelem kritéria „umělosti“, stačí, když

[2] Při rakovině hrtanu může dojít až k chirurgickému odstranění hlasivek, neboli k tzv. totální laryngektomii. V tomto případě nebude pacient již nikdy schopen mluvit svým hlasovým ústrojím. To je příkladem situace, kdy je možné využít moderní technologie k produkci umělé řeči a možnost vrátit se částečně zpět do normálního života a komunikace.

posluchač nerozezná, jestli je původcem řeči člověk nebo stroj.

Je obecně velkou snahou vytvořit co nejkvalitnější a přirozeně znějící syntetickou řeč. Úsilí napodobit lidskou řeč vynakládají lidé pravděpodobně z důvodu zjednodušení komunikace mezi člověkem a strojem. Taková komunikace by v mnoha ohledech lidem usnadnila práci a zvýšila pracovní efektivitu.[3]

Zařízení, které provádí vytváření umělé řeči, se nazývá **syntetizér řeči** (viz obrázek 2.4). To na základě vstupní informace generuje syntetickou řeč.



Obrázek 2.1: Jednoduché blokové schéma typického syntetizéru k produkci syntetické řeči

2.3 Metody tvorby syntetické řeči

Tvorba syntetické řeči samozřejmě není vynálezem posledních deseti let. V [3] je uvedeno, že vůbec první pokusy o vytvoření umělé řeči se datují již k roku 1779. O tento významný vynález se zasloužil německý profesor Christian Kratzenstein. V té době se jednalo o mechanický syntetizér, který byl později nahrazen syntetizérem elektronickým. Zlom ve vývoji syntetické řeči nastal zhruba v polovině šedesátých let minulého století, kdy s příchodem číslicových počítačů začaly vznikat první digitální syntetizéry. O problematiku syntetické řeči se začalo více zajímat a vzniklo několik různých metod, jak umělou řeč vytvářet.

2.3.1 Přehled metod syntézy řeči

Z hlediska způsobu modelování použitého při tvorbě syntetické řeči lze metody tvorby rozdělit na tři základní typy:

- **Formantová syntéza** řeči je dnes již spíše historickým milníkem v tvorbě umělé řeči. Ačkoliv se jedná o historicky velice úspěšnou metodu, v dnešní době ji zcela nahradila metoda kontatenační. Metoda formantové syntézy je založena na simulaci

[3] Komunikace mezi člověkem a strojem však není pouze otázkou syntézy řeči, nýbrž je také snaha o interakci s člověkem, čímž se zabývají disciplíny jako je rozpoznávání řeči, hlasové dialogové systémy či strojové učení a rozpoznávání.

procesu, kterým vytváří řeč člověk - modeluje se zdroj buzení a hlasový trakt. Simulací vibrací hlasivek, formantů a generováním signálů se vytváří syntetická řeč. Formant označuje lokální maximum ve spektru složených tónů a je závislý na aktuálním rozpoložení a konfiguraci řečových orgánů a geometrických vlastnostech hlasového traktu. Významný formant v syntéze řeči představuje **frekvence základního hlasivkového tónu** označovaná jako frekvence F_0. Velice zjednodušeně můžeme časový průběh F_0 přirovnat k intonaci. Sledováním F_0 můžeme také snadněji odhalit a eliminovat chyby v syntetizované řeči (např. při nespojitosti průběhu F_0).

- **Artikulační syntéza** používá k produkci syntetické řeči fyzikální model vytváření řeči člověkem (tzv. artikulační model), který zahrnuje model hlasivek a jednotlivých artikulátorů, tedy orgánů k produkci řeči. Tento matematicky popsaný systém tvoří dynamiku jednotlivých artikulátorů a simuluje činnost celé řečové soustavy k produkci syntetické řeči. Ačkoliv se jedná o metodu velmi atraktivní, v současné době není matematický popis reálného řečového systému kvůli své komplexnosti na takové úrovni, aby produkoval přirozeně znějící řeč a kvalitou se vyrovnal konkatenační syntéze řeči.

- **Konkatenační syntéza** reprezentuje v současné době nejhojněji používanou metodu přístupu k tvorbě syntetické řeči. Hlavní myšlenka konkatenační syntézy je založena na předpokladu, že jednotlivé zvuky z dané řeči lze reprezentovat konečným počtem řečových jednotek. Ty jsou uloženy v inventáři řečových jednotek, který je obvykle realizován jako segmentovaný řečový korpus. Syntetická řeč pak vzniká konkatenací právě těchto řečových jednotek. Jelikož v této práci je navrhována metoda pro hodnocení syntetické řeči vzniklé touto metodou, je konkatenační syntéza podrobněji popsána v následujícím oddílu.

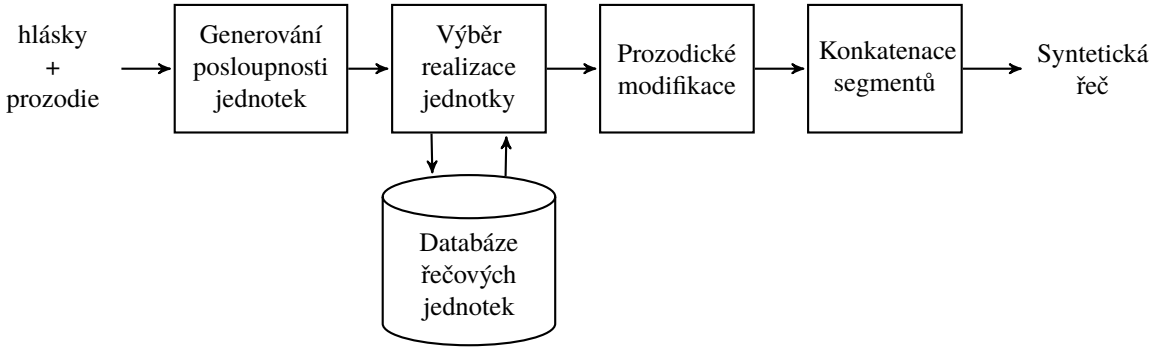
2.4 Konkatenační metoda syntézy řeči

Konkatenační syntéza (z anglického concatenative synthesis) vzniká konkatenací[4] řečových jednotek z připraveného inventáře řečových jednotek. Tyto jednotky ovšem reprezentují části přirozené řeči, což tuto metodu výrazně odlišuje od zmíněné formantové a

[4]Konkatenace znamená spojování segmentovaných řečových jednotek do výsledného řečového signálu. Konkatenace tedy znamená spojování či řetězení, což vystihuje a pojmenovává metodu konkatenační syntézy.

artikulační metody, ve kterých byl řečový systém zcela nahrazen modelem řečového systému. U konkatenační syntézy je zjednodušeně řečeno výsledná řeč generována z původních úseků mluvené, přirozené řeči, které jsou uloženy v řečovém korpusu.

Při řetězení řečových jednotek vybraných z inventáře může syntetizér tyto jednotky modifikovat, nebo je nechat beze změny. Syntetizéry modifikující řetězené řečové jednotky upravují jejich prozodické či spektrální charakteristiky, čímž se snaží minimalizovat nespojitosti na hranicích segmentů v generované řeči. V případě, že syntetizér řečové jednotky nemodifikuje a řetězí jednotky ve tvaru, ve kterém jsou uloženy v řečovém inventáři, je ve většině případů zapotřebí obrovský inventář řečových jednotek. Tato metoda je na rozdíl od metody s modifikací jednotek výpočetně výrazně náročnější, ale produkuje přirozeněji znějící řeč.



Obrázek 2.2: Blokové schéma tvorby řeči konkatenační syntézou

Základním stavebním kamenem konkatenační syntézy je řečový korpus a jeho tvorba, o něž se opírá veškeré následující zpracovávání signálu a samotná produkce řeči syntetizérem.

Řečové jednotky

Při návrhu syntetizéru a při tvorbě řečového korpusu je prvním důležitým rozhodnutím volba řečových jednotek, ze kterých pak bude vytvářena syntetická řeč. V současnosti existuje celá řada řečových jednotek, od vět, slov a slabik až po hlásky, difony, trifony a další, velmi malé jednotky řeči. V systému TTS, který byl použit pro tento výzkum, představují zvolené jednotky tzv. difony. Ty lze jednoduše definovat jako segment řeči začínající v polovině předchozí hlásky a končící v polovině hlásky druhé [3].

2.4.1 Řečový korpus a inventář řečových jednotek

Řečový korpus obsahuje akustické jednotky, které jsou nezbytné pro sestavení inventáře řečových jednotek. Jednotky obsažené v řečovém korpusu jsou také detailně popsány pro účely následného zpracování a korpus tak mimo samotné řečové signály obsahuje jejich fonetickou a ortografickou anotaci, glotální a parametrický popis, intenzitu, frekvenci základního hlasivkového tónu F_0, trvání řeči aj.

Inventář řečových jednotek vzniká segmentací řečového korpusu. V ojedinělých případech je tato databáze vytvářena manuálně, avšak ve většině případů, kdy se jedná o velmi rozsáhlý soubor dat z důvodu detailnějšího popisu řeči, je inventář vytvářen automaticky.

2.4.2 Korpusově orientovaná syntéza

Díky velmi rychle se rozvíjejícímu výpočetního výkonu počítačů se začala hojně používat tzv. **korpusově orientovaná syntéza** řeči. Ta k syntéze řeči využívá rozsáhlé (segmentované) řečové korpusy, které obsahují detailně popsané[5] řečové jednotky. Pro dosažení kvalitní a přirozeně znějící syntetické řeči je zapotřebí vytvořit kvalitní řečový korpus, neboť právě na základě zvoleného řečového korpusu závisí i kvalita výsledné generované řeči. Vhodně zvoleným algoritmem výběru řečových jednotek z připraveného inventáře pak probíhá vlastní syntéza řeči.

U korpusově orientované konkatenační syntézy řeči rozlišujeme následující dva přístupy její realizace:

Syntéza s pevným inventářem (anglicky Single Unit Instance, SUI) představuje syntézu, jejíž kvalita je limitována omezeným výběrem instancí řečových jednotek – každá řečová jednotka má vždy jen jednu instanci. SUI syntéza využívá většinou malé inventáře, což se projevuje i na kvalitě výstupní řeči. Značnou výhodou jsou ovšem její menší výpočetní nároky a objem zdrojových dat. Proto se tato syntéza nasazuje například do chytrých telefonů nebo jiných zařízení omezených úložnou kapacitou. Z důvodu omezeného množství řečových jednotek se u této metody využívá spektrální a prozodické modifikace řečových jednotek, aby na sebe řetězené jednotky lépe navazovaly. To ale může mít za následek zhoršení přirozenosti výsledné řeči.

[5] Při popisu je kladen důraz zejména na prozodické, fonetické a spektrální vlastnosti jednotlivých řečových jednotek, např. difonů.

Syntéza výběrem jednotek představuje syntézu, jejíž kvalita je závislá na výběru jedné z mnoha instancí řečových jednotek (anglicky Multiple Unit Instance, MUI) – zde je využit algoritmus výběru jednotek (anglicky Unit Selection). Syntéza výběrem jednotek je použita právě v této práci, a proto je podrobněji popsána v následujícím oddílu.

2.4.3 Syntéza výběrem jednotek

Syntéza řeči výběrem jednotek patří mezi nejvýznamnější představitele korpusově orientované konkatenační syntézy řeči [3, 9]. Výběr jednotek opět vychází z pečlivě segmentovaného řečového korpusu, kde jsou detailně anotované a prozodicky popsané jednotlivé řečové jednotky. Výsledná řeč je poté generována na základě požadovaných jednotek (zpravidla hlásek či difonů) doplněných o prozodické vlastnosti (F_0, hlasitost a trvání).

