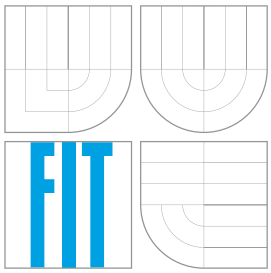
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



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

KONVERZE HLASU
VOICE CONVERSION

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

DAVID HODAŇ

VEDOUCÍ PRÁCE
SUPERVISOR

Doc. Dr. Ing. JAN ČERNOCKÝ

BRNO 2016

Abstrakt

Práce se zabývá problematikou konverze hlasu, což je transformace parametrů řeči jednoho řečníka tak, aby zněl jako někdo jiný. Je proveden rozbor metod odrážejících současný stav technik konverze. V teoretické části je nejprve přiblížen způsob tvorby řeči s důrazem na atributy identifikující a charakterizující hlas. Jsou popsány metody modifikace hlasu s jejich výhodami a úskalími, jež předurčují oblast použití daných metod. Dále jsou probrány způsoby transformace hlasu mezi zdrojovým a cílovým mluvčím. Na základě popsaných poznatků je vytvořen software demonstrující jednu z cest jak tohoto cíle dosáhnout. Konverze je rozdělena z pohledu trénování a syntézy. Součástí práce je program konverze hlasu, který byl vytvořen v programovém prostření MATLAB. Postupně je v práci popsán jeho návrh, implementace a zhodnocení dosažených výsledků.

Abstract

Voice conversion is the process of transformation of speech parameters belonging to one speaker in such a way that his/her speech sounds as spoken by someone else. This thesis presents a short summary of several techniques currently used for conversion. First, the theory of voice creation with an emphasis on key atributes that characterize and identify a speaker's voice is described. Methods for voice modification are discussed, together with the advantages and pitfalls that predetermine the use-cases for suitable application of these methods. A high-level overview of how speech is transformed between the source and the target speakers is presented. This description is subsequently used to design a voice conversion system that is aimed to demonstrate one of the possible approaches to the conversion problem. The process of conversion consists of two phases: training and synthesis. As part of this project, a computer program for voice conversion based on the MATLAB programming environment has been developed. Its design, implementation and results are discussed.

Klíčová slova

zpracování řeči, konverze hlasu, syntéza hlasu, fonémový rozpoznávač, LPC, DTW

Keywords

speech processing, voice conversion, voice synthesis, phoneme classifier, LPC, DTW

Citace

HODAŇ, David. *Konverze hlasu.* Brno, 2016. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Černocký Jan.

Konverze hlasu

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Jana Černockého. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

.....
David Hodaň
18. května 2016

Obsah

Seznam obrázků

Kapitola 1

Úvod

Konverze hlasu je proces adaptace charakteristik řeči jednoho řečníka takovým způsobem, aby jeho hlas zněl jako hlas jiného člověka – cílového řečníka. Ačkoliv jsou lidský hlas a lidská řeč dvě rozdílné kategorie, často v technické praxi dochází k překrývání těchto pojmů, např. rozpoznáváním řeči a rozpoznáváním hlasu se většinou myslí tentýž proces. Podobně hovoříme rovněž o konverzi hlasu nebo konverzi řeči.

Lidský hlas je technický prostředek, nosič zvuku, zatímco lidská řeč je způsob použití tohoto prostředku sloužící ke komunikaci. Avšak tvrzení, že pouze řeč nese informace, tedy myšlenky vtělené ve slova a věty, by bylo mylné. I hlas jako takový přenáší informace ve své aktuální kvalitativní podobě. Hlas bývá emočně zabarvený, posluchač snadno rozezná např. hlas úzkostlivý, rozzlobený, odevzdaný, netrpělivý, ironický, smutný apod. Roli zde hraje tón, rychlost, melodie, rytmus atd.

Systémy rozpoznávání a syntézy řeči jsou v současné době již komerčně dostupné běžnému spotřebiteli. V zařízeních, jako jsou mobilní telefony, se uplatňují rozpoznávače řeči. Ovládání hlasem se nabízí jako alternativa k dálkovému ovládání televizí nebo setů handsfree. Čtečky elektronických knih již běžně předčítají text poměrně věrnými hlasy. Dá se říci, že současné systémy zvládají již velmi dobře syntetizovat řeč, avšak ještě nedokáží věrně napodobit emočně zabarvený hlas [19]. Konverze hlasu je jednou z cest, jak syntetizovanou řeč ještě více přiblížit posluchačům.

Již dnes jsou patrné trendy budoucího rozvoje v této oblasti [25]. Spotřební elektronika by mohla poskytovat nastavení hlasu majitele zařízení, členů jeho rodiny, nebo známých herců. V počítačových hrách se může uplatnit možnost dokonaleji personifikovat herní postavu s hráčem. Lidem, kteří přišli z nějakého důvodu o hlas, by bylo možné nejen vrátit schopnost opět se dorozumět syntetizovanou řečí, což již dnes možné je, ale vrátit jim i jejich vlastní hlas. Ve filmovém průmyslu při dabingu by se mohla uplatnit transformace hlasu mezi jazyky. Jde o obtížný úkol, jehož zvládnutí dosud nebylo vyřešeno.

V oblasti ochrany osobních dat v bankovnictví s sebou rozvoj konverze hlasu přináší nové výzvy. Bude třeba podnikat kroky pro lepší zajištění dat a majetku proti prolomení hlasového zámku. V podvrhování řeči (tzv. spoofing) a obraně před ním již probíhá výzkum [29].

Tématem této práce je konverze řeči; jde o podmnožinu spadající do rozsáhlejší kategorie obecné transformace řeči, kde se řeší i transformace řeči umělého původu nebo převod psaného textu na řeč (tzv. syntéza text-to-speech). U konverze řeči se nemění obsah toho, co chtěl mluvčí sdělit, mění se pouze hlas promluvy. Tato práce ukáže, jakým způsobem lze konverze dosáhnout. Postup budu demonstrovat na programu, který jsem vytvořil v prostředí MATLAB. Jeho vstupem jsou trénovací promluvy, zdrojová a cílová, a dále nahrávka,

která má být konvertována. Výstupem je pak transformovaná nahrávka.

Kapitola 2 Hlas a jeho vznik se věnuje základní charakteristice hlasu s důrazem na ty vlastnosti, které mají bezprostřední vliv na zpracování hlasu technickými prostředky. V kapitole 3 Od modelu k filtru jsou přiblíženy teoretické základy pro řešení problematiky pomocí metod digitálního zpracování signálů. Řečové ústrojí je možné modelovat jako filtr, což je demonstrováno na sérii schémat a podloženo matematickým základem. V kapitole 4 jsou popsány metody vedoucí k realizaci systému pro konverzi. Je zmíněn současný stav techniky v tomto oboru. Kapitola 5 se věnuje konkrétnímu řešení, jež bylo realizováno v podobě počítačového programu. Bližší rozbor jeho implementace a popis jednotlivých použitých metod se nachází v kapitole 6. Následné vyhodnocení úspěšnosti v dosažení vytyčeného cíle je provedeno v kapitole 7. Jsou zde popsány experimenty, jež vedly k posuzování správnosti řešení v průběhu realizace. Podobnost výsledné zkonvertované nahrávky s originálem cílového mluvčího je demonstrována na sérii grafů. V závěru je naznačena cesta možného budoucího vývoje.

Kapitola 2

Hlas a jeho vznik

Lidský hlas může být nosičem mluveného slova nebo zpěvu. Hlas vypovídá o fyziologických vlastnostech jedince, o jeho fyzickém i duševním stavu, o sociálním původu nebo i o příslušnosti k zeměpisnému regionu apod. Vzniká v orgánech, které jsou součástí dýchacího ústrojí lidského těla.

Existují metody, jež nám umožňují toto ústrojí matematicky popsat a modelovat. Pomocí elektronických digitálních zařízení jsme potom schopni hlasové ústrojí velmi dobře simulovat.

2.1 Anatomie hlasového ústrojí

Dýchací ústrojí je složeno z dolních a horních cest dýchacích. Na předělu mezi těmito dvěma částmi jsou hlasivky. Dýchací ústrojí dobře reprezentuje obrázek 2.1 nazvaný Fyzikálně akustické schéma dýchacího systému [5].

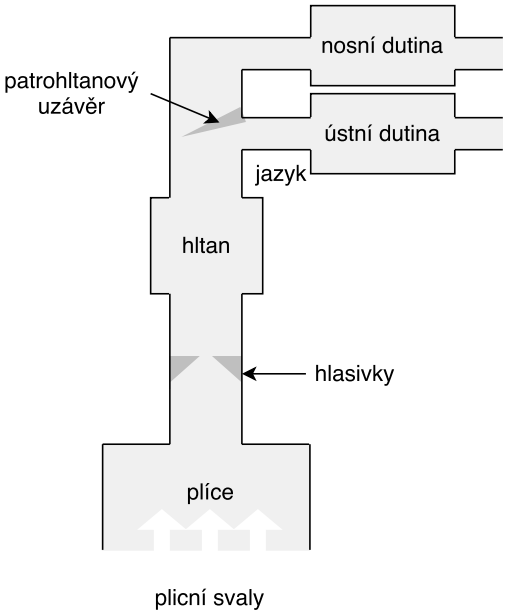
Pro vznik zvuku jsou v dolních cestách dýchacích nejdůležitější plíce s plicními svaly, které si můžeme představit jako analogii napájecího zdroje hlasového ústrojí. Produkují proud vzduchu, jehož proměnná energie je následně zpracovávána v navazujících orgánech podílejících se na tvorbě hlasu.

O části ústrojí od hlasivek nahoru hovoříme jako o artikulačním ústrojí. Jde o hrtan, hltan, dutinu ústní, dutinu nosní a nosohltan. Významnou úlohu sehrávají též zuby a rty.

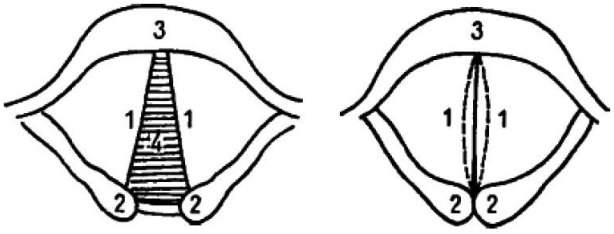
2.1.1 Hlasivky

Hlasivky jsou zobrazeny na obrázku 2.2 z pohledu, jenž se získá např. lékařským přístrojem, laryngoskopem. Jsou tvořeny dvěma pružnými vazy pokrytými sliznicí, napnutými mezi chrupavkami. Muži mají v průměru délku hlasivek 22 mm, ženy 18 mm.

Hlasivky tvoří úzké místo na cestě vzduchu z plic. Při dýchání jsou hlasivkové vazy ochablé, aby vzduch mohl volně proudit. Vzduch prochází trojúhelníkovou štěrbinou zvanou glottis. Při mluvení a zpěvu se zapojí drobné hrtanové svaly. Hlasivky se napnou a hlasová štěrbina se zúží. Tím vznikne překážka proudícímu vzduchu a hlasivky se rozvibrují podobně jako u dechového hudebního nástroje, jakým je např. hoboj. Výsledkem vibrací je zvuk, který se dále šíří vzhůru do artikulačního ústrojí. Tento zvuk – tzv. hrtanový hlas – nemá lidské zbarvení [8].



Obrázek 2.1: Fyzikálně akustické schéma dýchacího systému.



Obrázek 2.2: Schéma hlasivek. Na levém obrázku jsou znázorněny hlasivky (pod číslem 1) při dýchání, hlasivková štěrbina (4) je otevřená. Číslo 2 a 3 značí hlasivkové chrupavky a štítnou chrupavku. Na pravém obrázku je stejný pohled na hlasivky, tentokrát při mluvení, kdy je štěrbina mezi hlasivkami zúžená. Obrázek byl převzat z publikace Fonetika a fonologie [8], původním autorem je Gerhart Linder [9].

2.1.2 Artikulační ústrojí

To, jak primární zvuk z hlasivek vypadá a jak je zpracován artikulačním traktem, je možné interpretovat pomocí Helmholtzovy teorie tvorby samohlásek [20].

Hermann von Helmholtz popsal hrtanový hlas jako složený tón se základní harmonickou složkou a s vyššími harmonickými složkami zastoupenými ve svém spektru. Artikulační ústrojí představuje kaskádu rezonančních dutin s různými rezonančními frekvencemi. Při průchodu jednotlivými dutinami se zvýrazní harmonické složky s frekvencí podobnou rezonanční frekvenci a ostatní se naproti tomu potlačí. Tvar a objem rezonátorů se při mluvení mění. Změny ve vzájemné poloze jazyka, rtů a zubů se projeví změnou charakteru hlasu, zejména jeho barvy.

Z pohledu modelování tvorby řeči lze na hlasové ústrojí pohlížet jako na zdroj a filtr [4]. Hlasivky jsou zdrojem prvotního hlasu a rezonanční dutiny představují filtr modifikující hlas zdroje. Zvuková intenzita vyšších harmonických složek se vzrůstající frekvencí klesá

přibližně o 6 dB na oktávu.

2.2 Základní frekvence

Základní frekvence hrtanového zdrojového hlasu vycházejícího z hlasivek se označuje jako F_0 (říká se jí „pitch“). Výška hlasu závisí na velikosti hlasových orgánů - hlasivek, věku, zdravotním stavu a pohlaví jedince. Základní frekvence se pohybuje u mužů v rozmezí asi 80 až 200 Hz, u žen v rozmezí 150 až 350 Hz a u dětí v rozmezí 200 až 500 Hz.