Syntéza výběrem jednotek využívá velké množství instancí každé řečové jednotky (v našem případě difonu). Algoritmus výběru jednotek se pak snaží nalézt nejvhodnější jednotku pro požadovaný kontext. Výběr takové jednotky je zajišťován minimalizací tzv. ceny cíle a ceny konkatenace.

cena cíle (target cost) značí, jak se daná instance řečové jednotky u_i liší od požadované jednotky t_i. Celková cena cíle je dána součtem subcen z požadovaných vlastností řečové jednotky (F_0, trvání, pozice ve slově atp.).

$$C^t(t_i, u_i) = \sum w_k^t C_k^t(t_i, u_i),$$

kde w_k^t je váhová funkce cíle.

cena konkatenace (concatenation cost) naopak značí, jak dobře se daná jednotka u_i řetězí se sousední jednotkou u_{i-1}. Celková cena konkatenace je složena ze subcen (lokální spojitost F_0 apod.).

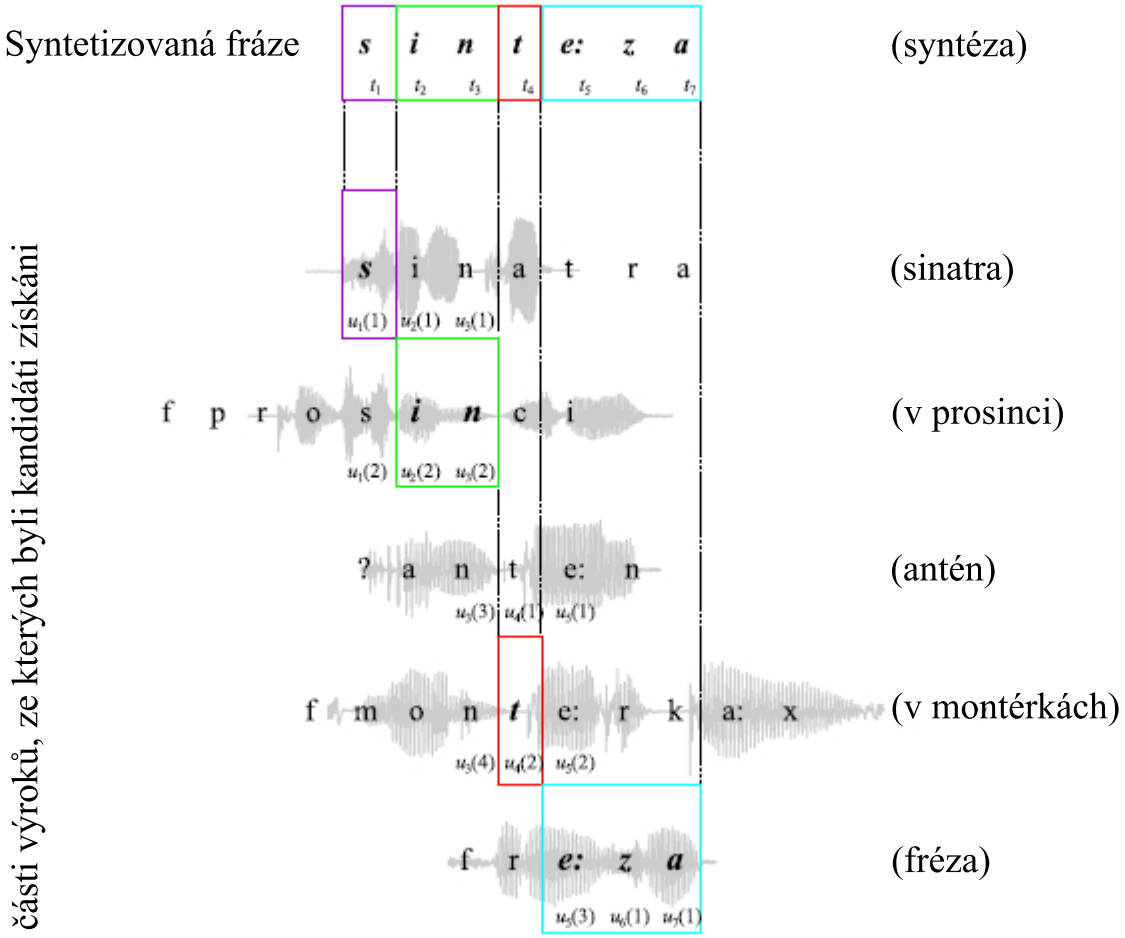
$$C^c(u_{i-1}, u_i) = \sum_k w_k^c C_k^c(u_{i-1}, u_i),$$

kde w_k^c je váhová funkce konkatenace.

Pro nalezení optimální posloupnosti jednotek je třeba minimalizovat celkovou cenu danou součtem cen cíle a cen konkatenace, tedy

$$C(t_1^N, t_2^N) = \sum_{i=1}^{N} C^t(t_i, u_i) + \sum_{i=2}^{N} C^c(u_{i-1}, u_i)$$

Takto nalezená posloupnost řečových jednotek je poté zřetězena do jedné souvislé zvukové stopy, která představuje výslednou syntetickou řeč. Na obrázku 2.3 je ilustrován příklad české syntézy výběrem jednotek. Tento obrázek a popis byl převzat z práce [8, strana 32], která se metodou Unit Selection podrobně zabývá.



Obrázek 2.3: Jednoduché znázornění principu syntézy slova *„syntéza“* vygenerované pomocí syntézy výběrem jednotek.

Z uvedených informací vyplývá důležitá informace využitá v této práci. Resyntézou věty, která byla nahrána do řečového korpusu, získáme identickou větu, tedy přirozenou promluvu. Je-li algoritmus výběru správný, mají jednotky resyntetizované řeči nulovou cenu cíle a cenu konkatenace. Nahrávka je pouze zřetězena zpět do své původní podoby.

Výhody a nevýhody syntézy výběrem jednotek

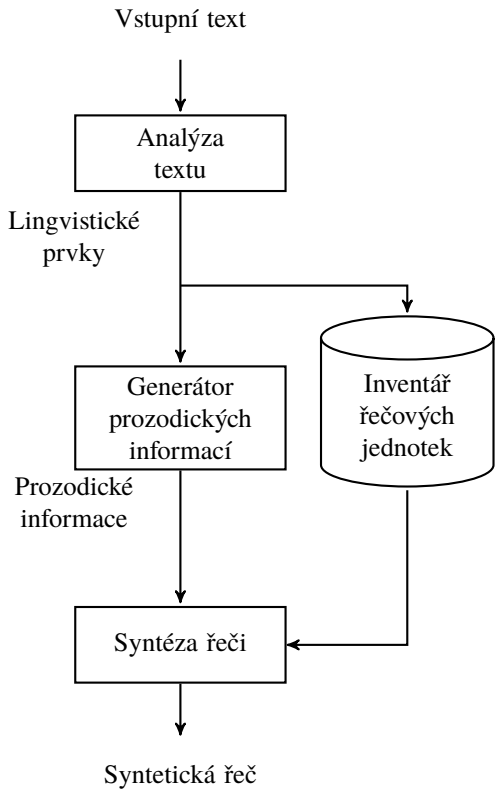
Kvalita syntézy výběrem jednotek je výrazně lepší, než u SUI syntézy, ovšem za cenu vyšší výpočetní náročnosti oproti metodě SUI. Syntéza Unit Selection využívá obrovské řečové inventáře – není tedy potřeba provádět žádné signálové modifikace, které zvukový signál v jisté míře zkreslují a snižují tím přirozenost generované řeči. Výhodou této syntézy je přirozenost prozodie při správném výběru řečových jednotek. Velmi značnou nevýhodou je naopak náchylnost na výskyt artefaktů v místě konkatenace jednotek, možné střídání dobrých a špatných úseků řeči a malá flexibilita

2.5 Systémy syntézy řeči z textu

Jedním z nejrozšířenějších využití syntézy řeči je syntéza řeči z textu, nebo také konverze textu na řeč (hojně používá zkratka TTS z anglického Text-to-Speech) [3]. Cílem TTS systému je umožnit vygenerování libovolné promluvy ze vstupní informace, která je ve formě textu. V ideálním případě by měl být systém TTS schopný produkovat přirozeně znějící libovolnou promluvu, která nebude vyžadovat zvýšenou pozornost při poslechu. Je ale zřejmé, že řešení takové úlohy je velice komplikované.

2.5.1 Architektura TTS

Na obrázku 2.4 je uvedeno jednoduché schéma architektury systému převodu textu na řeč. Na tomto schématu je znázorněn proces analýzy a zpracování vstupní informace, kterou je třeba provést před finální syntézou řeči. Prvním krokem je zpracování přirozeného jazyka, kde jsou na základě vstupní informace generovány prozodické informace a posloupnosti hlásek. Poté jsou speciálním algoritmem vybrány vhodné akustické jednotky z inventáře řečových jednotek a výsledná promluva vygenerována syntetizérem řeči.



Obrázek 2.4: Základní architektura jednoduchého TTS systému

Příkladem českého systému převodu textu na řeč je systém ARTIC.

2.5.2 TTS systém ARTIC

Ačkoliv jsou počátky české konkatenační syntézy datovány již od roku 1972 [3], významný vývoj české TTS syntézy nastal až v roce 1990, kdy pracovníci Ústavu radiotechniky a elektroniky Akademie věd České republiky vytvořili český parametrický konkatenační systéms TTS s názvem EPOS, který využíval manuálně vytvořený inventář řečových jednotek [12].

Od roku 1999 je na Katedře kybernetiky, Fakultě aplikovaných věd Západočeské univerzity v Plzni taktéž vyvíjen moderní TTS systém vysoké kvality ARTIC [8].[6] Jedná se o vícejazyčný TTS systém založený na korpusově orientované konkatenační syntéze řeči [2, 3].

[6]Název ARTIC je akronymem vzniklým z „ARtificial Talker In Czech“ a jedná se o český TTS systém

Kapitola 3

Analýza chyb v syntetické řeči s výběrem jednotek

Jelikož je navrhovaná metoda hodnocení kvality syntetické řeči připravována pro konkatenační syntézu výběrem jednotek, omezíme následující analýzu artefaktů v syntetické řeči pouze na chyby vyskytující se v řeči vygenerované touto metodou. Studiem možných chyb v konkatenační syntéze vytvoříme podklad pro klasifikaci chyb, které budou manuálně zanášeny do resyntetizované (přirozené) řeči. Manuálním vložením chyb získáme důležitou informaci, a to, v jakém místě a čase se daná chyba vyskytuje. Posluchači takto porušené řeči bude během poslechu zaznamenávána mozková aktivita pomocí EEG. Na výstupním signálu bude sledováno okolí známého výskytu chyby se snahou potvrdit hypotézu vzniku nestandardní[1] odezvy mozkové aktivity v okolí vneseného řečového artefaktu. Začneme studium chyb konkatenační syntézy definicí řečového artefaktu.

3.1 Definice řečových artefaktů

Řečové artefakty vznikají povětšinou v místech konkatenace řečových jednotek při vytváření řeči, která zní kvůli přítomnosti těchto rušivých elementů nepřirozeně. Jednotky jsou z řečového korpusu sice vybírány určitou hodnotící funkcí tak, aby na sebe co nejlépe navazovaly, ovšem ne vždy je zaručena stoprocentní spolehlivost tohoto výběru. Důležitým faktorem je kvalita přirozené řeči, která byla nahrána do řečového korpusu. Chyby v syntetické řeči mají výrazný vliv na její kvalitu, tedy i srozumitelnost a příjemnost poslechu.