2.3 Formanty

Jednotlivé rezonanční frekvence vokálního traktu se nazývají formanty. Rezonance ústní dutiny produkuje tzv. hlavní formant F1. Ten může člověk měnit v poměrně širokém rozsahu asi od 175 Hz (tón F) do 3700 Hz (tón B_4), a to polohou jazyka vůči zubům a měkkému patru. Na tvorbě hlavního formantu se též podílejí rty. Neměnná hrtanová dutina má rezonanci s frekvencí asi 400 Hz (tón G_1). Ta tvoří spolu s nosohltanovou dutinou tzv. vedlejší formant F2.

2.4 Samohlásky

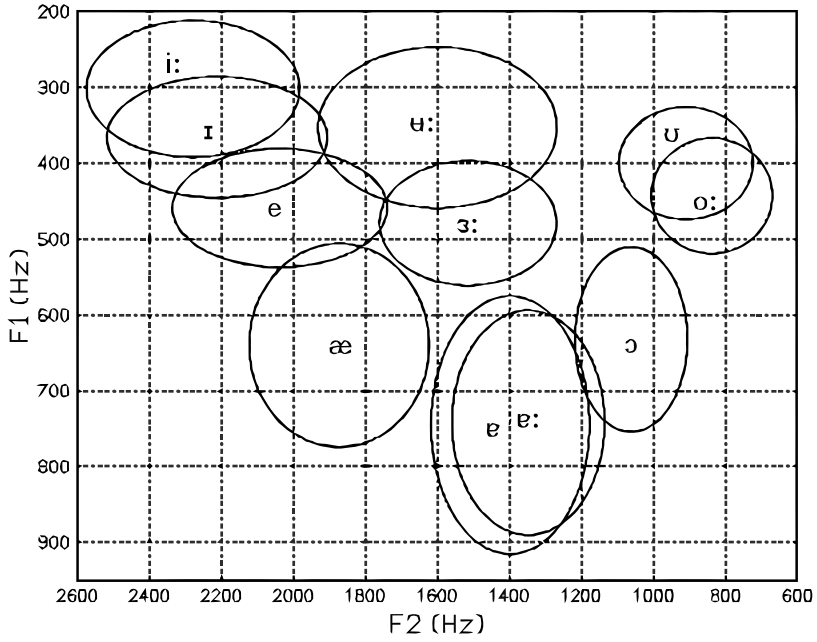
První dva formanty jsou určujícími pro samohlásky (v češtině *a, e, i, o, u*), zatímco vyšší formanty dodávají hlasu individuální zabarvení. Samohlásky mají nejužší vazbu na základní hrtanový hlas. České samohlásky jsou znělé a ústní. Protikladem k významu ústní jsou samohlásky nosové, např. v polštině a francouzštině (na jejich tvorbě se podílí též dutina nosní).

Jestliže se u samohlásek zvýší základní frekvence, protože např. zpěvák přešel z hlubokých tónů na vysoké, dojde k následujícímu jevu: protože vyšší harmonické složky zvuku jsou násobkem základní frekvence, zvyšováním se od sebe harmonické složky rychle oddalují, nařeďují se a nekryjí se dobře s formanty, samohlásky přestávají být správně akusticky definovány. To je důvodem, proč u vysokých hlasů znějí samohlásky podobně [6].

Graf na obrázku 2.3 byl převzat z práce Vowel Perception in Australian English od Roberta Mannela z university v Sydney [12]. Je na něm vokální mapa základních samohlásek australské angličtiny vyslovovaných dospělými muži. Na vodorovné ose jsou příslušné formanty F2, na svislé formanty F1. Z grafu je patrno, jak se liší frekvence formantů pro jednotlivé samohlásky. Čeština oproti angličtině má menší počet samohlásek.

2.5 Souhlásky

Souhlásky vznikají tak, že vzduch proudí přes různá zúžení podél artikulačního ústrojí. Mohou být znělé a neznělé. Při tvorbě znělých souhlásek se uplatní kmitání hlasivek. Jinak jsou souhlásky kromě periodických zvuků charakterizovány též neperiodickými šumy. Zvláštní kategorií jsou pak sykavky, na jejichž tvorbě se hrtanový hlas vůbec nepodílí. Jsou tvořeny šumem vznikajícím v úžině mezi jazykem a dásňovým obloukem.



Obrázek 2.3: Vokální mapa základních samohlásek australské angličtiny

2.6 Fonémy

Foném je nejmenší stavební jednotka zvuku řeči v konkrétním jazyce. Každý jazyk je vybaven vlastní specifickou sadou fonémů. Jejich množství se liší u různých jazyků a také dle použitého modelu.

Fonémová sada českého jazyka je tvořena 39 fonémy, z toho je 13 vokalických (samohláskových) a 26 konsonantických (souhláskových). Na přesnou definici toho, co je a není samostatný foném, nepanuje jednotný názor. Někteří odborníci zpochybňují některé z fonémů, např. /*ó*/, neboť ten má význam jen ve vztahu k cizím slovům.

Každá hláska není fonémem. Fonémem je jen ta, která je schopná rozlišit význam zvuku. V některých jazycích mohou mít dvě hlásky – odlišné zvuky – stejnou funkci, v jiných jazycích nikoliv, záleží na tom, tvoří-li stejný foném. Záměnou fonému se může změnit význam slova (např. pes — pas). Jeden foném může mít více zvukových realizací, tzv. alofony nebo též alofonní varianty fonému. Např. český foném /n/ je realizován dvěma alofonními variantami, kdy na konci slova *mlýn* zní jinak než uprostřed slova *branka*. Zde také velmi záleží na tom, kdo vyslovuje. Rozborem fonémů mluvčího lze usuzovat na celou řadu osobnostních charakteristik, např. na jeho sociální původ nebo postavení.

Vlastností fonémů je, že na základě distinktivních rysů vytvářejí tzv. fonologické protiklady, což jsou zvukové rozdíly umožňující rozlišení významu. Distinktivní rysy zahrnují vlastnosti vokalické, konsonantické a prozodické.

Vokalické vlastnosti (otevřenost, timbre, rezonance) utvářejí fonémy charakteristické nepřítomností překážky při vyslovování.

Konsonantické vlastnosti (místo artikulace, znělost, napjatost, přídech, rekurze, mlaskavost, zdvojenost) utvářejí fonémy charakteristické vytvořením artikulační překážky a jejím odstraněním. Prozodické vlastnosti (kvantita, přízvuk, intonace) samy o sobě nevytvářejí fonémy, ale jsou na fonémy vázány.

V českém jazyce jsou nejčastějšími fonémy krátké samohlásky *e, o, a, i*, neznělé souhlásky *t, s, k* a sonantické fonémy *n, l, m, r* (seřazeno vždy dle četnosti výskytu). Ze znělých souhlásek se uplatňují zvláště *v, d, z, b*. Dále jsou častými dlouhé samohláskové fonémy *í, á*. Slabika má v českém jazyce délku 1 až 6 fonémů.

Sada fonémů fonémového rozpoznávače vyvinutého skupinou BUT Speech@FIT[1], jehož je využito v této práci, zahrnuje v českém jazyce 45 fonémů. Avšak ne všechny z nich jsou klasickými fonémy, některé jsou jen pomocné a slouží k detekcím nehlasových anomálií.

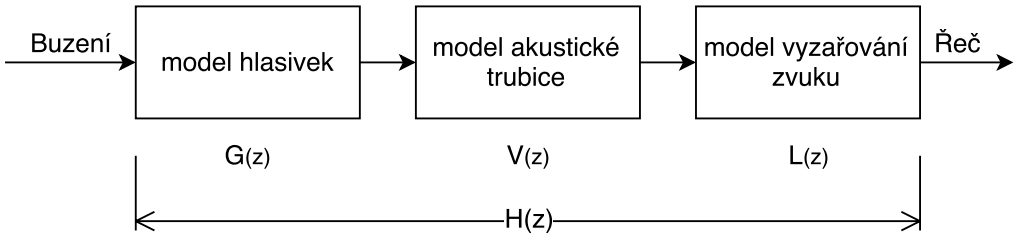
Sady fonémů pro jednotlivé jazyky jsou různé. Případný převod řeči z jednoho jazyka do druhého představuje problém.

[1] http://speech.fit.vutbr.cz/cs

Kapitola 3

Od modelu k filtru

Za účelem rozborů a modelování tvorby řeči si lze celý hlasový trakt včetně hlasivek a rtů představit jako kaskádu tří dílčích modelů, jak je znázorněno na obrázku 3.1.



Obrázek 3.1: Blokový model hlasového traktu

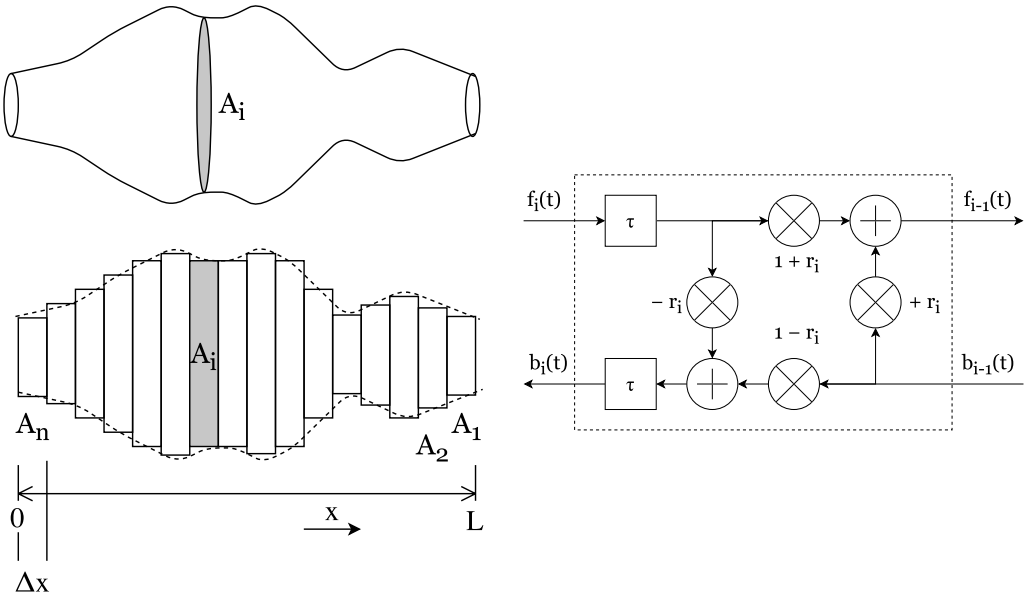
Buzením je proud vzduchu z plic o určité energii. Model hlasivek je dolní propustí 2. řádu, model vyzařování zvuku je horní propustí. Prostřední částí je model akustické trubice, jehož rozboru bude věnována větší pozornost v následující sekci. Jde o kaskádu dvoupólových rezonátorů. V rovině z lze stanovit celkovou přenosovou funkci modelu $H(z)$ jako součin přenosových funkcí:

$$H(z) = G(z) \cdot V(z) \cdot L(z) \tag{3.1}$$

3.1 Model akustické trubice hlasového traktu

Z tohoto rozdělení na zdroj (buzení) a filtr vychází celá řada digitálních i analogových modelů [22]. Akustická trubice je často modelována pomocí sekvence dutin zjednodušeně válcového tvaru. Jedním ze základních teorií vycházejících ze systému trubic je model Kellyho a Lochbauma [28] z šedesátých let dvacátého století. Jedná se o převedení v té době populárních analogových válcových modelů do podoby digitálního filtru.

V dalších úvahách bude navázáno na obrázek 2.1, kde lze provést na modelu zjednodušení a uvažovat zavřený patrohltanový uzávěr. Tím vzniká dutina na obrázku 3.2. Spojitou dutinu je možné aproximovat soustavou bezeztrátových trubic s příčnou plochou A_1 až A_n. U Kellyho a Lochbauma mají segmenty stejnou délku: Δx. V dílčích elementech trubice jsou rychlost šíření zvuku a tlak jednotlivými proměnnými v parciálních diferenciálních rovnicích. Tyto rovnice vedou na popis i-tého segmentu soustavy trubic, přičemž existují hraniční podmínky na přechodech mezi segmenty.



Obrázek 3.2: Na levém obrázku je aproximace akustické trubice pomocí válcových segmentů. Vpravo je model průchodů a odrazů vlny v jednom segmentu [28].

Počet těchto přechodů se rovná řádu systému digitálního filtru, který trubici odpovídá. Šíření dopředných a zpětných zvukových vln řeší Kellyho-Lochbaumovy rovnice [28]. Dopředná vlna $f_{i-1}(t)$ vstupující do segmentu se sčítá se zpětnou vlnou $b_{i-1}(t)$ částečně odraženou od rozhraní mezi segmenty.

Část filtru odpovídající jednomu trubicovému elementu je znázorněna na obrázku 3.2 vpravo. Filtr popisující celý model trubice hlasového traktu bude potom sestaven zřetězením dílčích struktur jednotlivých segmentů za sebe. Výsledná soustava zahrnující též model hlasivek a model vyzařování vznikne přidáním degenerovaných struktur zleva pro hlasivky a zprava pro vyzařování, což odpovídá obecnému schématu na obrázku 3.1.

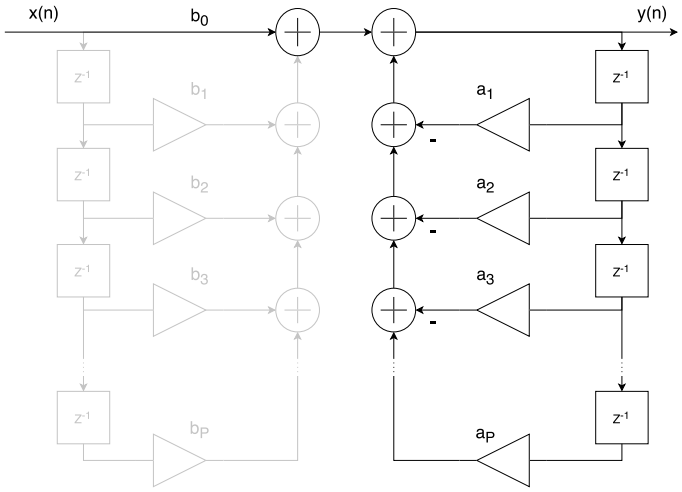
Analogové modely artikulačního ústrojí postupně ustoupily spektrálním modelům – ať už neparametrickým jako vocoder, nebo parametrickým modelům založeným na teorii zdroje a filtru, jako je populární metoda Linear Predictive Coding. Spektrální modely se prosadily v kódování telefonních přenosů, kde se uplatnila jejich nižší náročnost na množství přenášených dat za zachování poměrně vysokého stupně věrnosti. Linear Predictive Coding (LPC) je stěžejní metodou využívanou v této práci, proto bude detailně představena v následující sekci.

3.2 Linear Predictive Coding

LPC (česky Lineární prediktivní kódování) je velmi často používaná metoda, která popisuje model akustické trubice pomocí celopólového čistého IIR filtru [1], viz obrázek 3.3.

Východiskem metody je lineární predikce. Jde o matematickou operaci, která staví na tom, že ve kvazistacionárním časovém úseku lze odhadnout budoucí vzorek časového signálu jako lineární kombinaci určitého počtu vzorků minulých.

Na vstupu nemusí být jen základní hrtanový hlas z hlasivek, bývá sem též přidáván zdroj šumového signálu pro generování složitějšího zvuku lépe odpovídajícího reálnému hlasovému traktu, avšak zde v popisovaném modelu tento zdroj naznačen není.



Obrázek 3.3: Schéma obecného IIR filtru

Celková přenosová funkce $H(z)$ pro IIR filtr bude ve tvaru

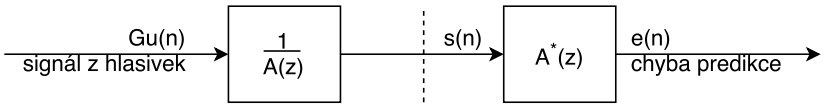
$$H(z) = \frac{1}{A(z)} \tag{3.2}$$

Snahou bude určit koeficienty funkce

$$A(z) = 1 + a_1 z^{-1} + a_2 z^{-2} + a_3 z^{-3} + \cdots + a_P z^{-P} \tag{3.3}$$

ve jmenovateli přenosu $H(z)$. Metodou, jak koeficienty identifikovat, je následující postup.

Prvním krokem identifikace koeficientů je zavedení inverzního filtru do přenosu $1/A(z)$ v podobě $A^*(z)$, přičemž koeficienty a_i a a_i^* by v ideálním případě měly být shodné. Uvedená operace je na obrázku 3.4.

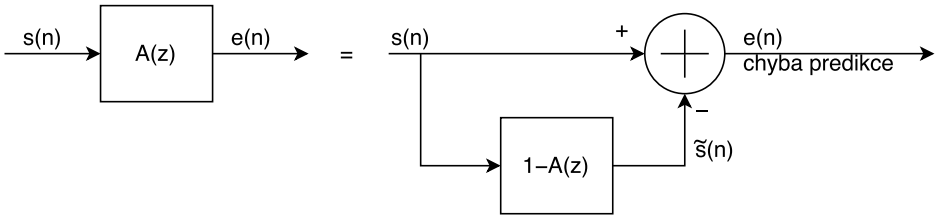


Obrázek 3.4: Aplikace inverzního filtru

Kdyby opravdu shodné byly, potom by výsledný přenos soustavy byl 1. Dá se dokázat, že koeficienty a_i se u stacionárního vstupního signálu dají identifikovat pomocí a_i^*, když se budou a_i^* postupně měnit takovým způsobem, že výsledkem bude minimum energie na výstupu soustavy. Minimum energie se dá vyjádřit jako funkce kvadrátu odchylky předpovězeného vzorku od skutečného vzorku.

Za předpokladu, že jsou koeficienty správně identifikovány, lze považovat funkci $A^*(z)$ za $A(z)$. Tedy na výstupu členu $A(z)$ lze hledat odchylku $e(n)$. Platí transformování dle obrázku 3.5.

Člen $1 - A(z)$ je možno chápat jako prediktor (odtud název metody lineární predikce), neboť skutečně po rozepsání polynomu $A(z)$ pomocí diferenční rovnice je vzorek signálu $\tilde{s}(n)$ vyjádřen pouze pomocí lineární kombinace vzorků minulých bez účasti vzorku přítomného. Chyba predikce je potom



Obrázek 3.5: Substituce vedoucí ke vzniku prediktoru

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \left[-\sum_{i=1}^{P} a_i s(n-i) \right] = s(n) + \sum_{i=1}^{P} a_i s(n-i) \tag{3.4}$$

Druhým a závěrečným krokem identifikace koeficientů je vyřešení sustavy lineárních rovnic, ke kterým se snadno dá dostat pomocí parciálních derivací vztahu $\sum e^2(n)$ položených rovno nule (hledání lokálních minim). Takto se získají konečné koeficienty a_i. K tomuto účelu se používá celá řada metod [11], např. autokorelační metoda s algoritmem Levinson-Durbin, kovarianční metoda.

V praxi znamená pořízení koeficientů IIR filtru velkou výhodu v tom, že velké množství dat z navzorkovaného hlasového signálu lze nahradit mnohem menším objemem dat pomocí právě získaných koeficientů, přičemž původní hlasový signál se dá pomocí inverzního procesu rekonstruovat. Došlo tedy ke kompresi hlasového signálu. Pro potřeby rekonstrukce je ovšem ještě zapotřebí uchovat základní frekvenci, intenzitu vstupu a některé další parametry hlasového signálu. U komprese pro komunikační účely se běžně pracuje s kompresním poměrem 50 až 60, pro srovnání u známého formátu MP3 je kompresní poměr 11.

LPC koeficienty nejsou vhodné pro následné zpracování přímo a v praxi se v původní podobě nepoužívají. Jsou citlivé na přesnost kvantování, nemají omezenou velikost, malá chyba může vést až k nestabilitě filtru. Nejsou podobné ničemu, z čehož by byla patrná jejich praktická reprezentace. Z těchto důvodů se koeficienty pro praktické použití nejdříve podrobí transformaci.

3.3 Line Spectral Frequencies

LSF (Line Spectral Frequencies) nebo také LSP (Pairs) [1] jsou odvozeny od LPC filtru $A(z)$ přičtením, či odečtením části jeho zpětné vazby. Tím je modelován hlasový trakt s otevřenými a zavřenými hlasivkami. Při reálné řeči nejsou hlasivky pouze otevřeny, či uzavřeny, ale často jsou i v přechodném stavu. Právě tento stav je zachycen LSF koeficienty. Z toho můžeme odvodit, že LSF budou z hlediska frekvenčního uzavírat formanty z obou stran. Vznikají tak dva polynomy $P(z)$ a $Q(z)$, které jsou o jeden řád vyšší, než $A(z)$:

$$P(z) = A(z) - z^{-(P+1)} A(z^{-1}) \tag{3.5}$$

$$Q(z) = A(z) + z^{-(P+1)} A(z^{-1}) \tag{3.6}$$

Jejich kořeny leží v komplexní rovině na jednotkové kružnici. Tyto kořeny se navzájem střídají, tj. žádné dva kořeny stejného polynomu neleží na kružnici vedle sebe. Ke každému kořenu jednoho polynomu existuje antisymetrický, či symetrický kořen druhého polynomu.

Stačí proto znát pouze polovinu jejich kořenů, k zachování veškeré informace o LSF. Samotné koeficienty jsou představovány úhly k jejich kořenům v komplexní rovině. Blízkost koeficientů vyjadřuje výraznou rezonanci ve spektru, což je způsobené blízkostí kořene $A(z)$ k jednotkové kružnici.

Pro zpětnou syntézu LPC koeficientů stačí použít vzorce:

$$A(z) = \frac{P(z) + Q(z)}{2} \tag{3.7}$$

Výhodou LSF koeficientů je jejich nízká citlivost na změnu. Představují stabilní IIR filtr, dokud je dodrženo pravidlo následnosti koeficientů – seřazení dle velikosti. Hodnoty koeficientů nabývají hodnot od 0 do π.

LSF koeficienty odrážejí reprezentaci lidského hlasového traktu, respektive jeho modelu, jenž je popsán výše. Základem jsou dvě rezonanční podmínky hlasového traktu, a sice zaprvé za situace, kdyby teoreticky byl trakt na svém začátku, tedy u hlasivek, zcela uzavřen a zadruhé, kdyby byl hlasový trakt úplně otevřen. Realita je ovšem někde mezi těmito dvěma extrémy, hlasivky jsou během řeči stále částečně otevřeny a stupeň otevření se velmi rychle mění. Od výše uvedených dvou rezonančních podmínek jsou odvozeny dvě množiny rezonančních frekvencí s takovým počtem v každé množině, který odpovídá počtu spojených trubic v modelu hlasového traktu. Rezonanční frekvence příslušející každé z podmínek jsou buď co do pořadí vždy liché nebo vždy sudé hodnoty příslušného spektra, navzájem se střídají a monotónně narůstají. Přitom skutečná rezonance, tedy vrcholek formantu, leží někde mezi nimi. Formanty jsou tedy ve spektru obklopeny párem LSP, a lze vysledovat, že čím ostřejší je vrcholek, tím těsnější je pár LSP.

Kapitola 4

Konverze řeči

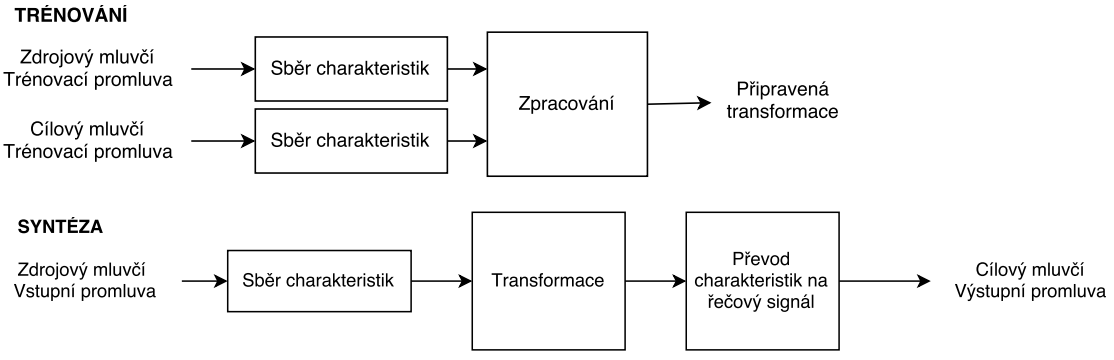
Konverze řeči zahrnuje množinu způsobů, jimiž lze měnit lidský hlas, či zpěv. Jak konverze řeči, tak syntéza řeči, využívají řadu stejných matematických metod. Z pohledu syntézy je konverze užitečná část systému sloužící ke změně barvy syntetizovaného hlasu. Ze strany konverze může syntéza poskytovat účinné nástroje modelující různé aspekty řeči.

Existuje více různých metod přístupu k transformaci řeči. První z nich vznikaly už v 80. letech 20. století [2]. Postupem času se metody vyvíjely až do podoby, kdy syntetizované hlasy znějí věrně.

Byly vytvořeny modely vnímání a produkce hlasu. Modely vnímání slouží k lepšímu porozumění, které charakteristiky hlasu jsou pro posluchače důležité pro jeho identifikaci, a které naopak mohou být z hlasu vypuštěny, neboť lidské ucho na ně není citlivé. Modely produkce rozlišují dva mechanismy tvorby řeči. První popisuje stránku stylu řeči, emocí, nálady, apod. Druhý popisuje průběh vzniku řeči uvnitř artikulačního ústrojí.

Transformace řeči si dává za cíl úpravu zvukové stránky řeči, tedy nehledí na její význam, či syntaxi. Každý člověk má svůj charakteristický hlas. Informace v něm obsažená může být analyzována a také modifikována. Transformace řeči se tedy může použít k napodobení určitého hlasu, nebo naopak jeho zastření, tak aby hlas nebylo možné identifikovat.

Proces konverze má 2 hlavní části: trénovací a syntetizační. Jejich schéma je na obrázku 4.1. Při trénování se porovnávají dvě nahrávky a zjišťuje se způsob konverze/závislost mezi nimi. Zjištěná závislost se aplikuje na vstupní zvukovou stopu prvního řečníka, která se tímto převede na nahrávku cílovou.



Obrázek 4.1: Obecné schéma konverze řeči

4.1 Trénování

Při konverzi řeči je potřeba učinit několik kroků. Tím prvním je porovnání dvou různých řečových nahrávek se stejným obsahem, které byly namluveny dvěma řečníky. Tedy pokud je cílem vytvoření konkrétní transformované nahrávky, která zní jako namluvená řečníkem A, pak je potřeba získat dostatečně dlouhý vzorek řeči mluvčího A. Poté řečník B nahraje jednak stejný vzorek namluvený svým hlasem a jednak vstupní zdrojovou nahrávku. Všechny nahrávky jsou předloženy programu ke zpracování. Při zpracování je potřeba trénovací signály na sebe navázat tak, aby každý foném jedné nahrávky odpovídal stejnému fonému druhé nahrávky. Za tímto účelem je možné využít mapovacího algoritmu.

4.1.1 DTW

DTW (Dynamic Time Warping) – dynamické borcení času – je často používaný algoritmus v oborech rozpoznávání řeči, rukopisu, jednoduchých obrazů, bioinformatice, finančnictví a obecně „dolování dat“ - Data Mining [17]. Je založen na porovnávání a zarovnávání dvou vektorů dat stejné povahy, přičemž zmiňované vstupní vektory zpravidla mají různou délku. Data jednoho vektoru jsou spárována s odpovídajícími daty druhého vektoru. Některé prvky mohou být kopírovány, nebo ignorovány. Tím je dosaženo namapování souvisejících úseků dat, čímž je „zdeformována časová osa“ – ve vstupních vektorech se zbortila „časová souvislost“ vzniku jejich složek. V praxi nemusí jít o reprezentaci času, ale prostě o libovolné vektory.