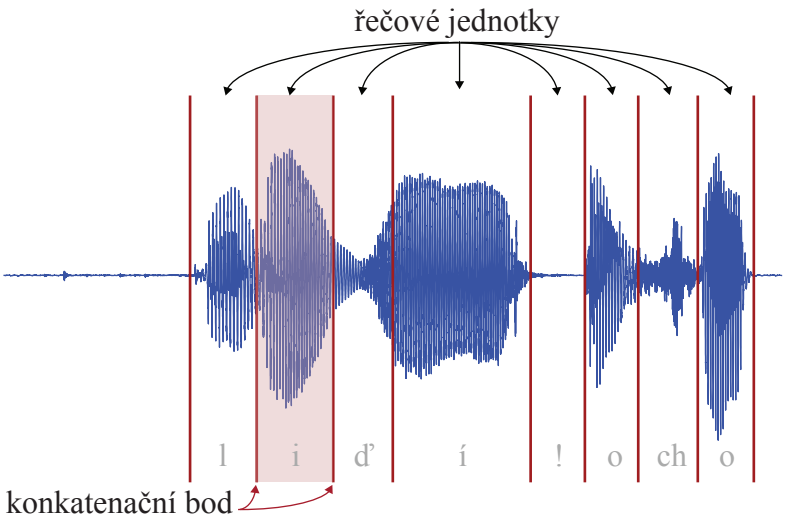
[1] Termínem „nestandardní“ je v našem případě myšleno srovnání výsledků s bezchybnou resyntetizovanou řečí, nikoliv nestandardní fyziologický signál.

3.2 Vznik chyb v syntetické řeči

Chyby v syntetické řeči mohou vzniknout v různých úrovních procesu jejího vytváření. Chyba může být do syntetické řeči vnesena již při tvorbě řečového korpusu, například změnou barvy hlasu řečníka nahrávajícího potřebná zvuková data, popřípadě následnou segmentací korpusu. To nám velice rozšiřuje body ve zpracování a tvorbě řeči, kde může dojít ke vzniku řečového artefaktu a tím pádem i znatelnému propadu kvality výsledné syntetické řeči. Z pohledu členění řeči při vytváření tak pro přehled můžeme chyby rozdělit do dvou základních skupin,[2] a to na chyby na segmentální a na suprasegmentální úrovni.

3.2.1 Chyby na segmentální úrovni

Mezi chyby na segmentální úrovni jsem klasifikoval takové artefakty, které se vyskytují pouze v rámci jednoho segmentu. Může se tak jednat například o chybu v algoritmu výběru jednotek, špatnou segmentaci nebo lokální nespojitost F_0 v místě konkatenace řečových jednotek (pro lepší představu a porozumění je segmentovaná řeč spolu s konkatenačními body zobrazena na obrázku 3.1).



Obrázek 3.1: Segmentovaná nahrávka

[2] Klasifikace chyb do dvou skupin, tj. segmentální a suprasegmentální, může být v určitých ohledech a do jisté míry zavádějící, proto u každého rozdělení uvádím, podle čeho a jaké chyby jsem se rozhodl do dané skupiny klasifikovat

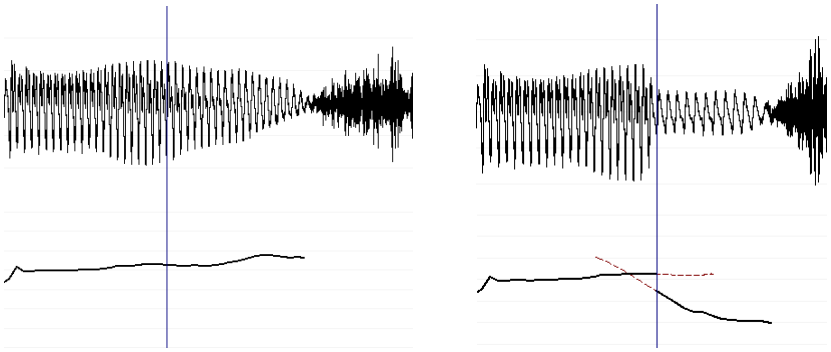
Chybná segmentace řečových jednotek

Řečový artefakt může vzniknout už při samotné segmentaci, kde při jejím špatném provedení řečový signál neodpovídá dané jednotce a ve výsledné řeči tak dojde ke zřetelně slyšitelné chybě. To je způsobeno chybou návazností signálů spojovaných řečových jednotek.

Nalézt místo výskytu chyby tohoto typu není obtížné a člověk znalý syntetické řeči musí manuálně danou jednotku opravit (například opětovnou, manuální segmentací daného úseku, nebo jednotku z inventáře odstranit).

Nespojitost základního hlasivkového tónu

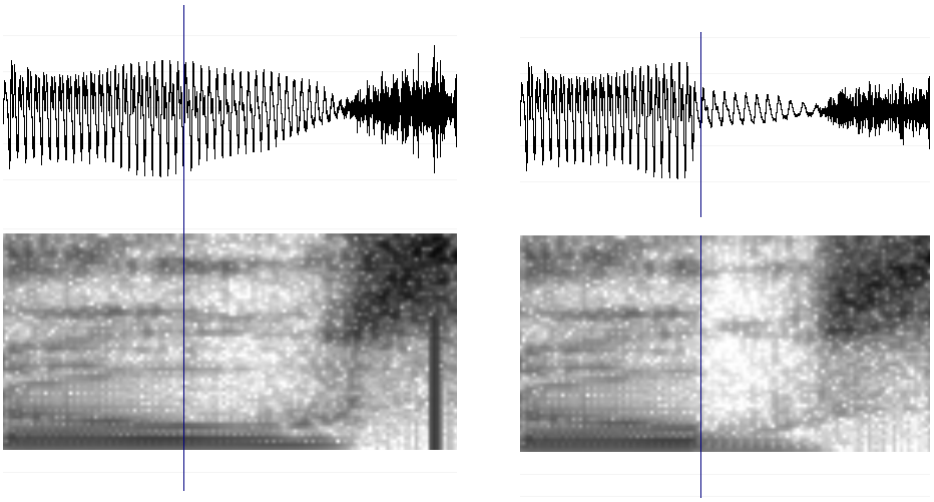
Velmi častým a zřejmě nejvýrazněji slyšitelným artefaktem v řečové syntéze je chyba zapříčiněná nespojitostí průběhu frekvence základního hlasivkového tónu F_0. Tato chyba vzniká v místech konkatenace řečových jednotek, které nemají stejnou frekvenci F_0, resp. není v těchto místech spojitá. Vyskytuje se pouze u znělých hlásek a je nejvíce slyšitelná u samohlásek [4, 11]. Frekvence F_0 má značný vliv na vznik řečového artefaktu, proto je součástí hodnotící funkce algoritmu výběru řečových jednotek u metody Unit Selection. Tento typ chyby lze dobře odhalit analýzou průběhu frekvence F_0, jak je zobrazeno na obrázku 3.2.



Obrázek 3.2: Nespojitost základního hlasivkového tónu – Průběh zvukové vlny a frekvence F_0 originálního (vlevo) a porušeného (vpravo) místa konkatenace (druhá jednotka byla vyměněna za nejméně vhodného kandidáta)

Nespojitost ve spektru

Podobně jako u nespojitosti průběhu frekvence F_0 i u nespojitosti ve spektru platí, že čím větší je rozdíl v místě konkatenace, tím větší je pravděpodobnost výskytu rušivého artefaktu, viz obrázek 3.3.



Obrázek 3.3: Nespojitost ve spektru – Průběh spektra originální části promluvy (vlevo) a porušené části promluvy (vpravo), kde je na první pohled vidět ostrý přechod ve spektru.

Změna tempa a délky trvání jednotky

Řečový artefakt může vzniknout i v situaci, kdy byla jako vhodný kandidát vybrána řečová jednotka, která sice byla svým průběhem frekvence F_0 spojitá a veškeré podmínky při výběru splňovala, ovšem trvání či délka této jednotky je v rámci slova neadekvátní.

3.2.2 Chyby na suprasegmentální úrovni

Chyby na suprasegmentální úrovni se projevují až v rámci skupin segmentů, tedy na úrovni slabik, slov, jejich spojení nebo na celých větách. Mohou vzniknout například jiným rozpoložením hlasu mluvčího, který nahrává věty do řečového korpusu. Fyzické i psychické rozpoložení řečníka může ovlivnit tempo či barvu jeho hlasu, což má později následky při tvorbě syntetické řeči. Z tohoto důvodu je velice důležité, aby samotné nahrávání bylo technicky optimalizované[3] a hlas řečníka byl po celou dobu nahrávání pokud možno stejně zabarvený. Ve většině případů se proto pro nahrávání potřebných zvukových dat volí profesionální řečník či moderátor, který je ze zkušenosti zvyklý na delší promluvy a dokáže promlouvat neutrálním hlasem.

Z pohledu vzniku chyb na suprasegmentální rovině je většinou těžko patrné, co ji způsobilo. Algoritmem vybraná jednotka při tvorbě řeči může být dle parametrů vhodná, jednotka sama zní přirozeně, ovšem při poslechu celého slova či věty již zní nepřirozeně a

[3] Je důležité a nezbytné, aby byla nahrávací technika kvalitní (kvůli zamezení velkého šumu) a prostředí, ve kterém se nahrává, nehlučné.

působí rušivě. Oblast suprasegmentálních jevů a jejich vliv na přirozenost strojové syntézy řeči je detailně popsána v práci [7].

3.2.3 Shrnutí poznatků

Analýzou artefaktů vyskytujících se v konkatenační syntéze řeči jsme získali podklady k první části přípravy dat pro experiment. Na základě získaných znalostí bude třeba definovat chyby, které budeme manuálně vnášet do resyntetizované řeči a místo jejich výskytu bude sledováno na EEG. Tím získáme nad výskytem chyb kontrolu. Navrhované chyby by měly odpovídat definované klasifikaci, což bude třeba empiricky ověřit.

Kapitola 4

Hodnocení kvality syntetické řeči

Než se začneme podrobněji zabývat návrhem metody hodnocení kvality syntetické řeči pomocí EEG a klasifikací chyb před přípravou dat, bude vhodné si krátce shrnout běžně používané metody hodnocení kvality syntetické řeči. Takovou metodou bychom také mohli empiricky ověřit námi definované klasifikace chyb.

4.1 Poslechové testy

Hodnocení kvality syntetické řeči může být pojato ve dvou přístupech. Poměřujeme-li původní promluvu s rekonstruovanou řečí, jde o poměrně jednoduchou záležitost a lze subjektivně ohodnotit kvalitu rekonstruované řeči. V případě hodnocení kvality libovolné promluvy generované TTS systémem však musíme postupovat jinak, neboť nemáme, nebo nemusíme mít k dispozici původní promluvy řečníka.

Nejběžněji používanou metodou pro hodnocení kvality syntetické řeči jsou standardizované testovací metody - **poslechové testy**. Pomocí těchto testů skupiny dobrovolníků subjektivně hodnotí syntetickou řeč. Navíc nás tyto testy mohou upozornit na chyby syntetizéru a částečně pomáhají tyto chyby eliminovat. Hodnocení formou poslechových testů není triviální záležitost, neboť se jedná o multidimenzionální problém (je třeba hodnotit prozodické vlastnosti, srozumitelnost, přirozenost, styl promluvy, rychlost promluvy atp.). Z hlediska zaměření můžeme poslechové testy rozdělit do dvou skupin:

1. **Testy srozumitelnosti**, které se zaměřují zejména na koartikulaci a na to, jak dobře posluchač rekonstruované řeči rozumí. Testy se tedy omezují na segmentální úroveň řeči (viz sekce 4.1.1).

2. **Testy přirozenosti**, které hodnotí řeč jako celek, hodnotí tedy suprasegmentální kvalitu rekonstruované řeči (viz sekce 4.1.2).