Metoda DTW pracuje s cenovou maticí o rozměru $N \times M$, kde je N počtem prvků prvního vektoru a M počtem druhého. Jednotlivé prvky matice představují lokální vzdálenosti od počátku do odpovídajícího bodu matice a tím i relativní vzdálenost mezi složkami vektorů. Účelem je minimalizace celkové ceny cesty mezi protějšími vrcholy matice. Nalezená cesta definuje konkrétní zarovnání sekvencí obou vektorů.

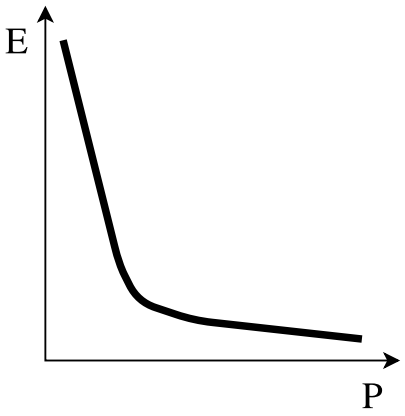
4.1.2 Linear Predictive Coding

Z hlediska konverze řeči hraje LPC v používaných postupech klíčovou roli. U praktických realizací se řád filtru P volí v závislosti na použité frekvenci vzorkování. Čím vyšší frekvence, tím větší P. U určitého zlomového množství další zvyšování počtu koeficientů již nepřináší žádoucí efekt v kvalitě predikce reprezentované poklesem energie chyby prediktoru. Závislost je naznačena na obrázku 4.2.

LPC je použito jak v podsystému trénování, tak i v podsystému syntézy. U trénování jsou koeficienty LPC získány při analýze vstupních trénovacích promluv, reziduální signál zde nemá žádný význam. Naopak u syntézy, kde jsou koeficienty LPC získány při analýze vstupní zdrojové promluvy, se reziduální signál uloží, aby byl později použit jako buzení pro syntézu zmodifikované výstupní promluvy.

4.2 Syntéza

Při syntéze se využívá nové zdrojové nahrávky a natrénovaných konverzních charakteristik. Na jejich základě se syntetizuje výsledná nahrávka. Základní metody, jak tohoto dosáhnout jsou uvedeny v následujících kapitolách.



Obrázek 4.2: Závislost chyby predikce na řádu filtru

4.2.1 Metody

Jak již bylo zmíněno, tak se při produkci hlasu uvažují dva hlavní směry. Rozlišuje se mechanismus vzniku hlasu na fyzické úrovni a způsob modulace hlasu na úrovni psychické. Studie ukazují, že oba tyto mechanismy hrají důležitou roli při identifikaci hlasu, a tím pádem i při jeho modifikaci.

Například imitátoři napodobují hlavně dikci lidí, které imitují - rychlost jejich řeči, intonaci, důraz kladený na určité části slov. Musí odpovídat i použitá slovní zásoba a slovní spojení. Už hůře se napodobuje základní frekvence a napodobení formantů je pro člověka již prakticky nemožné, neboť ty jsou určeny tvarem artikulačního ústrojí.

Většina transformačních metod naopak pracuje převážně s modelem tvorby řeči ve fyzickém slova smyslu. Transformace charakteristik ústrojí je algoritmicky jednodušší než je transformace většiny prozodických vlastností hlasu. Některé prozodické vlastnosti se projevují pouze v určitých částech řeči – například důraz na konci slov, nebo vět [15]. Transformace delších úseků řeči je tedy složitější, neboť řeč zahrnuje více prozodických vlastností. Lidé vnímají kontext celé řeči, a proto u delšího signálu mohou jednodušeji odhalit, že byl signál modifikován.

Navázání jednotlivých segmentů vyžaduje zvláštní pozornost, neboť prozodické rozdíly (perioda základního tónu, energie), skoky ve frekvencích jednotlivých formantů a ve fázi se projeví nepřirozeně znějícím hlasem [24].

4.2.2 Modifikace zdroje

Mezi nejjednodušší modifikace patří úprava zdroje řeči – model hrtanu. Takto je možné měnit hlavně prozodické vlastnosti hlasu, jako je základní frekvence, délka zvuků, či jeho energie [15]. Změny v základní frekvenci vedou ke změnám délky hlasového signálu. Z toho důvodu je zároveň často použita i kompenzující změna délky zvuku, která do signálu kopíruje existující vzorky, nebo je z něj odstraňuje. Člověk dokáže v určitých mezích měnit základní frekvenci svého hlasu, přičemž tato schopnost se dá trénováním zlepšovat. Změny energie lze docílit prostým vynásobením signálu určitým koeficientem.

TD-PSOLA (Time-Domain Pitch Synchronous Overlap and Add) [3] vyhledává příznaky period a na jejich základě signál segmentuje. Obvyklá délka segmentů je 2 až 4 periody. Tyto vzájemně se překrývající segmenty mohou být přiblíženy, nebo oddáleny, což vede ke

změně základní frekvence. Druhou možností je kopírování a mazání segmentů, které změní délku signálu. Po dokončení úprav jsou překrývající se části segmentů sečteny a je z nich vytvořen nový signál. Obě zmíněné úpravy zachovávají kvalitu signálu při nízkém faktoru změn.

4.2.3 Modifikace filtru

Složitější modifikace se zabývají změnami ve spektru řeči. Za tímto účelem se používají různé filtry, u nichž se změna koeficientů může projevit změnou spektrální obálky. Přehled metod z této sekce vychází z knihy Predicting Prosody from Text for Text-to-Speech Synthesis [15].

HNM (Harmonic plus Noise Model) [23] je technika reprezentace řečového signálu. Podle tohoto modelu se signál skládá z harmonické a šumové složky. Harmonická složka modeluje pseudoperiodické jevy pomocí sumy sinusových křivek. Šumová složka modeluje neperiodické děje autoregresním modelem.

Fourierova řada může aproximovat libovolný hlasový signál pomocí funkcí sinus a kosinus. Z toho vychází sinusoidní modely [13]. Ty charakterizují hlas na úrovni frekvencí, amplitud a fází funkcí sinus a kosinus odvozených ze signálu. Takto jsou reprezentovány pouze harmonické složky signálu a pro rekonstrukci neharmonických složek je nutné použít model jiný.

Vocoder STRAIGHT[7] byl vytvořen na základě modelu sluchového aparátu. Díky tomu jeho parametry reprezentují převážně informace, které sami lidé používají při identifikaci mluveného slova. Mezi tyto parametry patří základní perioda, vyhlazený spektrogram a mapa period.

VQ-mapping (Vector Quantization) je technika vytvořená přímo za účelem konverze řeči. Je založená na vytváření mapovacích databází mezi databázemi hlasových parametrů dvou řečníků. Díky tomu je při mapování zachována individualita řeči. Konverze probíhá ve dvou fázích. První je fáze učení, kdy jsou generovány potřebné databáze. Ve druhé fázi se syntetizuje řeč za použití vytvořených mapovacích databází.

Interpolační metody se využívají především ke zlepšení schopnosti převodu mluvené řeči do textu. Charakteristiky cílového řečníka jsou interpolovány na základě předem připravených charakteristik více různých řečníků. Lze tedy říci, že je cílový řečník charakterizován podobností k ostatním řečníkům.

Tento přehled není vyčerpávající, představuje pouze několik hlavních běžně používaných metod. Více podrobností včetně výsledků a srovnání lze nalézt například v přehledu *Voice Conversion: A Critical Survey* [10].

4.3 Současný stav v oboru

V současnosti metody konverze řeči procházejí stálým vývojem, neboť výsledky konverzních systémů nedosahují zaměnitelnosti syntetizované nahrávky s originální cílovou nahrávkou. Vznikají nové metody, které jsou často odvozeny od jiných starších metod. Velmi dobrých výsledků dosahují postupy kombinující více metod současně [27].

4.3.1 Nástroje

Existuje řada softwarových nástrojů – toolkitů, které pomáhají se zpracováním řeči. Pro většinu výše popsaných metod jsou takové nástroje již k dispozici. Při vytváření programu

pro konverzi řeči je vhodné jejich využití zvážit.

SPTK (Speech Signal Processing Toolkit)[1] je soubor nástrojů vytvořených pro práci s řečovými signály v prostředí unixových systémů. Obsahuje rozsáhlou sadu metod implementujících například analýzu a syntézu koeficientů (např.: LPC, LSF, PARCOR, cepstrum, MGC), ale i množství dalších metod pro práci s řečovými daty.

HTS (HMM-based speech synthesis system) toolkit [30] je nástroj podporující syntézu řeči, jež je založen na skrytých Markovových modelech. Systém využívá databáze řečových úseků a na ní modely trénuje.

AHOcoder[2] je nástroj k analýze a syntéze řeči, který byl vyvinut za účelem parametrizace řeči. Signál je popsán třemi skupinami parametrů. Jsou to charakteristiky základní frekvence, spektra a buzení. Zmíněné koeficienty mohou sloužit jako doplněk k HTS.

Signal Processing Toolbox[3] je sada nástrojů programového prostředí MATLAB. Obsahuje funkce vhodné pro vytváření, zpracování, modifikaci a zobrazení signálu. Tyto funkce jsou výkonově optimalizované k práci s maticemi a existuje k nim podrobná dokumentace.

[1] Dostupné pod upravenou licencí BSD na http://sp-tk.sourceforge.net

[2] http://aholab.ehu.es/users/derro/software.html

[3] http://www.mathworks.com/help/signal/

Kapitola 5

Návrh systému

Základ systému je postaven na již existujících metodách pro konverzi řeči. Jednotlivé metody je potřeba do systému správným způsobem zakomponovat a navzájem provázat. Cílem práce je systém navrhnout, implementovat a dále rozvíjet.

5.1 Požadavky na systém

Výsledný systém konverze řeči by měl být schopen zpracovat dva trénovací řečové signály o délce, která je pro trénování dostatečná. Předpokládá se, že signály obsahují tutéž promluvu a jsou dobré kvality. Na základě vztahů trénovacích signálů by měl systém zajistit funkci změny parametrů od zdrojového k cílovému mluvčímu. Změny jsou aplikovány na vstupní zdrojovou promluvu, z níž je syntetizována výstupní cílová promluva, která by měla obsahovat charakteristiky cílového mluvčího.

5.2 Přehled

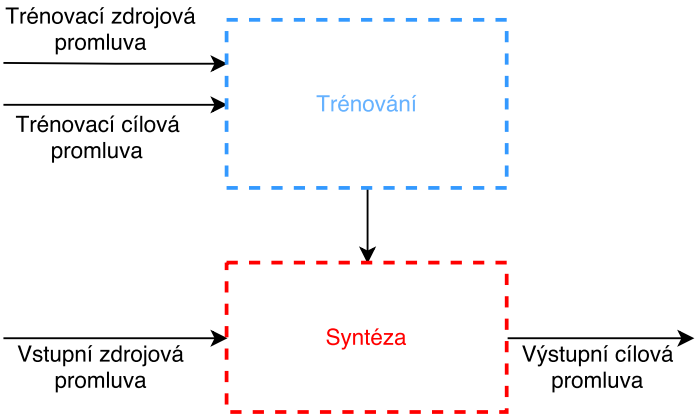
Navrhovaný systém konverze řeči se skládá ze dvou hlavních částí. První částí je trénovací podsystém, ve kterém jsou získány informace o převodu mezi zdrojovým a cílovým řečníkem. Vstupem podsystému jsou dvě nahrávky, jedna od každého řečníka. Obě nahrávky obsahují tutéž sekvenci slov, jejichž porovnáním trénovací podsystém získá převodové matice, které se použijí v druhé části systému – v podsystému syntézy. Zde probíhá převod vstupní zdrojové řeči na výstupní cílovou řeč. Přenos dat mezi podsystémy probíhá pouze jedním směrem. Trénování je na syntéze nezávislé a data, jež jsou vygenerovaná v tomto bloku a jež charakterizující provázání mezi řečníky, postupují do syntézy.

Oba podsystémy jsou vyobrazeny na diagramu 5.1, ze kterého je patrné jejich vzájemné postavení.

Podrobné blokové schéma je na obrázku 5.8 (strana 28). Pro názornost je zde opět modrou čárkovanou čarou zvýrazněn podsystém trénování a červenou čárkovanou čarou podsystém syntézy.

5.2.1 Podsystém trénování

Vstupem podsystému trénování jsou nahrávky zdrojového a cílového řečníka, obsahující tutéž sekvenci slov. K trénování se používá dvojice delších nahrávek cca 1 minuta na rozdíl od přístupu, kdy je vloženo množství nahrávek krátkých. Z obou nahrávek jsou extrahovány posteriorní pravděpodobnosti jednotlivých fonémů pomocí fonémového rozpoznávače



Obrázek 5.1: Obecný diagram podsystémů konverze řeči

PHNREC[1] (vyvinutého výzkumnou skupinou BUT Speech@FIT [21]). Rozpoznávač vytvoří pro každou nahrávku matici, ve které každý řádek obsahuje vyhodnocené pravděpodobnosti výskytu jednotlivých fonémů ve vzorku trvajícím 10 ms (přičemž počet sloupců odpovídá počtu fonémů, který rozpoznávač rozlišuje pro daný jazyk). S maticemi posteriorních pravděpodobností pracuje další prvek systému – DTW. V algoritmu dynamického borcení času se mezi jednotlivými řádky matic hledá jejich podobnost v čase. Výsledkem je matice, v níž je vyznačena tzv. nejkratší cesta. Blok bude popsán v samostatné sekci 5.3, kde bude též charakteristický obrázek zmíněné výsledné matice získaný z konkrétních dat zpracovávaných v této práci (obrázek 5.4 na straně 25).

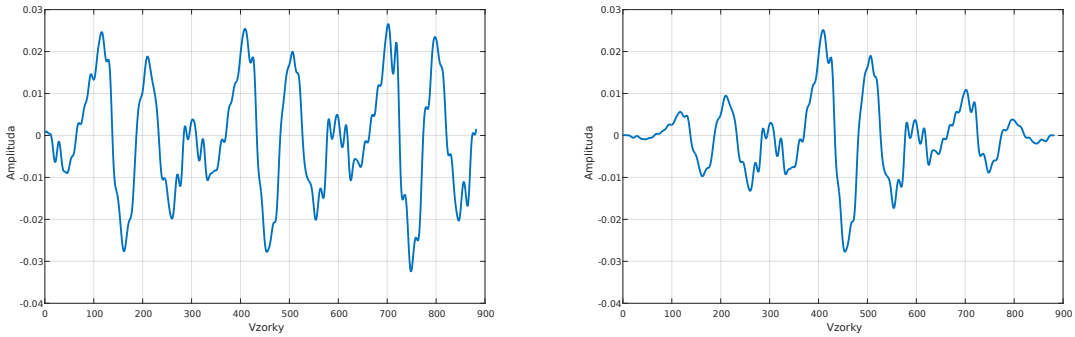
Vstupní nahrávky jsou zpracovány též v paralelní větvi. Na jejím začátku jsou nahrávky normalizovány v blocích Normalizace. Jde o ustřednění nahrávek a vyrovnání jejich energií. Nahrávky dále postupují do bloků Segmentace, kde dochází k jejich rozložení na vzorky o délce 20 ms s překrytím 10 ms. V takto krátkých úsecích se spektrální obálka nemění, a proto je možné na signál pohlížet, jako na statický. Zároveň jsou tyto rámce dostatečně dlouhé na to, aby se v nich objevily alespoň dvě periody základní frekvence. Tyto hodnoty jsou standardně používány v systémech zpracování řeči [16]. Výsledkem segmentace jsou matice segmentů o velikosti $(f_s/1000 \cdot 20) \times (L \cdot 1000/10 - 1)$, kde f_s je vzorkovací frekvence a L délka nahrávek v sekundách.

Výstupy obou paralelních větví dále vstupují do bloku Navázání. Na základě fonetické struktury segmentů dojde k navázání segmentů tak, aby byly vzájemně izomorfní. Efektem je též sjednocení počtů segmentů uvnitř obou matic – počet řádků. Stále se pracuje se dvěma maticemi příslušejícími zdrojovému a cílovému řečníkovi. Protože dochází k modifikaci počtu segmentů, musí se všechny operace změn velikosti odrazit i na matici fonémů, jak v diagramu naznačuje čárkovaná vazba A.

Aby nedocházelo ke zkreslování spekter řeči po aplikaci následného kódovacího algoritmu, jsou segmenty signálů vynásobeny Hammingovým oknem, které eliminuje nežádoucí ostré přechody mezi navazujícími segmenty. Na obrázku 5.2 je znázorněn segment signálu před a po aplikování Hammingova okna.

Následující bloky Ořezání ticha odstraní ze signálu (matic) takové stejnolehlé páry segmentů, u nichž energie alespoň jednoho členu leží pod definovaným prahem. Z toho plyne, že rozměr matic charakterizující délku nahrávek bude i po zmenšení shodný. Současně se

[1] Fonémomový rozpoznávač PHNREC je dostupný na http://speech.fit.vutbr.cz/cs/software/phoneme-recognizer-based-long-temporal-context



Obrázek 5.2: Příklad signálu před a po aplikování Hammingova okna

analogická operace provede nad fonémovými daty, jak v diagramu naznačuje čárkovaná vazba B.

V tomto stádiu jsou segmenty připraveny pro analýzu v bloku LPC (Linear Predictive Coding) - lineární prediktivní kódování. Jsou normalizovány, setříděny a tvarově přizpůsobeny. LPC algoritmus z každého segmentu extrahuje LPC koeficienty. Tím dojde k výraznému zkomprimování signálu při zachování podstatné části hlasové informace. Segmenty signálu se již v dalším zpracování nebudou používat. Místo nich budou použity menší matice koeficientů.

Navazující modul LSF (Line Spectral Frequencies) pouze provede modifikaci koeficientů do tvaru vhodnějšího pro budoucí zpracování.

Poté se nad maticí trénovací zdrojové promluvy provede pseudoinverze v bloku PINV. Nad maticí trénovací cílové promluvy se žádná podobná operace neprovádí.

Dalším logickým blokem je Extrakce. Jejím úkolem je zpracovat, setřídit a sdružit vektory koeficientů v trénovacích maticích. Jejím vstupem je dvojice matic trénovacích koeficientů a jedna matice fonémů. Výstupem je množina normalizačních faktorů a tři transformační matice reprezentující vztahy mezi samohláskami, souhláskami znělými a souhláskami neznělými obou řečníků. Obecně platí vztahy:

$$Z_t^+ C_t = W, \quad Z_v W = C_v \tag{5.1}$$

kde Z_t je matice zdrojových trénovacích koeficientů, C_t je matice cílových trénovacích koeficientů a W je obecně matice transformační. Druhý vztah již naznačuje souvislost se syntézou, neboť Z_v je matice zdrojových vstupních koeficientů a C_v matice výstupních koeficientů.

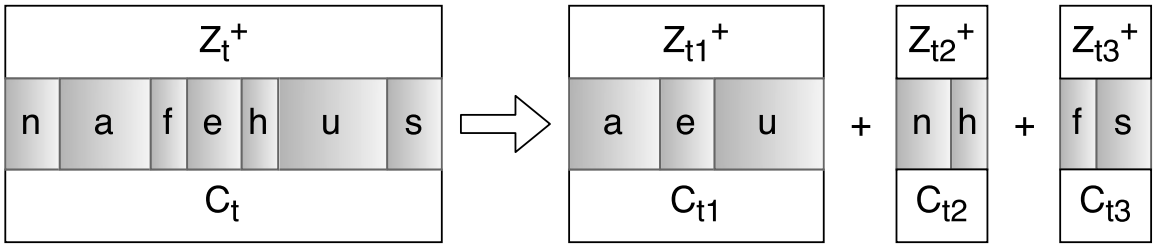
Rozčlenění matic je patrno v podrobnějších rovnicích:

$$Z_{t1}^+ C_{t1} = W_1, \quad Z_{t2}^+ C_{t2} = W_2, \quad Z_{t3}^+ C_{t3} = W_3 \tag{5.2}$$

kde matice s indexem 1, 2, 3 představují skupinu samohlásek, souhlásek znělých a souhlásek neznělých v tomto pořadí. Samotné třídění segmentů dle fonémů je naznačeno na obrázku 5.3.

5.2.2 Podsystém syntézy

Vstupy do podsystému syntézy jsou vstupní zdrojová promluva určená ke konverzi a množina transformačních charakteristik. Na vstupní promluvu jsou aplikovány podobné operace



Obrázek 5.3: Rozdělení matic koeficientů dle hláskových skupin

jako na trénovací promluvy. Jsou to: normalizace, segmentace, násobení Hammingovým oknem, generování LPC koeficientů s převodem na LSF koeficienty. V této sekvenci operací navazuje úprava Hammingovým oknem přímo na Segmentaci a chybí zde blok ořezání ticha.

Při LPC analýze dochází nejen k extrakci koeficientů, ale zároveň se na rozdíl od trénovací fáze uchovává i reziduální signál. Ten je zpracován v bloku Desegmentace, kde se segmenty s reziduálním signálem spojí do souvislého celku.

Dalším logickým blokem je Konverze. Jejím úkolem je převedení koeficientů zdrojových na koeficienty cílové. Vstupy Konverze jsou koeficienty LSF vstupní zdrojové promluvy, transformační charakteristiky z trénovací části a vstupní fonémy. Vstupními fonémy zde rozumíme zpracované posteriorní pravděpodobnosti získané z fonémového rozpoznávače analýzou vstupní promluvy. Výstupem jsou koeficienty LSF pro cílovou výstupní promluvu.

Následuje zpětný převod LSF koeficientů na LPC koeficienty.

Završením procesu je generování výstupní cílové promluvy. K tomu je použit filtr s nekonečnou impulzní odezvou s blokovým označením IIR (Infinite Impulse Response). Do filtru vstupuje reziduální signál a jako jeho koeficienty se použijí LPC koeficienty získané v předchozím kroku.

5.3 DTW

Cílem logického bloku DTW je zmapování dvou trénovacích nahrávek a vytvoření správné informace o jejich časovém navázání. Na vstupu DTW je matice fonémových posteriorních pravděpodobností, na výstupu matice minimální cesty.

Výstupní matice se dá vizualizovat např. ve formě jako na obrázku 5.4, data sloužící k vizualizaci byla získána přímo na základě reálného měření. Minimální cestu představuje křivka spojující levý horní roh obdélníku s pravým dolním rohem. Prohledávaná oblast je zúžena, neboť byl definován požadavek podobnosti obou trénovacích promluv. Z toho plyne, že minimální hledaná cesta nebude ležet daleko od diagonály a výpočty v odlehlejších oblastech matice by byly neefektivní – na obrázku žlutě.

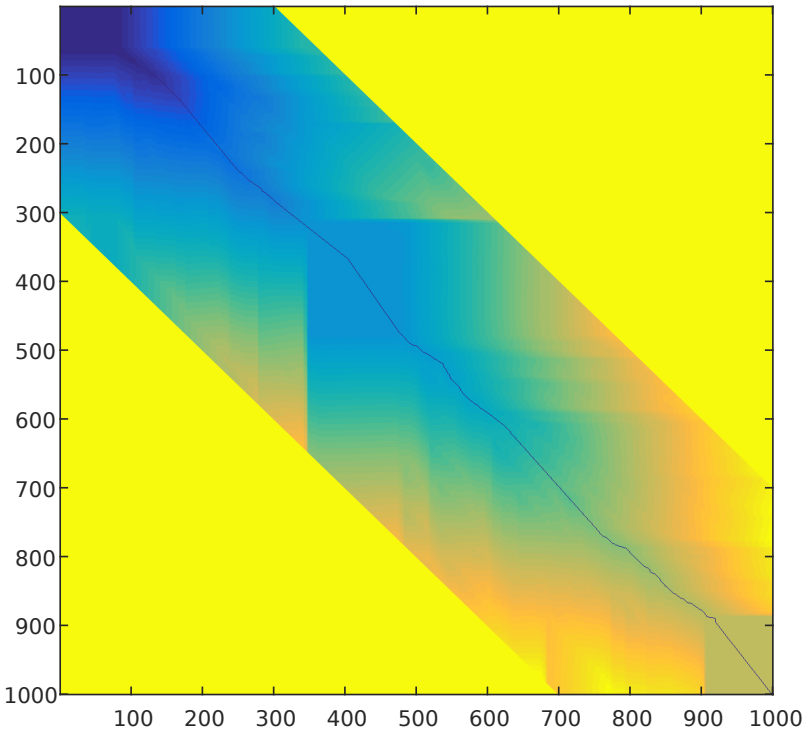
Je možné použít více způsobů pro výpočet podobností vektorů. Mezi hlavní způsoby patří euklidovská vzdálenost a kosinová vzdálenost. Euklidovská vzdálenost udává vzdálenost mezi dvěma body v euklidovském prostoru, jak lze vidět v rovnici 5.3.

$$d = \sqrt{\sum_{i=1}^{n} (A_i - B_i)^2} \tag{5.3}$$

Kosinová vzdálenost vyjadřuje úhel mezi vektory a je definována vztahem v rovnici 5.4. Pokud je úhel mezi vektory nulový – největší podobnost, tak je kosinová vzdálenost rovna 1.

Hodnota klesá s narůstající vzdáleností až do 0 pro nezáporné vektory a do −1 pro obecné vektory.

$$\cos(\theta) = \frac{AB}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{5.4}$$



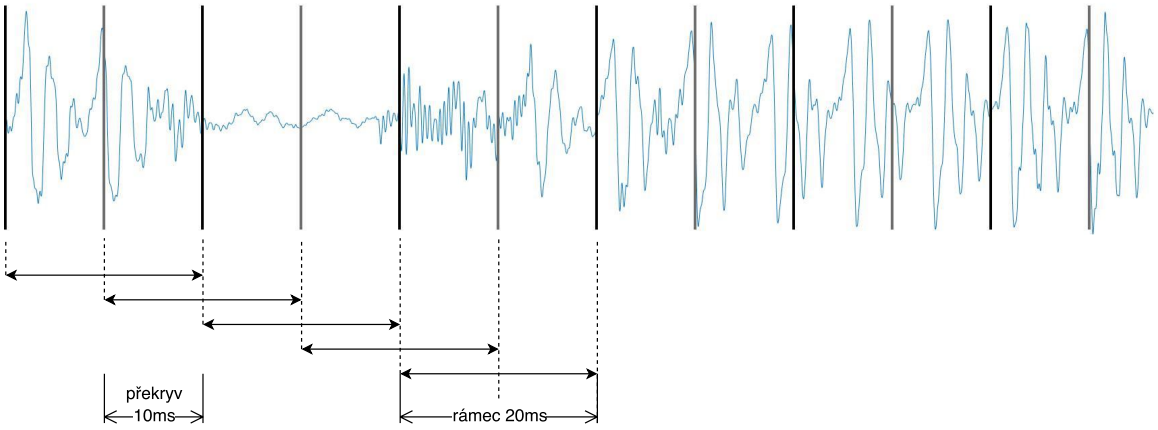
Obrázek 5.4: DTW matice vzdáleností vektorů (kosinová vzdálenost) se znázorněnou minimální cestou

5.4 Segmentace

Blok segmentace slouží k rozdělení souvislé nahrávky na krátké úseky stejné délky vhodné ke zpracování. Za vhodnou délku rámců bylo zvoleno 20 ms [16], jak je znázorněno na obrázku 5.5. Jednotlivé rámce se překrývají o polovinu své délky. Z toho plyne, že ve skutečnosti je signál zpracováván po 10 ms, což odpovídá i délce úseků zpracovávaných v rozpoznávači PHNREC.