4.1.1 Testy srozumitelnosti

Testy srozumitelnosti hodnotí kvalitu syntetické řeči na segmentální úrovni, tj. úrovni jednotlivých fonémů či slov. Mezi testy srozumitelnosti patří MRT testy (z anglického Modified Rhyme Test), případně jejich modifikace DTR (Diagnostic Rhyme Test) a SUS testy (z anglického Semantically Unpredictable Sentences) [3, 5].

- **MRT testy** se skládají ze skupin několika rýmujících se jednoslabičných slov, respektive ze stejně znějících jednoslabičných slov. Všechna slova v každé skupině se liší pouze první nebo poslední souhláskou. Slova ve skupinách mají tzv. CVC strukturu, neboli souhláska-samohláska-souhláska (z anglického Consonant-Vowel-Consonant). Úkolem posluchačů je napsat slovo, které slyšeli v syntetizované nahrávce nebo identifikovat slovo, které slyšeli, v dané skupině slov. V každém kroku testování si před hodnocením posluchač danou nahrávku smí jedenkrát přehrát.

lev	les	lem	led	lep	len
suk	puk	kuk	luk	muk	fuk

Tabulka 4.1: Příklad českých slov navržených pro MRT test[1]

- **SUS testy** se stejně jako MRT testy soustředí na hodnocení srozumitelnosti syntetické řeči. Princip těchto testů spočívá v syntakticky správných, ale bezvýznamných větách. To nutí posluchače, který kvalitu syntézy hodnotí, snažit se porozumět každému slovu ve větě. Tím pádem tato metoda potlačuje to, že by si mohl posluchač část slov nebo význam věty domyslet.

Jezte	**lesní**	**stoly**	**i**	**vejce**.
sloveso	*příd. jm.*	*podst. jm.*	*spojka*	*podst. jm.*
Čtecí	**bojovník**	**byl**	**běžný**	**parník**.
příd. jm.	*podst. jm.*	*sloveso*	*příd. jm.*	*podst. jm.*

Obrázek 4.1: Příklady sestavených vět v SUS testu (rozkazovací struktura) navrženého pro hodnocení českých TTS systémů (příklady převzaty z [3, strana 629])

[1] Slova byla převzata z [5, strana 2], kde je možné se o MRT testu dočíst více.

4.1.2 Testy přirozenosti

Testy přirozenosti hodnotí kvalitu syntetické řeči na suprasegmentální úrovni, tedy celkovou[2] kvalitu syntetizovaných slov či vět. Hodnotí se tedy to, jestli zní věta jako celek přirozeně a posluchač nemusí vynakládat větší úsilí k porozumění při poslechu syntetické řeči. Mezi tyto testy patří MOS (Mean Opinion Score) a CCR (Comparison Category Rating) testy.

- **MOS testy** jsou velice rozšířené a pro hodnotící subjekt jednoduché. Posluchač hodnotí kvalitu vygenerované řeči na pětistupňové škále kvality syntetické řeči (viz tabulka 4.2), nebo na podobné škále, ovšem zaměřené na námahu poslechu vygenerované řeči, respektive úsilí vynaložené k porozumění umělé řeči (viz tabulka 4.3).

Hodnocení	Kvalita řeči
1	špatná
2	horší
3	slušná
4	dobrá
5	výborná

Tabulka 4.2: Škála pro hodnocení kvality syntetické řeči

Hodnocení	Námaha při poslechu řeči
1	zcela neadekvátní
2	značná
3	"průměrná"
4	bez značné námahy
5	bez jakékoli námahy

Tabulka 4.3: Škála pro hodnocení příjemnosti poslechu syntetické řeči

- **CCR testy** slouží především pro porovnávání dvou TTS systémů, případně dvou různých verzí stejného systému nebo metod použitých pro tvorbu syntetické řeči (označme tyto dva TTS systémy například jako X a Y). Vztah kvality mezi systémem X a Y tak můžeme hodnotit například na sedmistupňové škále (viz tabulka 4.4) nebo na zjednodušené stupnici (viz tabulka 4.5). V obou případech má posluchač možnost si vygenerovanou řeč při hodnocení přehrávat tolikrát, kolikrát potřebuje.

[2]Jako celkovou kvalitu řeči můžeme považovat hodnocení plynulosti projevu, námahy poslechu, přirozenosti, kvality intonace a dalších aspektů, které umělou řeč přibližují k řeči přirozené.

3	-	Systém X je mnohem lepší, než systém Y
⋮	-	⋮
-3	-	Systém X je mnohem horší, než systém Y

Tabulka 4.4: Škála pro porovnávání kvality dvou TTS systémů

1) Dávám přednost X
2) Zhruba stejné
3) Dávám přednost Y

Tabulka 4.5: Zjednodušená stupnice pro porovnávání kvality dvou TTS systémů (stupnice může být i pouze dvoubodová – Preferuji X/Preferuji Y)

4.1.3 Nevýhody poslechových testů

Ačkoliv jsou poslechové testy nejrozšířenější a nejčastěji používanou metodou pro hodnocení kvality syntetické řeči, mohou být nedostačující pro účely optimalizace procesu syntézy řeči za účelem zvýšení její přirozenosti. Pokud se budeme snažit poslechovými testy zjistit, zda je nové nastavení parametrů TTS systému lepší, než předchozí, je nutné pro každý parametr nebo každou jeho změnu vytvořit nový poslechový test. To je náročné nejen pro autory testu, ale zejména pro dobrovolníky, kteří test podstupují. Hodnocení pomocí EEG by tento problém velice zjednodušilo tím, že by dobrovolník pouze při měření na EEG poslouchal syntetickou řeč vygenerovanou TTS systémem bez nutnosti vědomého a verbálního hodnocení. Analýzou naměřeného signálu by se pak na základně množství signálů značících výskyt chyby rozhodlo, zda je nové nastavení parametrů lepší, či nikoliv.

4.2 Experimentální metoda EEG

4.2.1 Metoda EEG a její význam

Zatímco v běžném testování kvality syntetické řeči se používá nějaký typ výše zmíněných poslechových testů nebo jejich modifikace, alternativní metoda hodnocení pomocí EEG přivádí k této problematice nový úhel pohledu. Tato poměrně nová metoda se nesnaží pouze modifikovat poslechové testy, nýbrž je tu snahou pomocí neurobiologického přístupu nalézt souvislost mezi kvalitou TTS systému a lidským kognitivním a emočním vnímáním. Tím tento přístup otevírá dveře novým metodám pro hodnocení kvality syntetické řeči [6].

4.2.2 Hodnocení celkové kvality TTS pomocí EEG

Principem této alternativní metody je sledování neurofyziologického signálu získaného z měření zkoumaného subjektu, následnou analýzou získaného signálu a hledáním souvislostí mezi kvalitou syntetické řeči a tímto signálem.

Ve studii [6] se kanadští vědci snažili pomocí alternativní EEG metody hodnotit celkovou kvalitu TTS systémů ve srovnání s vybranými poslechovými testy dle normy ITU-T Recommendations P-85 od Mezinárodní telekomunikační unie [13]. Vědci v této studii potvrdili vztah mezi odezvou EEG P300[3] a vnímáním kvality TTS systémů. Hodnocení kvality syntetické řeči se tak přesunulo i do neurobiologických oblastí, což může v budoucnu přispět ke zvýšení přirozenosti a zvýšení kvality současně vyvíjených systémů převodu textu na řeč.

4.3 Navrhovaný experimentální protokol

4.3.1 Plánovaný průběh experimentu

Výzkum uvedený v [6] se nám stal inspirací k provedení vlastního experimentu pomocí EEG. Důležitým cílem bude ověřit hypotézu, zda se v naměřeném EEG signálu při poslechu porušené syntetické řeči bude vyskytovat nestandardní odezva mozku v okolí výskytu této chyby. Sledováním naměřených aktivit na EEG se budeme snažit ověřit tuto hypotézu.

Pokud by se hypotéza potvrdila, využila by se detekce těchto specifických změn v signálu k hodnocení kvality dané syntetické řeči. Mohl by to být základ nové objektivnější testovací metody ke zvyšování kvality a přirozenosti současně vyvíjených TTS systémů. Navíc by tato metoda mohl být rychlejší a efektivnější, než jsou poslechové testy. Plánovaná osnova navrhovaného experimentu je následující:

1) Resyntéza 300 nahrávek obsažených v řečovém korpusu
2) Manuální zavedení právě jedné chyby do každé ze 150 připravených nahrávek. Tato chyba představuje v experimentu **stimul**.
3) Příprava skriptů pro řízení experimentu na EEG s připravenými stimuly
4) Sledování stimulů na výstupním signálu z EEG a analýza získaných dat

[3] P300 je označení pro pozitivní maximum EEG signálu objevující se zhruba 300ms po onsetu (začátku) stimulu (v našem případě stimul = chyba v syntetické řeči).

4.3.2 Klasifikace typů chyb pro experiment

V kapitole 3 jsme analyzovali možné chyby vyskytující se v konkatenační syntéze výběrem jednotek. Na základě toho teď vytvoříme klasifikaci chyb, které budou manuálně vkládány do resyntetizované řeči pro následné měření. Touto klasifikací chyb se poté budeme řídit v praktické části úlohy.

Rozdělení míry chyb pro experiment

Původně bylo plánováno klasifikovat tři různé typy chyb – malé, střední a velké. Po vyzkoušení a následném subjektivním zhodnocení jsme ale došli k závěru, že střední chyby bychom ve většině případů klasifikovali spíše jaké malé, resp. velké. Z tohoto důvodu jsme se rozhodli vytvořit klasifikaci pouze dvou typů chyb – malých a velkých. Pro experiment jsme tedy definovali termíny:

Originál: Resyntetizovaná řeč bez jakékoliv modifikace je v rámci experimentu označována jako originální nahrávka (zkratkou *org* nebo *Original*). Tyto nahrávky budou později sloužit ke komparaci EEG signálu naměřeného u porušené a neporušené řeči.

Malá chyba: Méně slyšitelné chyby, v experimentu označované zkratkou *sgl* nebo také *Small Glitch*, byly zanášeny na místa konkatenace dvou řečových jednotek modifikací jednoho z difonů. Vlastním subjektivním hodnocením jsme zjistili, že tento typ chyby je nejlepší zanášet na krátké samohlásky a nosovky.

Velká chyba: Zřetelně slyšitelné chyby, označované v experimentu zkratkou *bgl* nebo také *Big Glitch*, představují artefakty razantně narušující přirozenost a kvalitu vygenerované řeči. Povětšinou se jedná o artefakty vzniklé nespojitostí základní frekvence F_0. Opět subjektivním hodnocením při zanášení chyb jsme dospěli k názoru, že tento typ chyby je nejlepší zanášet na dlouhé samohlásky, kde je intenzita tohoto artefaktu násobena právě délkou řečové jednotky.

4.3.3 Poslechový minitest

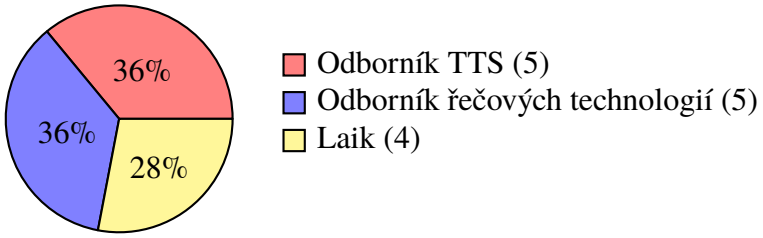
K empirickému ověření klasifikace chyb zanášených do syntetické řeči (popsané výše) byl na základě znalostí získaných studiem poslechových testů na začátku této kapitoly

vytvořen jednoduchý, dvacetibodový poslechový minitest. K jeho realizaci byla využita webová služba Google Forms. Pro účely tohoto testu bylo resyntetizováno 20 heuristicky vybraných vět, do nichž byla na základě vytvořené klasifikace (4.3.2) zanesena vždy jen jedna chyba jednoho typu. Cílové skupiny dobrovolníků testu byly laici,[4] odborníci na systémy TTS a odborníci řečových technologií obecně. Účastníci testu měli na hodnotící škále (viz tabulka 4.6) rozhodnout, jaké klasifikaci odpovídá chyba zanesená do testovací nahrávky, tedy jestli se jedná o malou či velkou chybu, případně o tom nelze rozhodnout.

(1) Malá chyba
(2) Velká chyba
(3) Chyba není slyšitelná nebo nejde klasifikovat

Tabulka 4.6: Škála použitá u poslechového testu k ověření správné klasifikace chyb zanesených do syntetizovaných vět

Celkem se testu zúčastnilo 14 osob, jejichž distribuce je zobrazena na obrázku 4.2. Shodou okolností jsou počty vybraných skupin vesměs vyrovnané. Počet účastníků poslechového testu není velký, ovšem pro účely ověření klasifikace byl tento počet dostačující.



Obrázek 4.2: Graf znázorňující distribuci účastníků poslechového testu

Výsledky testu

Výsledky testu v převážné většině potvrdily správnou klasifikaci námi uměle zanášených chyb do resyntetizované řeči. Tabulka 4.7 ukazuje případy, kdy účastníci poslechového testu aposteriorně potvrdili apriorní klasifikaci chyby (shoda klasifikace je větší než 85%, označeno ■), vyvrátili apriorní klasifikaci chyby (zanesená chyba neodpovídá klasifikaci, označeno ■), nebo byl výsledek hodnocení nejasný (chyba byla klasifikována do obou skupin ve stejném poměru, označeno ■).

[4]Za laika je zde považován člověk, který má malé či nemá žádné povědomí o problematice syntetické řeči a chybách v ní

☐	**Text věty**	**Typ**	**Velká chyba**	**Malá chyba**
■	To je otázka inkvizice a křížových výprav	SGL	10	4
■	Vladimír Růžička znovu jednou nadchl	SGL	2	12
■	Uvedl to primátor Plzně Jiří Šneberger	BGL	13	1
■	Jméno nového šéfa banky však nechtěl říci	SGL	1	13
■	S fanoušky v zádech býváme těžko k poražení	BGL	13	1
■	Německá inflace stlačí cenu dluhopisů	SGL	7	7
■	Vytvořím širší bázi pro náš vztah k Moskvě	BGL	13	1
■	Proti tomu však mluví různá svědectví	SGL	7	7

Tabulka 4.7: Seznam vět vyvracejících/potvrzujících správnost klasifikace zanášených chyb (odpovědi, které nerozdělují chybu do žádné skupiny nejsou zobrazeny)

Zvýrazněný text v tabulce 4.7 vyznačuje místo,[5] kam byl artefakt uměle vnesen. Výsledky testu potvrdily apriorní subjektivní klasifikaci, čímž byl heuristicky definován způsob vytváření chyb zanášených do resyntetizované řeči pro účely experimentu. Přehled všech vět použitých v poslechovém minitestu najdete na konci této práce v příloze A. Na základě takto připravené klasifikace můžeme nyní začít s přípravou dat pro experimentální měření.

[5] Místo vneseného artefaktu je označeno zvýrazněnými hláskami daného difonu.

Kapitola 5

Příprava experimentu

5.1 Příprava experimentálních dat

Prvním krokem k úspěšnému experimentu je vytvoření kvalitního, dostatečně bohatého a dobře strukturovaného souboru dat. Aby bylo možné zavádět specifické chyby do resyntetizované řeči, je nezbytné vybrat taková vstupní data, která budou vyhovovat účelu experimentu. Do resyntetizované řeči pak bude zvolenými prostředky manuálně zanesena vždy jedna chyba a do tabulky zaznamenány veškeré potřebné informace pro měření (onset/offset[1] chyby, text nahrávky, klasifikace zanesené chyby apod.).

Pro připravená zvuková data budou vytvořeny skripty, které tato data budou zkoumanému subjektu vhodnou formou interpretovat a současně komunikovat s EEG zařízením. V poslední řadě bude nutné zajistit integritu celého systému, aby ve finální formě fungoval podle představ a poskytoval předpokládané výstupy a výsledky.

5.1.1 Použité technologie

Pro zpracování vstupních dat v textové podobě byl použit programovací jazyk Visual C# a regulární výrazy knihoven Microsoft .NET frameworku. K syntéze vybraných vět a zanášení chyb do řečové syntézy byl použit program Prokus (více o programu v 5.1.3). Nahrávky byly upravovány ve freeware programu Audacity. Skripty pro řízení experimentu na EEG byly vytvořeny v aplikačním prostředí Presentation, pro které mi byla v rámci této práce poskytnuta licence.

[1] Pro pořádek zde uvedu význam použité terminologie: onset = začátek/počátek, offset = konec

5.1.2 Výběr vhodných vět pro experiment

Abychom mohli vytvořit resyntézu řeči, je pro nás důležité získat přesné znění frází, které byly nahrány do řečového korpusu při jeho tvorbě. Výstupem pak dostaneme identickou řeč řečníka, který tyto fráze nahrával, tedy přirozenou a neporušenou řeč. Je tedy nutné vybírat data uložená přímo v řečovém korpusu. Použitý řečový korpus systému ARTIC ovšem obsahuje obrovské množství nahraných frází (v řádech tisíců vět). Výběr vhodných vět bude proto filtrován a budou vybrány pouze pouze věty vhodné k experimentu. Kvůli technickým podmínkám měření jsme se rozhodli vybírat data splňující následující podmínky:

1. Syntetizovaná věta by měla být jednoduchá, krátká a nemělo by se jednat o souvětí.
2. Délka nahrané věty by měla být v rozmezí čtyř až pěti sekund. Tuto podmínku jsme zavedli zejména kvůli měření na magnetické rezonanci v možném budoucím pokračování tohoto výzkumu.

Zdroj dat

Věty potřebné pro resyntézu byly vybírány ze zkráceného souboru ASF,[2] který obsahuje značné množství informací o každé větě uložené v řečovém korpusu. Data tohoto souboru byla filtrována tak, že u každé věty byla nalezena poslední řečová jednotka představující ticho (označeno _SIL_). Pokud byl čas konce této jednotky v rozmezí 4 až 5 sekund, byla vybrána jako vhodný kandidát a uložena do seznamu kandidátů (tabulka Excelu). Tabulka 5.1 poskytuje náhled na strukturu zdrojového souboru a sledované části dat.

[2] Soubory ASF jsou rozšířením formátu Master Label File (MLF) a obsahují například i prozodická slova. Soubor je v případě této práce generován přímo TTS systémem ARTIC.

"oznam00001_00"

phones				word	mlfBegTime	mlfEndTime	modelScore
$	P	P	0	_SIL_	0.0	0.277	-58.033958
t	F	F	0	tehdy	0.277	0.313375	-70.565437
e	-	-	0	.	0.313375	0.3893125	-65.412216
h	-	-	0	.	0.3893125	0.498625	-65.102638
d	-	-	0	.	0.498625	0.5655	-81.895042
i	L	-	0	.	0.5655	0.6485	-69.920311
				⋮			
L	L	L	1.1	.	3.737875	3.907	-75.070724
$	P	P	0	**_SIL_**	3.907	**4.146**	-96.789253

Tabulka 5.1: Ukázka struktury ASF souboru s daty (zvýrazněné části značí sledované části zdrojových dat)

Takto vyfiltrované věty, které se staly potencionálními kandidáty pro experiment, byly vyhledány v anotačním souboru obsahujícím text vět nahraných do řečového korpusu, viz tabulka 5.2. Z těchto vět pak byly vybírány pouze věty jednoduché neobsahující čárky. Z celkového počtu 8876 vět tomuto vyhovovalo 782 vět, z nichž bylo heuristicky vybráno 300 vět, což jsme usoudili jako dostačující pro tento výzkum.

dopln00402_00	Jaká zařízení jsou v současné době používána {poûžývána}?
oznam00001_00	**Tehdy v sedmnáctém kole jsem udělal chybu a vypadl.**
oznam00016_00	Tímto krokem chce burza vrátit na trh likviditu {lykvydytu}.
oznam04392_00	Hospodářský růst, zejména v našich podmínkách, vypovídá o něčem důležitém.
	⋮

Tabulka 5.2: Ukázka struktury a vyhledávání v anotačním souboru

5.1.3 Zavádění chyb do řečové syntézy

Pro získané věty z anotačního souboru řečového korpusu bylo nutné provést jejich resyntézu a do ní pak následně vložit jednu chybu na základě klasifikace (4.3.2). K tomu byl použit program Prokus.[3]

[3]Prokus je kódové označení programu naznačující nejčastější činnost uživatele - „prokousávání se syntézou“ [4].

Resyntéza řeči v programu Prokus

Prokus je počítačový program, který slouží jak pro analýzu procesu syntézy řeči, tak i jako nástroj pro odhalování některých typů chyb v syntetické řeči [4]. V této aplikaci je možné využít TTS systém ARTIC a syntetizovat libovolný text na syntetickou řeč. K našemu účelu jsme však tuto aplikaci použili pouze k resyntéze vybraných frází a zavádění námi definovaných chyb. Program totiž umožňuje upravovat segmenty a vyměňovat jednotlivé řečové jednotky za jejich alternativní kandidáty.

Vytvořenou resyntetizovanou řeč bylo vždy třeba zkontrolovat, jestli syntetizér větu poskládat opravdu z původních úseků přirozené řeči. Pokud by tak nebylo, mohlo by to později vést ke zkreslení výsledků měření a ztrátě přehledu o chybách v dané řeči. V případě chybné resyntézy bylo nutné špatné úseky manuálně opravit (výměnou řečové jednotky) nebo zvolit frázi jinou.

Zavedení chyby

Abychom mohli naměřený signál porušené syntetické řeči porovnávat se signálem naměřeným při poslechu přirozené (neporušené) řeči a provádět komparaci těchto signálů, stačilo zanášet chybu pouze do poloviny připravených nahrávek. Do 150 nahrávek jsme tedy vnesli vždy jednu chybu zvolené kategorie. Tímto způsobem jsme připravili celkem 75 nahrávek s malou chybou a 75 s chybou velkou. V každé resyntetizované větě bylo třeba vybrat jednu řečovou jednotku, která byla následně nahrazena alternativní jednotkou s menší cenou cíle a cenou konkatenace.

Při zanášení zvolených chyb jsem na základě rozboru v (4.3.2) a výsledku poslechového minitestu (4.3.3) postupoval následovně:

Malá chyba byla zanášena výběrem řečové jednotky s vyšší cenou cíle a cenou konkatenace. Bylo důležité mít pod kontrolou, aby chyba nebyla příliš rušivá, což by bylo v rozporu s její klasifikací. Poškozenou nahrávku jsem si tak několikrát poslechl a alternativy v případě nespokojenosti měnil.

Velká chyba byla zanášena obdobně, jako chyba malá, jako alternativní jednotka však byla vybrána ta nejméně vhodná, tedy s největší cenou cíle a cenou konkatenace. V tomto případě se dalo dobře řídit skokovou nespojitostí průběhu frekvence F_0.

Stejně jako v předchozím případě jsem si nahrávky poslechl a zhodnotil vhodnost zvolené alternativy, kterou jsme v případě nespokojenosti opětovně změnil za jinou.

Současně se zanášením chyb byly zaznamenány veškeré potřebné informace o vnesené chybě do tabulky, jejíž struktura je zobrazena v tabulce (5.3). Tyto poznatky budou dále využity při měření na EEG. Mezi nejdůležitější zaznamenávané parametry patří **onset** chyby (čas začátku vnesené chyby v milisekundách), **offset** chyby (čas konce vnesené chyby v milisekundách), **klasifikace** zanášené chyby, **text** syntetizované věty a **unikátní název** nahrávky z důvodu jejich dalšího použití při tvorbě skriptů k řízení experimentu.