5.5 Extrakce

Blok Extrakce je zobrazen na obrázku 5.6. Má na vstupu matice koeficientů LSF, kde každý řádek matice představuje jeden vektor koeficientů (jejíž počet je předem zvolen na základě

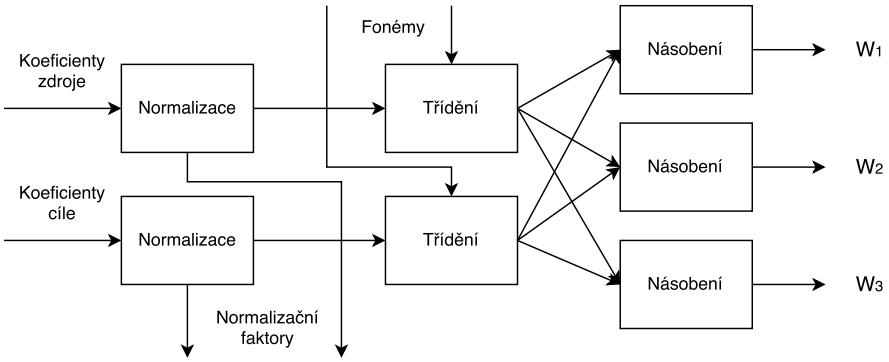


Obrázek 5.5: Segmentace zvukového signálu

přímé úměry se vzorkovací frekvencí). V bloku Extrakce nejprve dochází k normalizaci těchto koeficientů. Výpočet normalizace se uskuteční pomocí formule:

$$Z = \frac{X - \mu}{\sigma} \tag{5.5}$$

Použité normalizační faktory jsou předány do podsystému syntézy, kde je využívá blok Konverze pro normalizaci koeficientů zdrojové promluvy a denormalizaci koeficientů cíle. Koeficienty jsou dále roztříděny podle kategorií fonémů, které představují, čímž vzniknou tři matice pro zdrojovou a tři matice pro cílovou větev. Matice stejného typu se spolu vynásobí. Výsledkem celého procesu jsou transformační tři transformační matice W_1, W_2 a W_3 pro samohlásky a znělé a neznělé souhlásky.

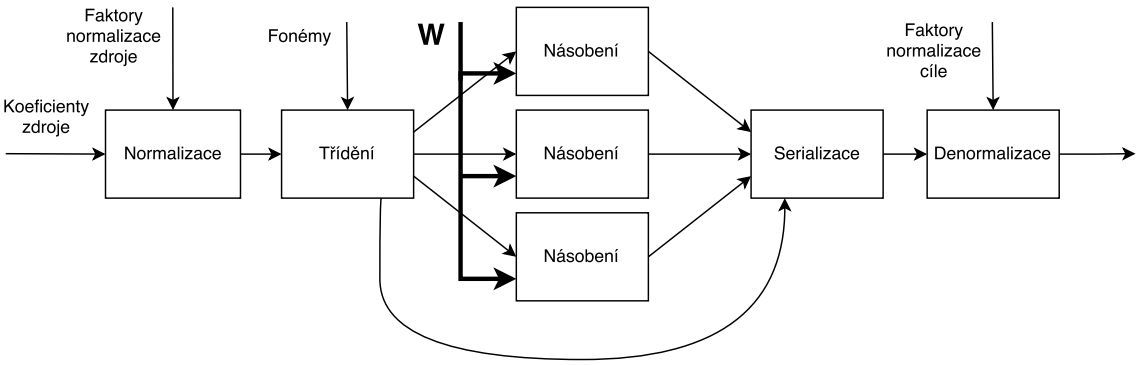


Obrázek 5.6: Blokové schéma extrakce

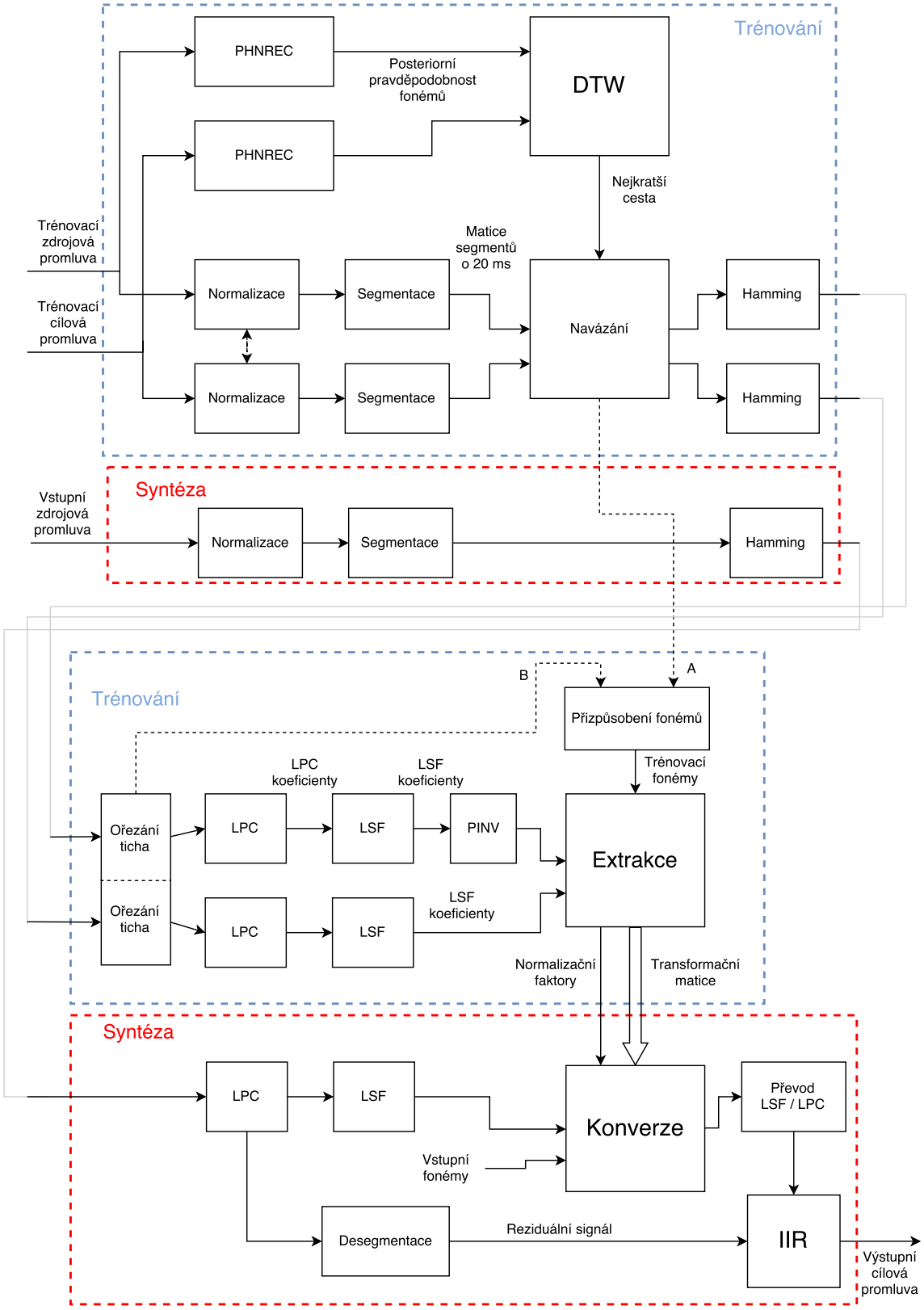
5.6 Konverze

Blok Konverze převádí LSF koeficienty vstupní zdrojové promluvy na LSF koeficienty výstupní cílové promluvy. Vstupní i výstupní LSF koeficienty jsou uloženy v řádcích matic. Obdobně jako u Extrakce jsou nejprve matice zdrojových koeficientů normalizovány, a to za použití normalizačních zdrojových faktorů obdržených z bloku Extrakce. Analogicky jako

u Extrakce jsou koeficienty roztříděny dle fonémů. Informace o jejich rozdělení je uložena pro jejich pozdější zpětné setřídění při serializaci. Matice zdrojových koeficientů LFS jsou vynásobeny transformačními maticemi W_1, W_2 a W_3 podle příslušnosti do stejných fonémových skupin. Tím vznikají matice cílových koeficientů. Následně proběhne již zmíněná serializace a původní tři matice jsou sloučeny do jedné. Ta musí být následně denormalizována, k čemuž poslouží denormalizační cílové faktory (opět z bloku Extrakce).



Obrázek 5.7: Blokové schéma konverze



Obrázek 5.8: Celkové schéma systému

Kapitola 6

Implementace

Podle návrhu popsaném v předchozí kapitole byl implementován program konverze řeči. V této kapitole bude konkrétní implementace detailně popsána. Důležité funkce budou kategorizovány dle jednotlivých bloků v diagramu a tokem dat mezi nimi.

Program byl implementován ve vývojovém prostředí MATLAB, které nabízí jednoduchou práci s maticemi a některé funkce pro zpracování signálu jsou již k dispozici.

Tělo programu konverze řeči se nachází v M-souboru `voice_converter.m`.

Cesty ke vstupním souborům s promluvami se nastavují přímo v hlavním souboru. Nahrávky jsou načteny funkcemi `audioread()`. Soubory posteriorních pravděpodobností ke všem nahrávkám načítá funkce `read_phnrec_file_r()` a ukládá je do matice posteriorních pravděpodobností fonémů.

6.1 Trénování

Signály obou trénovacích promluv jsou načteny do vektorů a je uložena informace o jejich vzorkovací frekvenci. Program počítá s tím, že je frekvence obou trénovacích nahrávek stejná. Použité demonstrační nahrávky mají vzorkovací frekvenci 44100 Hz. První operace provedené nad signály jsou ustřednění a vyrovnání jejich energie, které proběhnou ve funkcích `raw_center()` a `raw_normalize_to_raw_s()`. Jde o úpravy spadající do bloku Normalizace. Od vektorů nahrávek je odečtena jejich průměrná hodnota a poté je vektor cílové vstupní nahrávky vynásoben poměrem standardních odchylek obou nahrávek.

Matice posteriorních pravděpodobností fonémů má 135 sloupců. Zmíněné číslo zahrnuje rozlišení začátku, středu a konce většiny stavů, které fonémový rozpoznávač rozeznává. Program konverze hlasu vyžaduje použití české sady, neboť předpokládá určité rozložení fonémů uvnitř matice. Ticho může být v matici reprezentováno více způsoby, z toho důvodu je matice zpracována pomocí funkce `phnrec_vector_adj()`. Tato funkce slučuje ty údaje, jejichž rozlišení není pro aplikaci konverze relevantní.

Program používá dvě matice posteriorních pravděpodobností fonémů k vytvoření DTW matice cen a nalezení minimální cesty. DTW matice je vytvořena funkcí `dtw()`. V této funkci je cena jednotlivých pozic v matici je počítána pomocí kosinové vzdálenosti. Tuto funkci lze v kódu vyměnit za funkci `dtw_euclidean()`, která počítá cenu Euklidovskou vzdáleností. Vzdálenost od diagonály je omezena pásem Sakoe-Chiba [18] s odchylkou nastavenou na 3 s. Minimální cesta je nalezena funkcí `dtw_process()`, která prochází matici cen od začátku ke konci a vytváří matici cesty, v níž je cesta vyznačena. Obdobný proces je použit na nalezení druhé cesty procházením od konce k začátku. Jde o variantu obousměrného DTW [26],

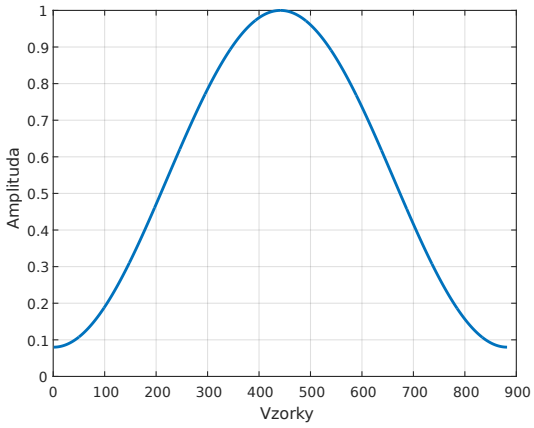
která je vhodná ke snížení počtu opakování stejného úseku. Výsledná cesta je vypočítána ve funkci `dtw_process_2D()`, která spojí obě nalezené cesty a v místech, kde jsou cesty rozděleny, spojí první a poslední pozici před a po rozdělení.

Normalizovaný signál se rozdělí na rámce o délce 20 ms. K tomu dochází ve funkci `vector_to_mat()`. Počet prvků odpovídajících 20 ms se vypočítá na základě vzorkovací frekvence. Výsledná matice má tedy počet sloupců roven vzorkovací frekvenci [v kHz] signálu vynásobené délkou rámce. Kvůli 10 ms překryvu má matice vysokou hodnotu redundance, protože kromě první poloviny prvního rámce a druhé poloviny posledního rámce je celí signál v matici obsažen dvakrát. Pokud není poslední rámec naplněn do konce, jsou jeho hodnoty doplněny o nuly (tzv. zero padding).

V této fázi programu jsou obě matice rámců trénovacích nahrávek navázány tak, aby rámce obsahující stejné fonémy byly v obou maticích uloženy na stejném místě (z hlediska řádků matic). Algoritmus sleduje cestu vytvořenou v bloku DTW a pro každý vztah 1:N kopíruje N-krát segment příslušející jedničce. Obě matice se tedy mohou pouze zvětšovat. Výsledné matice jsou stejné délky a řádky jedné matice odpovídají foneticky řádkům druhé.