ID	Corpus_ID[4]	Text	Glitch[5]	Onset	Offset	Wave_name
148	oznam02036_00	Švýcaři včera odešli už po pár minutách	org	-	-	record_148_org.wav
⋮	⋮	⋮	⋮	⋮	⋮	⋮
188	oznam02705_00	Holky v bazénu bloudí a nevědí o sobě	sgl	2.8438	2.9313	record_188_sgl.wav
189	oznam02720_00	Hůře se vyvíjí pouze situace ve stavebnictví	bgl	2.2188	2.3375	record_189_bgl.wav
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Tabulka 5.3: Struktura připravovaných dat pro další zpracování v programu Presentation (v praxi se jedná o strukturovaný soubor CSV, který je koncipován jako tabulka, kde jsou středníky ";"odděleny jednotlivé sloupce)

5.1.4 Úprava délky nahrávek

Připravené nahrávky bylo ještě třeba určitým způsobem finalizovat. Pro měření bylo nezbytné zajistit, aby každá z nahrávek měla přesně 5 sekund. K tomuto účelu byl do programu Audacity vytvořen plugin napsaný v jazyce Nyquist, který daný set nahrávek doplnil tichem do času pěti sekund.

Druhou důležitou úpravou nahrávek bylo nutné převzorkování. Z technologických důvodů EEG (a případné magnetické rezonance) bylo nutné audio data převzorkovat z frekvence 44100 Hz na 8000 Hz. Tento krok, ačkoliv s rizikem malého zkreslení z důvodu převzorkování, byl nezbytným krokem ke správnému chodu experimentu.

[4]Označení originální věty uložené v řečovém korpusu
[5]Označení typu chyby

5.2 Tvorba skriptů pro experiment

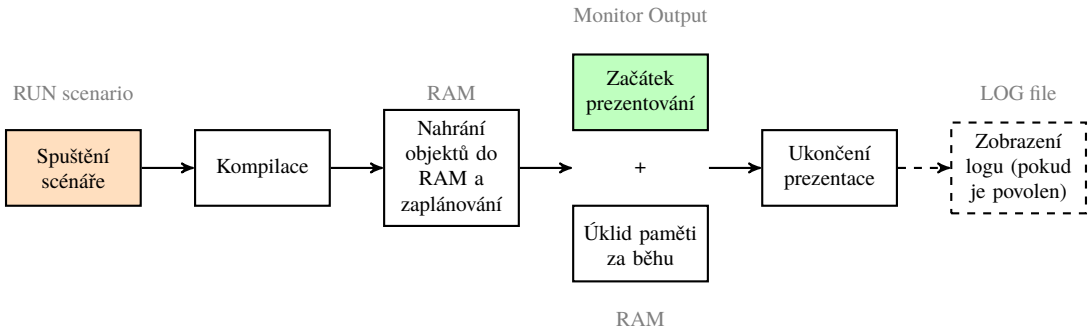
Po dokončení zvukových nahrávek pro měření bylo dalším cílem připravit skripty, které budou řídit průběh experimentu, zaznamenávat důležité události (čas, sledovaný signál, výstupy), čímž získáme potřebné materiály k další analýze zkoumané problematiky.

Skripty pro měření na EEG zařízení byly vytvořeny v Presentation, aplikačním vývojovém prostředí reálného času, běžícím na operačním systému Microsoft Windows, které slouží především pro výzkum neurobiologie a řízení neurobehaviorálních experimentů [1]. Z důvodu požadované časové přesnosti měření je zde kladen velký důraz právě na prostředí reálného času.

Toto prostředí bude využito pro spouštění stimulů, které v našem případě představují zavedené chyby nebo maskovací úsek v připravené resyntetizované řeči. Zároveň s tím bude do EEG zasílána značka (označováno též jako tag či marker) v čase onsetu stimulu. V případě neporušené řeči bude tento tag umístěn do poloviny nahrávky, tedy do času 2500ms. Tato značka přirozené řeči slouží jako „virtuální“ maskovací stimul, tj. stimul, který představuje neporušený zvolený úsek resyntetizované řeči pro pozdější komparaci se stimuly představujícími vnesené chyby. V okolí těchto značek budou při analýze naměřených aktivit na EEG sledovány výkyvy signálu a ověřována navržená hypotéza.

5.2.1 Popis aplikačního prostředí Presentation

Základními kameny experimentů tvořených v prostředí Presentation jsou tzv. **scénáře**. Jedná se o soubory obsahující zdrojový kód, který definuje objekty, stimuly, časování, vstupy a nastavení komunikace s externím zařízením (v našem případě EEG).

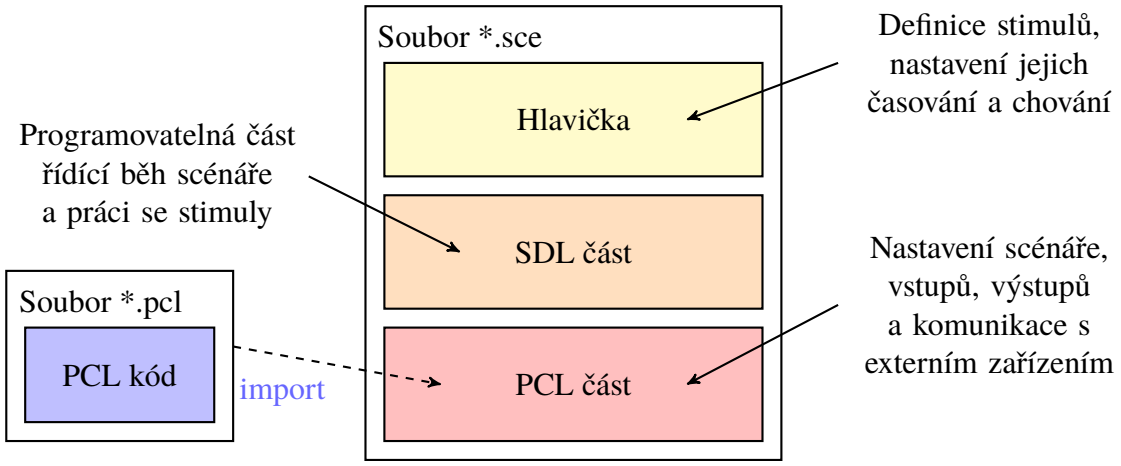


Obrázek 5.1: Princip spouštění a chodu scénářů/experimentů v prostředí Presentation (funkce, které je možné vypnout, jsou zobrazeny čárkovaně)

Jak je zobrazeno na obrázku 5.1, při spuštění scénáře (případně experimentu) nejdříve prostředí zkompiluje zdrojové soubory a ověří, jestli neobsahují kód, který by byl v rozporu se syntaktickými pravidly jazyka. Poté se vytvoří a uloží objekty definované v kódu do paměti RAM, odkud jsou za běhu odstraňovány, pokud jich není nadále potřeba (žádná další reference v budoucím kódu na ně neodkazuje). Jakmile jsou všechny potřebné objekty připravené v paměti, spustí se samotná prezentace dat (stimulů) a probíhá zaznamenávání událostí do výpisového souboru. Ten je možné po ukončení aplikace zobrazit, pokud nebyla tato možnost vypnuta.

Struktura scénářů

Každý scénář je funkčně rozdělen do tří částí (viz obrázek 5.2): hlavička obsahující nastavení dokumentu a experimentu, část obsahující definice stimulů a nakonec programovaná část s funkcionalitou experimentu.



Obrázek 5.2: Jednoduché schéma znázorňující strukturu scénářů v aplikaci Presentation.

Pro vytváření scénářů jsou v Presentation definovány dva jazyky:

SDL (Scenario Description Language) představuje deskriptivní jazyk používaný k definování stimulů a s nimi spojenými vlastnostmi použitých ve scénáři.

PCL (Presentation Control Language) představuje interpretovaný programovací jazyk používaný pro řízení scénáře. PCL část scénáře však není vyžadována a je možné vytvářet scénáře pouze s deskriptivní SDL částí.

Kód v jazyce PCL může být psát v souborech *.pcl separovaných od scénáře, což zpřehledňuje kód. Toho bylo využito i při návrhu experimentu a do těchto souborů byly psány

funkce zpracovávající soubory s daty společně s audio soubory, které se pak pouze volaly v hlavní scénáři experimentu.

5.2.2 Tvorba scénářů

V aplikaci Presentation bylo potřeba připravit scénář, který bude chod experimentu řídit. Kód by měl testovanému subjektu prezentovat připravené nahrávky a v čase onsetu stimulu zasílat na EEG zařízení značku rozlišující typ stimulu, tj. stimul značící chybu nebo maskovací stimul. Po domluvě s odbornými pracovníky NUDZ jsme se rozhodli umisťovat v této části výzkumu pouze značku pro porušenou a neporušenou resyntetizovanou řeč (nebudeme tedy prozatím rozlišovat typ chyby v nahrávce). U porušené řeči bude na EEG při onsetu stimulu značícího chybu umisťována značka s označením S_1, v případě neporušené řeči, která nemá z důvodu absence chyby žádnou informaci o onsetu porušeného místa, bude značka umístěna do poloviny nahrávky, tedy v čase 2500ms, s označením S_2. Umístěné značky budou poté sloužit při analýze naměřeného signálu (viz následující kapitola).

V první řadě byla při tvorbě scénáře připravena hlavička, ve které byly definovány základní údaje, cesty k audio souborům, nastavení výstupního zařízení a vlastností scénáře. Poté byla v SDL části připravena struktura stimulů včetně jejich časování a grafických výstupů na výstupní zařízení. Pro náš experiment byly jako stimuly brány nahrávky resyntetizované řeči společně s údaji o zavedených chybách – zejména o onsetu chyby. V této části skriptu je nejdůležitější vytvoření korektní struktury stimulů, jejich časování. Jejich obsah zde není relevantní, neboť bude měněn v programované PCL části scénáře.

Kvůli velkému množství zvukových dat byl proces načítání a aktivace stimulů naprogramován v PCL části, která dovoluje automatizačními prostředky (cykly) pracovat s objekty. Pro připravené struktury v SDL části byl tedy v jazyce PCL přidán kód, který bude načítat a obsluhovat objekty během aktivního měření na EEG. Současně s tím byla přidána možnost pseudonáhodného rozdělení dat, což může každé měření učinit jedinečným.

V první řadě bylo nezbytné načíst a zpracovat vstupní data uložená v tabulce CSV, která obsahuje všechny potřebné informace k následnému vytvoření jednotlivých stimulů (informace o chybě, umístění nahrávky atp.). Procedura načítání dat z externího souboru je navržena v algoritmu 1 formou pseudokódu.

Algoritmus 1 Čtení připravených dat ze souboru CSV

```
1:  procedure ReadCSV(s)                          ▷ s představuje vstupní soubor
2:      Open(s)
3:      Create(array, Length(s))                  ▷ Vytvoří pole délky s (počet řádků souboru)
4:      while ReadLine(s) ≠ ∅ do
5:          PutInto(readline(s), pole)
6:      end while
7:
8:      Close(s)
9:      return array                              ▷ Vrací pole obsahující informace ze souboru
10: end procedure
```

Po zpracování vstupních dat už zbývá poslední část, a to vytvoření procedury pro prezentaci dat měřenému subjektu – výstup na obrazovku, přehrávání audia a zaznamenávání do logu. Zjednodušený popis vytváření a prezentace stimulů je zobrazen na algoritmu 2.