Zároveň s navázáním matic probíhá označení segmentů dle kategorií fonémů, pod které spadají. Funkce `sort_phonemes()` využívá fonémových údajů a vytváří vektor fonémů o velikosti počtu řádků matic segmentů.

Všechny řádky matic jsou vynásobeny Hammingovým oknem (obrázek 6.1), čímž se zpřesňuje další analýza. Takto upravené segmenty lze též „složit“ a vytvořit nový testovací signál (funkce `recon_mat_win()`), na němž lze testovat účinnost již provedených změn (např. testování DTW). Hammingovo okno na řádky matic aplikuje funkce `mat_hamming()`. Operace zpřesňuje další analýzu krátkých segmentů – upravuje přechod mezi vzorky uvnitř okna a nulami vně.

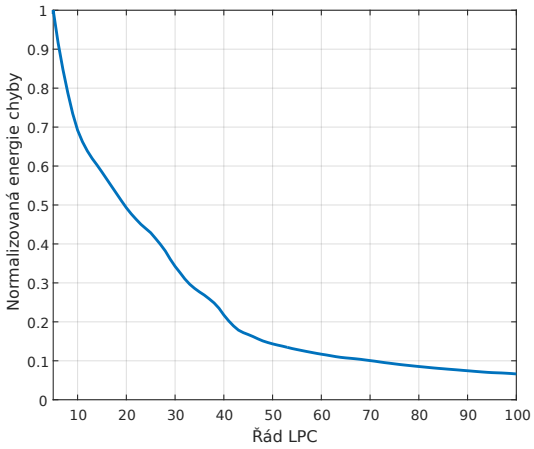


Obrázek 6.1: Hammingovo okno

Algoritmus maže segmenty s nízkou energií. Energie ticha (šumu) se velmi odlišuje od energie mluveného slova. Práh pro vymazání segmentu byl experimentálně stanoven jako jedna tisícina průměrné energie celé nahrávky.

Nad maticemi obsahujícími upravené testovací řečové segmenty je provedena LPC analýza ve funkci `mat_lpc()`. Druhým argumentem této funkce je počet LPC koeficientů, který má být vytvořen. Výchozí hodnota je 32, což je počet koeficientů vhodný pro signál o vzorkovací frekvenci 44100 Hz. Závislost predikční chyby na řádu LPC je pro signál o vzorkovací frekvenci 44100 Hz zobrazena na grafu 6.2. Počet sloupců výstupní matice je roven počtu

LPC koeficientů.



Obrázek 6.2: Změřená závislost chyby predikce na řádu koeficientů LPC pro $F_s = 44{,}1\,\text{kHz}$

Převod LPC koeficientů na LSF koeficienty se provádí ve funkci `lpc_to_lsf()`. Výstupní matice má stejnou velikost jako matice LPC koeficientů.

Funkce `coef_normalize()` je zodpovědná za normalizaci koeficientů. V matici jsou koeficienty každého segmentu uloženy v řádcích, proto normalizace probíhá po sloupcích. V každém sloupci jsou koeficienty stejného řádu. Průměry hodnot ve sloupcích jsou zaznamenány společně se směrodatnými odchylkami a na jejich základě je normalizace provedena.

Předtím, než proběhne výpočet transformačních matic, jsou segmenty s koeficienty rozřazeny do tří skupin podle typu hlásek, pod který spadají. Dělí se do kategorií samohlásky, souhlásky znělé a souhlásky neznělé. Z jedné matice koeficientů se tímto krokem stávají tři matice. Pro obě trénovací nahrávky (celkem šest matic) mají příslušné matice koeficientů stejné velikosti.

Po roztřídění koeficientů již nic nebrání vytvoření samotných transformačních matic W_1, W_2 a W_3. Matice trénovací zdrojové promluvy jsou podrobeny pseudoinverzi a vynásobeny svými protějšky z trénovací cílové promluvy. Tímto je ukončen blok trénování a transformační matice jsou přichystány k modifikaci zdrojových vstupních koeficientů na cílové výstupní.

6.2 Syntéza

Průběh zpracování vstupní zdrojové promluvy je velmi podobný zpracování trénovacích promluv popsaných v předchozí kapitole. Zdrojová promluva je uložena do vektoru a je uložena informace o její vzorkovací frekvenci. Proběhne normalizace nahrávky a její segmentace, která z vektoru vytvoří matici segmentů. Každý segment je vynásoben Hammingovým oknem. Dále probíhá LPC analýza a převod koeficientů na LSF. Při LPC analýze se navíc uchovává reziduální signál, který později poslouží jako buzení filtru výstupní cílové promluvy.

Segmenty LSF koeficientů jsou modifikovány normalizačními faktory zdroje. Stejným způsobem, kterým při trénování proběhla sloupcová normalizace v trénovací zdrojové matici LSF koeficientů, je modifikována matice vstupních zdrojových LSF koeficientů určených ke konverzi.

Obdobně jako při trénování je ke vstupní zdrojové promluvě navázána matice obsahující její fonetickou informaci. Na základě této fonetické informace jsou segmenty matice LSF koeficientů rozřazeny do stejných skupin, jako tomu bylo při trénování. Vznikají tři matice koeficientů, které jsou násobeny příslušnými transformačními maticemi. Výsledkem násobení jsou tři nové matice, jež obsahují syntetizované LSF koeficienty cílové. Dále proběhne sloučení vytvořených matic do matice jedné. Na této výsledné matici probíhá denormalizace, tedy modifikace denormalizačními faktory cíle. Denormalizace je implementována ve funkci `coef_denormalize()`.

V syntetizované matici se v tomto kroku nacházejí výstupní cílové LSF koeficienty. Ve funkci `lsf_to_lpc()` jsou LSF koeficienty převedeny na LPC koeficienty. Následně probíhá rekonstrukce cílové nahrávky ve filtru s nekonečnou impulzní odezvou. Jako budící signál je použit upravený reziduální signál zdrojové nahrávky a je doprovázený LPC koeficienty opatřenými při syntéze. Koeficienty filtru jsou průběžně upravovány a je vytvářena jedna souvislá nahrávka.

Syntéza končí uložením syntetizované cílové výstupní nahrávky funkcí `audiowrite()`.

Kapitola 7

Experimenty a zhodnocení

Implementace programu konverze řeči probíhala postupně v iteračních krocích. V každém kroku byly stanoveny testy, které validovaly funkčnost částí programu. Hlavními částmi se staly celky vzniklé po završení segmentace, DTW, LPC analýzy, LSF analýzy, bloku Konverze a bloku Syntézy.

Výsledky testů byly zhodnoceny jak subjektivně, tak objektivně. Subjektivní testování zahrnovalo poslechy nahrávek, které byly zrekonstruovány po aplikaci příslušných modifikací. U objektivního testování bylo využito metod porovnávání spekter hlasů, měření energií apod. Vyhodnocování systémů konverze hlasu klade velký důraz hlavně na subjektivní stránku, neboť komplexnost vnímání řeči znesnadňuje nalezení objektivní vyhodnocovací metriky [25].

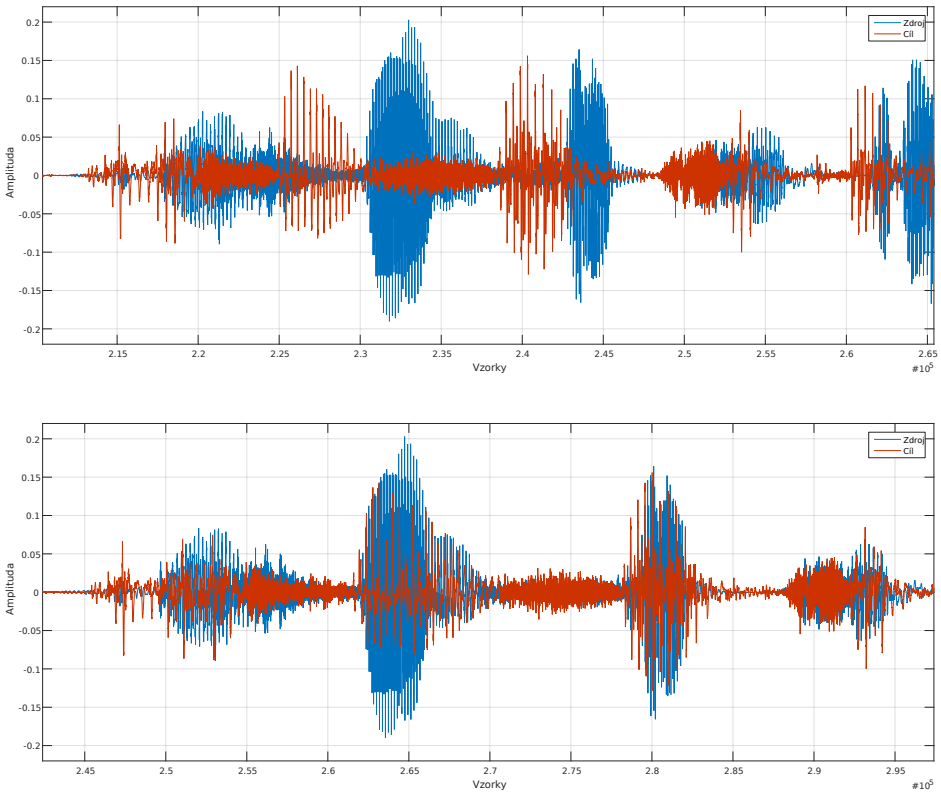
V následujících kapitolách budou popsány způsoby testování funkčnosti hlavních částí programu, které zároveň odpovídají blokům v diagramu na obrázku 5.8.

7.1 Segmentace

Mezi první implementované části patří segmentace, která má na vstupu vektor signálu a na výstupu matici segmentů. Cílem testování bylo ověřit, že nahrávka může být nasegmentována a opět složena do původního tvaru bez ztráty její kvality. Byla využita vlastnost Hammingova okna, kdy součet překrývajících se vzorků vynásobených Hammingovým oknem odpovídá původnímu signálu. Matice vzniklá segmentací byla proto použita na vstupu funkcí `mat_hamming()` a `recon_mat_win()`. Z výsledného vektoru zrekonstruovaného signálu byla vytvořena nahrávka znějící identicky se vstupní nahrávkou.

7.2 DTW

Dalším bodem kontroly bylo dynamické borcení času. Cílem testu bylo zjistit, zda došlo ke správnému provázání nahrávek. Byla vytvořena stereo nahrávka, kde levý kanál reprezentoval promluvu zdrojového řečníka a pravý kanál promluvu cílového řečníka. Poslechem bylo zjištěno, že obě promluvy si odpovídají časovým průběhem. V promluvě se vyskytly artefakty, ale ty byly lokalizovány pouze v částech ticha. Jelikož nejsou tyto tiché segmenty dále použity, lze konstatovat, že provázání proběhlo úspěšně. Na obrázku 7.1 jsou zobrazeny dva signály, zdrojový a cílový, o délce asi 1 s. Horní graf ukazuje signály před zarovnáním, dolní po zarovnání.



Obrázek 7.1: Porovnání řečových signálů před a po aplikaci DTW. (Oba mluvčí právě vyslovují slova „*. . . když jsem se při. . .* “.)

7.3 LPC

Byl testován převod matice segmentů na matici LPC koeficientů a zpět. Nahrávka byla zrekonstruována IIR filtrem za použití zmíněných koeficientů a příslušných rezidualních signálů. Tímto způsobem transformovaná nahrávka byla prakticky identická se zdrojovou nahrávkou.

7.4 LSF

Blok LSF byl otestován obdobně jako blok LPC. Z nahrávky byly extrahovány LPC koeficienty, které byly následně převedeny na LSF koeficienty. Po zpětném převodu a rekonstrukci nahrávky nebyl patrný rozdíl mezi poslechem zdrojové a rekonstruované nahrávky.

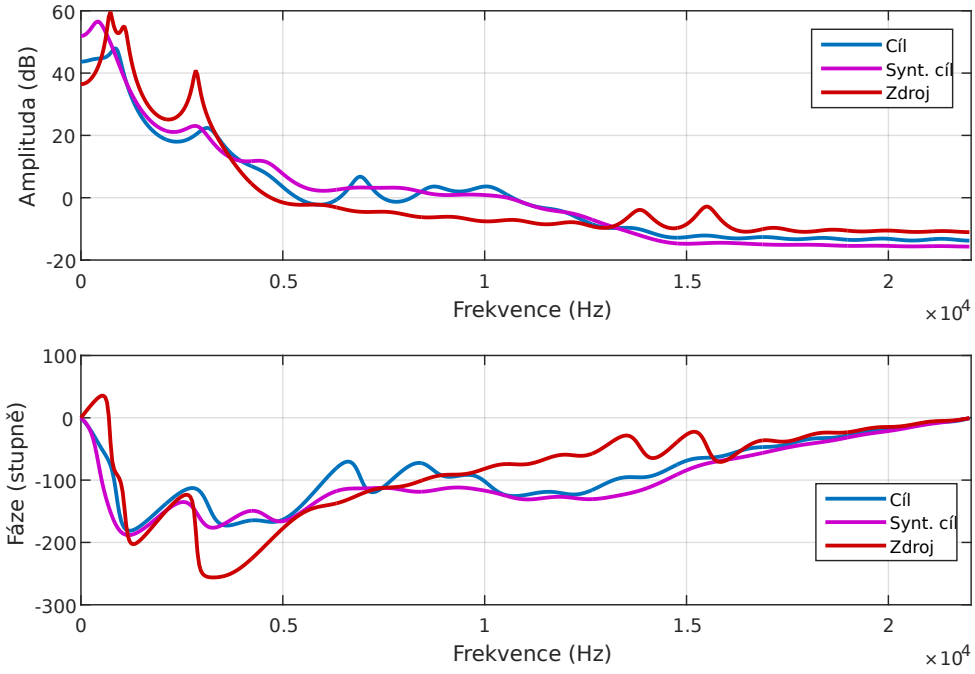
Zároveň bylo testováno zatížení chybou při malé změně LSF koeficientů. K prvkům matice LSF koeficientů byla přičtena náhodná hodnota. Pokud tato hodnota byla menší než 1E-3, tak na nahrávce nebyl patrný rozdíl. LSF nabývá hodnot $\langle 0, \pi \rangle$.

7.5 Extrakce a Konverze

Cílem testů bloků Extrakce a Konverze bylo zjistit, do jaké míry jsou transformační matice schopné transformovat LSF zdrojové koeficienty na cílové. V této fázi byla uvedena vstupní

zdrojová promluva spolu s očekávanou výstupní cílovou promluvou. Obě tyto promluvy vznikly z části trénovacích promluv, které však nebyly při trénování použity (byly ovšem zarovnány pomocí DTW). Díky tomu mohly být tyto promluvy využity pro zhodnocení účinnosti konverze.

Při testech se porovnávala spektrální obálka modifikované řeči zdrojového řečníka se spektrální obálkou očekávané řeči cílového řečníka. Na obrázku 7.2 lze vidět, že pro hlásku „A“ jsou si obálky transformovaného segmentu zdrojového řečníka a opravdového segmentu cílového řečníka dosti podobné.



Obrázek 7.2: Srovnání spektrálních obálek samohlásky „A“ u cílové, zdrojové asyntetizované nahrávky

Na základě těchto výsledků lze předpokládat, že při konverzi je do jisté míry provedena transformace formantů zdroje na cíl. Přesně vyčíslit úspěšnost konverze ovšem není možné, neboť ta závisí na konkrétních vstupních segmentech LSF koeficientů, které jsou při konverzi modifikovány.

7.6 Syntéza

Úkolem bloku syntézy je převod vstupní zdrojové promluvy na promluvu výstupní cílovou. Výstupní cílová promluva by měla znít co nejpodobněji jako promluva nahraná cílovým mluvčím. Na tuto skutečnost byly testy syntézy zaměřeny. Znovu byla použita nahrávka „očekávaná“ na výstupu a ta byla srovnávána s nahrávkou vzniklou – syntetizovanou.

Podobnost těchto nahrávek byla nízká. Výsledná nahrávka zněla jako kombinace obou řečníků. Tato skutečnost je částečně způsobena tím, že navržený program konverze řeči nemění prozodické vlastnosti nahrávek. Při poslechu nahrávky lze z intonace snadno rozeznat hlas zdrojového řečníka. Aby se vliv těchto složek hlasu při testech utlumil, byla navržena

další sada testů zaměřená na potlačení prozodické stránky hlasu.

Tato sada testů byla založena na použití jiného budícího signálu filtru IIR. Místo extrahování reziduálního signálu z nahrávky zdrojové byl tento signál opatřen přímo z cílové „očekávané“ nahrávky. Při praktickém použití konverze řeči by samozřejmě takováto nahrávka neexistovala, ale z důvodu testů je její použití výhodné. Tímto způsobem opatřený reziduální signál má v sobě obsaženy potřebné prozodické vlastnosti cílové řeči. Zároveň je tento signál zarovnán se zdrojovým signálem pomocí DTW. Problémy u tohoto typu testování jsou artefakty vzniklé při DTW a také skutečnost, že reziduální signál částečně obsahuje některé neprozodické informace o lidském hlasu.

S ohledem na zmíněné skutečnosti byla vytvořena nahrávka kombinací syntetizovaných LPC koeficientů a budícího signálu extrahovaného z „očekávané“ nahrávky. Hlas ve výsledné nahrávce je již zřetelně podobný hlasu cílovému.

Podíl DTW na zkreslení cílové nahrávky lze částečně určit porovnáním této nahrávky a nahrávky vytvořené podobným způsobem, ale s použitít LPC koeficientů „očekávané“ nahrávky. Artefakty nalezené v obou záznamech jsou následkem použití algoritmu DTW, neboť ostatní použité modifikace (bloky Normalizace, Segmentace, LPC) nemají dopad (viz. výše) na kvalitu výsledné nahrávky.

Kapitola 8

Závěr

V této práci byly popsány různé aspekty konverze řeči. Byla rozebrána hlasová stránka řeči včetně způsobu její tvorby. Důraz byl kladen na parametry charakterizující a identifikující lidský hlas. V kontextu syntézy řeči byl přiblížen základní model artikulačního ústrojí a bylo ukázáno, jakými způsoby lze tento model algoritmizovat. Na zmíněné poznatky bylo navázáno při popisu metod modifikace řeči. Zvýšené pozornosti bylo věnováno metodám přímo souvisejícím s konverzí hlasu. Metody byly rozděleny dle jejich nejčastějšího použití do dvou základních kategorií — trénování a syntéza. Byly vysvětleny jejich vzájemné odlišnosti a jejich podíl při konverzi. Byl vytvořen přehled nástrojů, které zjednodušují práci s transformačními metodami i s řečovým signálem samotným. Na základě získaných poznatků byl navržen systém konverze hlasu. Návrh ukazuje, jaké funkční bloky v systému existují a jaké jsou jejich vzájemné vztahy. U funkčních bloků bylo obecně vysvětleno jejich chování.

Podle popsaného modelu byl postupně vytvářen program konverze hlasu. Jako vývojové prostředí programu bylo zvoleno programové prostředí MATLAB, a to konkrétně jeho verze R2015a. MATLAB nabízí sadu užitečných nástrojů pro zpracování signálu a je optimalizovaný na práci s maticemi. Samotný program byl tvořen iteračně, tedy tak, že byla ověřována funkčnost jeho částí a na základě výsledků byl volen jeho další rozvoj. Důležité výsledky testů byly prezentovány v kapitole 7. Celková implementace vytvořeného programu konverze hlasu byla popsána v kapitole 6.

Vstupní i výstupní nahrávky jsou k dispozici na přiloženém DVD. Z výsledků je patrný posuv charakteristik hlasu zdrojového řečníka k cílovému. I přesto není syntetizovaná výsledná řeč příliš podobná cílovému mluvčímu. Podobnost se výrazně přiblíží při aplikaci budícího signálu extrahovaného přímo z cílové nahrávky. Tento krok není možné učinit v praxi (originální cílová nahrávka neexistuje), ale jeho výsledky ukazují důležitost prozodické stránky řeči.

Další možnost rozvoje vidím v modifikaci budícího signálu. Hlasový signál je velmi dynamický a proto je k němu potřeba i přistupovat dynamicky. Do systému by mohly být zavedeny statistické metody, jež by na základě mnoha faktorů odhadovaly, jakým způsobem by měl být budící signál upraven [31]. Segmentace by mohla probíhat s ohledem na základní frekvenci hlasu. Práce se segmenty proměnlivé velikosti je mnohem složitější, ale zpřesňuje přístup k datům [14]. Transformačních matic na převod LSF koeficientů by mohlo být vytvořeno více, případně by se program mohl pokusit vytvořit transformační matici pro každý foném zvlášť. V takovém případě by bylo nutné najít způsob, jakým co nejlépe odhadnout transformační matici, pro níž by neexistoval dostatek trénovacích dat. Zde by mohlo pomoci vytvoření tabulky příbuznosti fonémů, podle níž by byla data doplněna.

Literatura

[1] Benesty, J.; Sondhi, M. M.; Huang, Y. A.: *Springer Handbook of Speech Processing*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2007, ISBN 3540491252.

[2] Childers, D. G.; Yegnanarayana, B.; Wu, K.: Voice conversion: Factors responsible for quality. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'85.*, ročník 10, IEEE, 1985, s. 748–751, doi:10.1109/ICASSP.1985.1168479.

[3] Dutoit, T.: *An introduction to text-to-speech synthesis*. 1997, ISBN 9789401157308, doi:10.1007/978-94-011-5730-8.

[4] Fant, G.: *Acoustic Theory of Speech Production: With Calculations Based on X-Ray Studies of Russian Articulations (Description and Analysis of Contemporary Standard Russian)*. De Gruyter Mouton, reprint 2012 ed. vydání, Leden 1971, ISBN 9027916004.

[5] Flanagan, J. L.: *Speech analysis, synthesis, and perception.* Springer, 1965, ISBN 9783662008515.

[6] Joliveau, E.; Smith, J.; Wolfe, J.: Vocal tract resonances in singing: The soprano voice. *The Journal of the Acoustical Society of America*, ročník 116, č. 4, Říjen 2004: s. 2434–2439, ISSN 0001-4966, doi:10.1121/1.1791717.

[7] Kawahara, H.; Masuda-Katsuse, I.; de Cheveigné, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, ročník 27, č. 3-4, Duben 1999: s. 187–207, ISSN 01676393, doi:10.1016/s0167-6393(98)00085-5.

[8] Krčmová, M.: *Fonetika a fonologie [online]*. Elportál, Masarykova univerzita, druhé vydání, 2008 [cit. 2016-05-14].
URL http://is.muni.cz/elportal/?id=766384

[9] Linder, G.: *Einführung in die experimentelle Phonetik*. Berlin: Akademie-Verlag, 1969.

[10] Machado, A. F.; Queiroz, M.: Voice Conversion: A Critical Survey. In *Proceedings of SCM Conference 2010*, Barcelona, 2010.

[11] Makhoul, J.: Linear prediction: A tutorial review. *Proceedings of the IEEE*, ročník 63, č. 4, Duben 1975: s. 561–580, ISSN 0018-9219, doi:10.1109/proc.1975.9792.

[12] Mannell, R.: Vowel Perception in Australian English. Prosinec 2008.

[13] McAulay, R. J.; Quatieri, T. F.: Speech Processing Based on a Sinusoidal Model. *The Lincoln Laboratory Journal*, ročník 1, č. 2, 1988: s. 153–168.

[14] Moulines, E.; Charpentier, F.: Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, ročník 9, č. 5-6, Prosinec 1990: s. 453–467, ISSN 01676393, doi:10.1016/0167-6393(90)90021-z.

[15] Rao, K. S.: *Predicting Prosody from Text for Text-to-Speech Synthesis*. Springer, 2012, ISBN 1461413370, 9781461413370, doi:10.1007/978-1-4614-1338-7.

[16] Rao, K. S.; Vuppala, A. K.: *Speech Processing in Mobile Environments*. Springer, 2014, ISBN 3319031155, 9783319031156, doi:10.1007/978-3-319-03116-3.

[17] Ratanamahatana, A.; Keogh, E.: Everything you know about dynamic time warping is wrong. *3rd Workshop on Mining Temporal and Sequential Data, in conjunction with 10th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD-2004)*, 2004.

[18] Sakoe, H.; Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ročník 26, č. 1, Únor 1978: s. 43–49, ISSN 0096-3518, doi:10.1109/tassp.1978.1163055.

[19] Schröder, M.: Expressive Speech Synthesis: Past, Present, and Possible Futures. In *Affective Information Processing*, editace J. Tao; T. Tan, Springer London, 2009, ISBN 978-1-84800-306-4, s. 111–126, doi:10.1007/978-1-84800-306-4_7.

[20] Schroeder, M. R.: *Computer speech : recognition, compression, synthesis*. Springer-Verlag, druhé vydání, 2004, ISBN 3540212671.

[21] Schwarz, P.: *Phoneme recognition based on long temporal context*. Dizertační práce, Vysoké Učení Technické v Brně, Fakulta informačních technologií, Brno, 2008.

[22] Smith, J. O.: *Physical audio signal processing : for virtual musical instruments and audio effects [online]*, kapitola Voice Synthesis. W3K Publishing, 2010 [cit. 2016-05-14], ISBN 0974560723.
URL http://www.worldcat.org/isbn/0974560723

[23] Stylianou, Y.: Concatenative Speech Synthesis using a Harmonic plus Noise Model. In *in The 3rd ESCA/COCOSDA Workshop on Speech Synthesis, Jenolan Caves*, 1998, s. 261–266.

[24] Stylianou, Y.: Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Speech and Audio Processing*, ročník 9, č. 1, Leden 2001: s. 21–29, ISSN 1063-6676, doi:10.1109/89.890068.

[25] Sündermann, D.: Voice conversion: state-of-the-art and future work. In *Proceedings of the 31st German Annual Conference on Acoustics (DAGA '01)*, Mnichov, Německo, 2001.

[26] Tamulevičius, G.; Serackis, A.; Sledevič, T.; aj.: Bidirectional Dynamic Time Warping Algorithm for the Recognition of Isolated Words Impacted by Transient Noise Pulses. *International Journal of Computer, Electrical, Automation, Control and Information Engineering*, ročník 8, č. 4, 2014: s. 63–67.

[27] Tian, X.; Wu, Z.; Lee, S.-W.; aj.: System Fusion for High-Performance Voice Conversion. In *Interspeech*, 2015.

[28] Vary, P.; Martin, R.: *Digital speech transmission : enhancement, coding and error concealment*, kapitola 2.3.1 Acoustic Tube Model of Vocal Tract. John Wiley, Leden 2006, ISBN 9780471560180, s. 12–19.

[29] Wu, Z.; Li, H.: Voice conversion and spoofing attack on speaker verification systems. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2013 Asia-Pacific*, IEEE, Říjen 2013, s. 1–9, doi:10.1109/apsipa.2013.6694344.

[30] Zen, H.; Nose, T.; Yamagishi, J.; aj.: The HMM-based speech synthesis system (HTS) version 2.0. In *6th ISCA Workshop on Speech Synthesis (SSW6)*, Bonn, Německo: ISCA, Srpen 2007, s. 294–299.

[31] Zen, H.; Tokuda, K.; Black, A. W.: Statistical parametric speech synthesis. *Speech Communication*, ročník 51, č. 11, Listopad 2009: s. 1039–1064, ISSN 01676393, doi:10.1016/j.specom.2009.04.004.