Algoritmus 2 Prezentování stimulů na výstupu

```
1:  procedure PresentData
2:      data ← ReadCSV(s)
3:      ID ← 0
4:      while ID < Length(data) do
5:          Create(stimul, data[ID])              ▷ Vytvoření stimulu z dat
6:          Set(stimul, data[ID])                 ▷ Nastavení vlastností stimulu dle dat
7:          Load(stimul)                          ▷ Načtení stimulu
8:          Present(stimul)                       ▷ Prezentace stimulu (Monitor + Sluchátka)
9:          Unload(stimul)                        ▷ Úklid z paměti
10:
11:         ID ← ID + 1
12:     end while
13: end procedure
```

Pro experimentální měření byl takto připraven skript, který načte všech 300 připravených zvukových souborů a veškeré informace s nimi související. V intervalu 5 sekund se pak na obrazovku prezentuje text aktuálně přehrávaného zvukového souboru. Současně s tím je jako audio výstup přehrávána relevantní nahrávka. Do záznamu spuštěného měření se ukládají informace o tom, v jakém pořadí se soubory spouštějí, onset, offset a klasifikace dané chyby (pokud se v nahrávce vyskytuje).

Nejdůležitějším faktorem u jednotlivých stimulů je však přesné načasování pulsu, při kterém je do EEG zařízení zasílána značka porušené či neporušené nahrávky. Kdyby byl tento puls zasílán se zpožděním, došlo by ke zkreslení výsledků z důvodu posunu značky z místa onsetu chyby.

Kapitola 6

Průběh a vyhodnocení EEG experimentu

Na připravených datech bylo vykonáno několik měření, jejichž cílem bylo potvrdit zkoumanou hypotézu a zjistit, zda se ve sledovaném neurofyziologickém signálu vyskytuje nějaká nestandardní aktivita poukazující na výskyt chyby ve sledované syntetické řeči. V této kapitole je stručně popsáno, jak tento krátký experiment probíhal a jakých výsledků bylo dosaženo.

6.1 Zkoumané subjekty

Jako dobrovolníci pro měření na EEG byly vybrány náhodné subjekty bez zdravotních potíží, které souhlasily s podmínkami měření. Tyto osoby byly seznámeny pouze se základními principy experimentu a nebylo jim žádným způsobem sděleno, že budou poslouchat syntetickou řeč, navíc poškozenou uměle vnesenými chybami. Měření se účastnilo celkem 6 dobrovolníků, přičemž dva z nich měření ukončili předčasně.

6.2 Průběh experimentu

Zkoumaným subjektům byly nejprve připevněny speciální elektrody na povrch skalpu a připraven software pro zaznamenávání mozkové aktivity, resp. elektrických potenciálů na jednotlivých elektrodách. Celkem bylo subjektu aplikováno 20 skalpových elektrod, 2 elektrody v oblasti processus mastoideus[1] a elektrody pro snímání okulogramu.

V programu Presentation bylo nastaveno výstupní zařízení (sluchátka) a nastavena komunikace s EEG zařízením přes paralelní port. Další důležitou částí experimentu bylo připravení ambientního elementu, aby se měřený subjekt příliš nesoustředil na chyby v přehrávané řeči. Jako vhodným prostředkem k odvedení pozornosti bylo zvoleno sledování

[1] oblast za ušními boltci u spánkové kosti

videa, u kterého nebude zkoumaný subjekt nucen příliš přemýšlet a bude minimalizováno mrkání, které má rušivý potenciál na měřené signály.

Takto připravený experiment byl pak zahájen. Měření každého subjektu trvalo celkem 25 minut (300 vět v délce pěti sekund). Dva dobrovolníci ukončili experiment předčasně z důvodu příliš dlouhého měření.

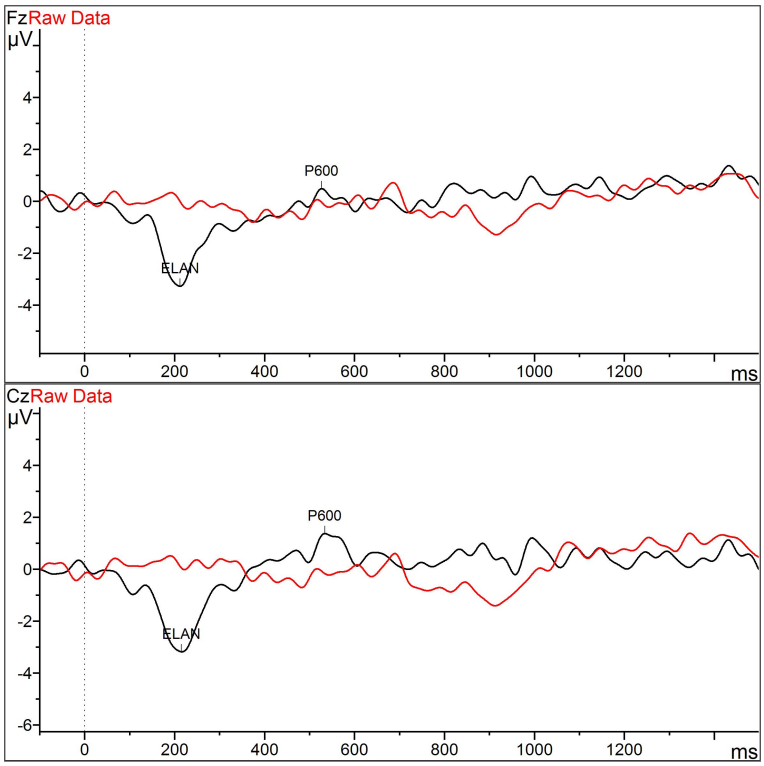
6.3 Vyhodnocení experimentu

Zaznamenaný signál z EEG byl po měření analyzován odbornou pracovnicí Národního ústavu duševního zdraví. Postup analyzování signálu naměřeného na EEG byl následující:

1) Filtrování zaznamenaného signálu s cílem odstranit nežádoucí složky (šum).
2) Eliminování artefaktů v signálu způsobených mrkáním měřeného subjektu. Toho bylo docíleno pomocí *ocular ICA korekce*.
3) Segmentace signálu na úseky -100 až 1500ms, přičemž v bodě 0ms se nachází značka zavedené chyby (případně statická maskovací značka originální věty v čase 2500ms).
4) Zprůměrování segmentovaných úseků do dvou křivek:
    1: Zprůměrovaná EEG aktivita při poslechu vět bez chyby (Original)
    2: Zprůměrované EEG úseky při poslechu vět se zaměněným fonémem (Small/-Big Glitch)

Grand average, neboli zprůměrování všech záznamů pro EEG aktivitu při poslechu originální a porušené věty, jsou zobrazeny na obrázku 6.1. Černá křivka představuje signál při poslechu věty s chybou, červená při poslechu neporušené věty.

U vět s chybou se přibližně v čase 210ms objevuje negativita, která by mohla být považována za tzv. ELAN (*Early Left Anterior Negativity*), což představuje negativní složku evokovaných potenciálů odrážející pochody spojené s automatickou detekcí změny v lingvistickém paradigmatu. V přibližně 600 milisekundách se objevuje pozitivum, označované jako P600, které představuje „syntaktický evokovaný potenciál“, tj. reflektuje reanalýzu věty, která není jednoduše srozumitelná z důvodu nějaké strukturální chyby.



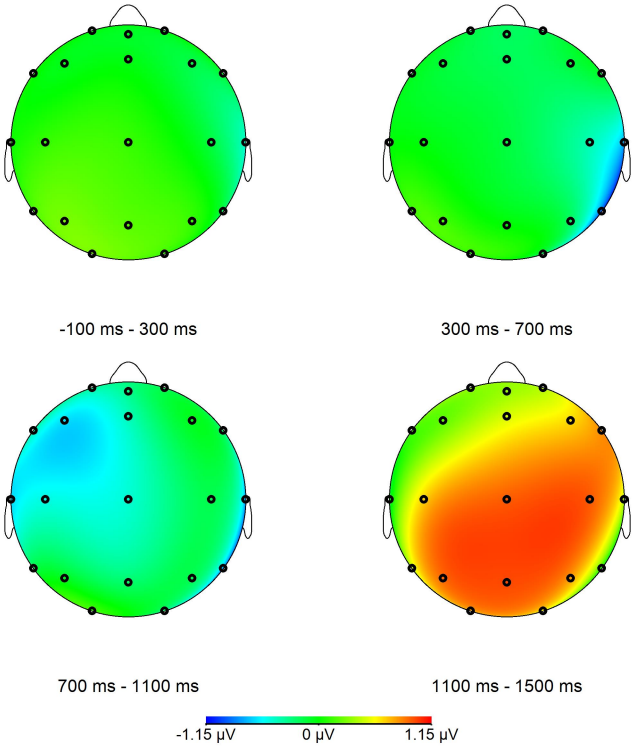
Obrázek 6.1: Grand average zaznemanané EEG aktivity s vyznačením ELAN a P600

Poté bylo zprůměrováno 5, respektive 10 záznamů do tzv. grand average ve formě distribuční mapy pro signál při poslechu originálních (obrázek 6.2) a porušených vět (obrázek 6.3). Modrá barva znázorňuje negativitu fronto-centrální (ELAN/MMN[2]) v čase do 300 milisekund, červená barva představuje pozitivitu centro-parieto až okcipitální (P600/P600-like) v čase 300 - 700 milisekund.

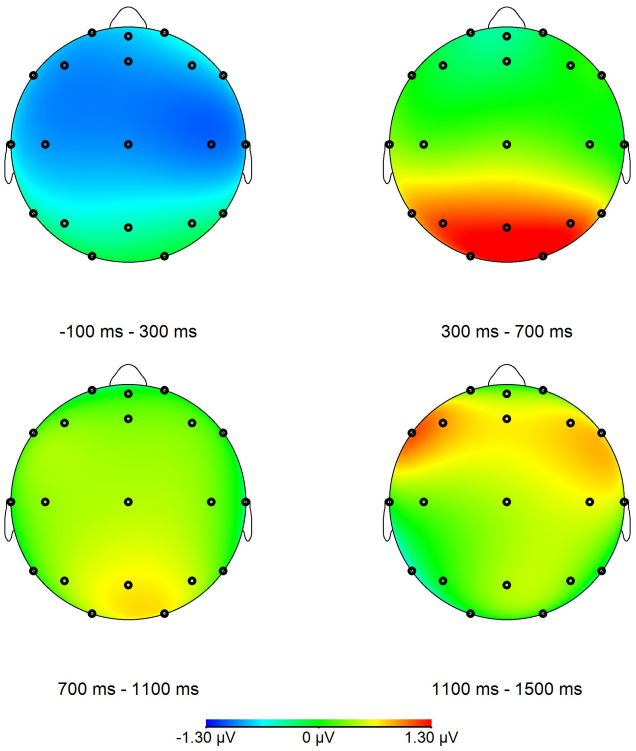
Výsledky měření

Naše hypotéza byla předběžně potvrzena, neboť se u většiny porušených vět objevují nestandardní změny mozkové aktivity popsané výše, zatímco u neporušených vět naměřeny nebyly. Výsledky těchto zkoumání se pak mohou stát základem nové metody detekce chyb v syntetické řeči a nové metody hodnocení kvality syntetické řeči s využitím EEG.

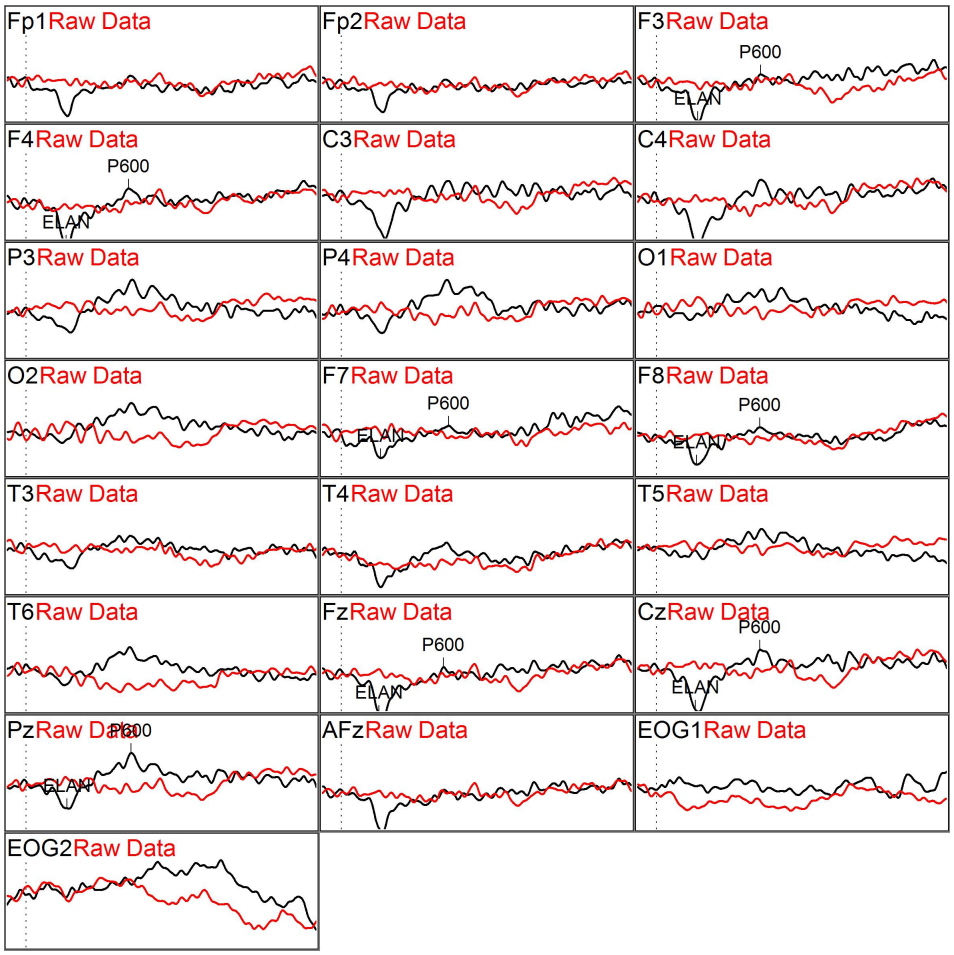
[2]Mismatch negativita (MMN) je složka podobná ELAN, jen není omezena pouze na lingvistické paradigma.



Obrázek 6.2: Mapa distribuce potenciálu zaznamenané EEG aktivity při poslechu neporušených nahrávek



Obrázek 6.3: Mapa distribuce potenciálu zaznamenané EEG aktivity při poslechu vět s uměle zavedenou chybou



Obrázek 6.4: Grand average naměřených signálů z jednotlivých elektrod. Černá barva značí signál porušené řeči, červená naopak signál řeči přirozené. Čárkovaná čára pak značí onset stimulu S_1, resp. S_2

Kapitola 7

Závěr

V této práci byl představen návrh základu nové metody pro hodnocení kvality syntetické řeči sledováním neurofyziologických signálů naměřených na EEG. Hlavním cílem práce bylo připravit data k experimentálnímu ověření hypotézy nestandardní změny naměřeného signálu mozkové aktivity v místech výskytu chyb v syntetické řeči. V teoretickém úvodu práce byl čtenář seznámen se základními poznatky z oblasti tvorby umělé řeči, zejména pak s konkatenační syntézou řeči výběrem jednotek. Poté následovala kapitola objasňující možné typy chyb vznikajících v konkatenační syntéze řeči s detailnějším popisem příčiny jejich vzniku. Dále byly čtenáři představeny běžně používané prostředky pro hodnocení kvality syntetické řeči formou poslechových testů a přiblížena experimentální metoda EEG.

Po teoretickém úvodu byl čtenáři představen návrh experimentálního protokolu a podrobněji popsán cíl této práce. V této části byly klasifikovány typy chyb, které byly v praktické části úlohy zaváděny do resyntetizované řeči pro účely experimentu. Navržená klasifikace byla na konci této kapitoly empiricky ověřena pomocí malého poslechového testu.

Po návrhu experimentu byl čtenáři podrobněji popsán postup při přípravě experimentálních dat a skriptů pro měření na EEG, přičemž bylo přiblíženo aplikační prostředí Presentation, které bylo k přípravě skriptů použito. V poslední kapitole byl popsán průběh a realizace samotného měření na EEG. Nakonec byl přiblížen postup při analyzování naměřených aktivit na EEG s popisem jevů vyskytujících se v signálu. Výsledky experimentu jsou shrnuty níže.

7.1 Vyhodnocení výsledků experimentu

Navrhovaná hypotéza byla při analýze získaných signálů z EEG potvrzena (viz 6.3). Analýza poukazuje na změnu aktivit mozku v okolí onsetu zavedených chyb (obrázek 6.1).

Tato skutečnost potvrzuje, že při poslechu syntetické řeči obsahující chyby člověk podvědomě registruje nepřirozenost řeči, kterou poslouchá. Tyto výsledky se tak mohou stát základem tvorby nové metody hodnocení kvality řeči, případně metody detekce chyb v syntetické řeči.

7.2 Návrh na pokračování výzkumu

Ačkoliv považuji realizovaný experiment za úspěšný, mou snahou v budoucnu by bylo využít EEG, případně magnetickou rezonanci, pro hlubší výzkum detekce a analýzy chyb v syntetické řeči. Za zvážení stojí také úprava paradigmatu pro měření. Bude potřeba prozkoumat, zda bude signál při poslechu neporušené nahrávky vykazovat stejné chování, bude-li nahrávka zařazena v měření vícekrát na různých místech. Na základě žádostí dobrovolníků podstupujících měření bude třeba také zvážit snížení rozsahu experimentálních dat, neboť byl experiment ve většině případů označen za příliš dlouhý.

Literatura

[1] NEUROBEHAVIORAL SYSTEMS, INC. *Neurobehavioral Systems* [online]. 2015 [cit. 2015-01-14]. Dostupné z: `http://www.neurobs.com/menu_presentation/menu_features/features_list`

[2] MATOUŠEK, Jindřich, ROMPORTL, Jan, TIHELKA, Daniel, TYCHTL, Zbyněk. *Recent Improvements on ARTIC: Czech Text-to-Speech System.* In: Interspeech 2004 - ICSLP, proceedings of the 8th International Conference on Spoken Language Processing. Korea: Jeju Island, 2004, s. 1933-1936.

[3] PSUTKA, Josef et al. *Mluvíme s počítačem česky*. Vyd. 1. Praha: Academia, 2006, 746 s. Česká matice technická, roč. 111, č. spisu 502. ISBN 8020013091.

[4] VÍT, Jakub. *Automatická detekce a vizualizace chyb konkatenační syntézy řeči*. Plzeň, 2013. 45 s. Diplomová práce. Západočeská univerzita. Fakulta aplikovaných věd. Vedoucí práce Doc. Ing. Jindrich Matoušek, Ph.D.

[5] MATOUŠEK, Jindřich, TIHELKA, Daniel. *The Design of Czech Language Formal Listening Tests for the Evaluation of TTS Systems*. In: LREC 2004, proceedings of 4th International Conference on Language Resources and Evaluation. Lisabon: European Language Resources Association, 2004. s 2099 - 2012.

[6] ARNDT, Sebastian et al. *Subjective Quality Ratings and Physiological Correlates of Synthesized Speech*. In: Fifth International Workshop on Quality of Multimedia Experience. Rakousko: Klagenfurt am Wörthersee, 2013. s. 152 - 157.

[7] ROMPORTL, Jan. *Zvyšování přirozenosti strojově vytvářené řeči v oblasti suprasegmentálních zvukových jevů*. Plzeň, 2008. 156 s. Disertační práce. Západočeská univerzita. Fakulta aplikovaných věd. Školitel Prof. Ing. Josef Psutka, CSc.

[8] TIHELKA, Daniel. *The Unit Selection Approach in Czech TTS Synthesis*. Plzeň, 2005. 103 s. Disertační práce. Západočeská univerzita, Fakulta aplikovaných věd. Školitel Prof. Ing. Josef Psutka, CSc.

[9] HUNT, Andrew, BLACK, Alan W. *Unit Selection in Concatenative Speech Synthesis System Using a Large Speech Database*. In: Proc. ICASSP. USA: Atlanta, 1996. s. 373-376.

[10] TIHELKA, Daniel, MATOUŠEK, Jindřich, KALA, Jiří. *Quality Deterioration Factors in Unit Selection Speech Synthesis*. In: Text, Speech and Dialogue, proceedings of the 10th International Conference TSD 2007. Německo: Berlín, Springer-Verlag, Hiedelberg, 2007. s. 508-515.

[11] LU, Heng, WEI, Si, DAI, Lirong a WANG, Ren-Hua. *Automatic error detection for unit selection speech synthesis using log likelihood ratio based SVM classifier*. In: Proc. Interspeech. Japonsko: Makuhari, 2010. s 162 – 165.

[12] DUBĚDA, Tomáš, HORÁK, Petr, VÍCH, Robert. *History of Speech Synthesis in the Czech Lands.* In: Proc. ESSP. Česká republika: Praha, 2005. s. 364-317.

[13] INTERNATIONAL TELECOMMUNICATION UNION. ITU-T Recommendation P.85, *A method for subjective performance assessment of the quelity of speech voice output devices*. Švýcarsko: Ženeva, červen 1994.

Příloha A

Seznam vět poslechového testu

Seznam vět poslechového testu

1. Rychlý vzestup Telefóniky mnohé překvapil
2. Jinak jako by člověk žvýkal chutnou dřevotřísku
3. To je otázka inkvizice a křížových výprav
4. Vladimír Růžička znovu jednou nadchl
5. Uvedl to primátor Plzně Jiří Šneberger
6. Nejlepší výkon předvedl Dvořák v dálce
7. Podle odborníků jim totiž hrozí leukemie
8. Jméno nového šéfa banky však nechtěl říci
9. Deseti mrtvým záchranářům už nikdo nepomůže
10. S fanoušky v zádech býváme těžko k poražení
11. Prodej automobilů trhá každý měsíc rekordy
12. Německá inflace stlačí cenu dluhopisů
13. Těch pár sýčků patří k místnímu koloritu
14. Vytvořím širší bázi pro náš vztah k Moskvě
15. Zároveň ovšem Mlynářův optimismus přibrzdil
16. V loňském roce zvítězil Kaučuk Kralupy
17. Tyto důvody však označil Giňa za zástupné
18. Proti tomu však mluví různá svědectví
19. Starší z nich byl dokonce soudním tlumočníkem
20. Mé knihy vyšly v bezmála čtyřiceti jazycích

Příloha B

Obsah přílohy na CD

Obsah CD

- Poslechový minitest
  - Originální nahrávky
  - Porušené nahrávky (obsahují malou a velkou chybu)
  - Tabulka s informacemi o souborech (XLS)
- Data pro experiment
  - Zkrácený ASF soubor řečového korpusu a anotační soubor (TXT)
  - Skript pro výběr a úpravu nahrávek (EXE)
  - Připravené nahrávky pro experiment
    - Soubor 300 nahrávek připravených pro experiment (WAV)
    - Soubory s informacemi o jednotlivých nahrávkách formou jednoduché tabulky (CSV)
- Scénáře a experiment programu Presentation
  - Soubor experimentu s nastavením (EXP)
  - Scénáře pro experiment (SCE)
  - Funkce definované pro použité scénáře (PCL)

Příloha C

Seznam obrázků, tabulek a algoritmů

Seznam obrázků

Seznam tabulek

Seznam algoritmů