Portál pre odborné publikovanie ISSN 1338-0087

Syntéza reči v mobilnom telefóne pomocou komprimovanej parametrickej databázy

Makovínyi Pavel · Informačné technológie, Študentské práce

28.11.2011

Táto práca je venovaná implementácii komprimovanej parametrickej databázy v aplikácii multimediálneho čítania SMS správ v mobilnom telefóne. V prvej časti je opísaná aplikácia a súčasný stav syntézy reči v tejto aplikácii. Ďalej je opísaná syntéza reči pomocou difón. V tretej časti je uvedený princíp HNM modelu a štvrtá časť popisuje štruktúru komprimovanej parametrickej databázy. Na záver je opísané generovanie syntetizovaného zvuku.

1. Úvod

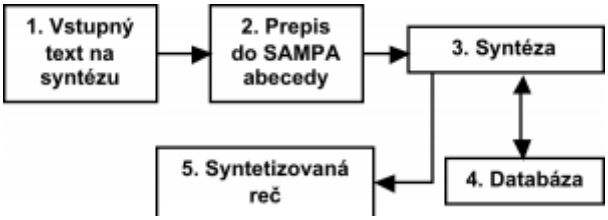
Aplikácia multimediálneho čítania SMS správ v mobilnom telefóne je pod vedením doc. Ing. Gregora Rozinaja, PhD. na katedre telekomunikácií vyvíjaná už niekoľko rokov. Aplikácia je postavená na technológií Java ME (predchádzajúci názov Java 2 Platform, Micro Edition – J2ME), ktorá je podporovaná väčšinou výrobcov mobilných telefónov, čo by malo zabezpečiť relatívne širokú použiteľnosť tejto aplikácie pre koncových používateľov. Okrem syntézy reči je implementovaná aj funkcia animácie hovoriacej tváre, táto časť projektu však v súčasnosti nie je ďalej vyvíjaná.

Syntéza reči bola pôvodne vytváraná z difónovej databázy zakódovanej pomocou PCM vzoriek vo WAV súbore. Takýto formát uchovania audio signálu nám však dáva len veľmi malé možnosti na úpravu prozódie syntetizovanej reči. Preto bolo potrebné použiť iný model, ktorý by takúto modifikáciu prozódie umožňoval. Tu sa ako vhodným nástrojom ukázal sínusoidálny model, pri ktorom sú vzorky jednotlivých foném a difónov zakódované pomocou sínusoíd. Sínusoidálny model bol použitý v predchádzajúcej práci na vytvorenie nekomprimovanej parametrickej databázy a v tejto práci bola implementovaná komprimovaná verzia parametrickej databázy uchovávajúcej hodnoty sínusoíd rozšírená aj o šumovú zložku – takzvaný HNM model (Harmonic plus noise model).

Druhou a nezanedbateľnou výhodou komprimovanej parametrickej databázy je zníženie veľkosti databázy v závislosti na úrovni kompresie, ktorá ovplyvňuje následne kvalitu syntetizovanej reči.

2. Syntéza pomocou difón

Syntéza reči prebieha postupne vo viacerých krokoch, ktoré sú znázornené na Obr. 1.



Obr. 1. Bloková schéma syntézy.

Táto práca je zameraná iba na časti 3 a 4 – Syntéza a Databáza. Pôvodná databáza bola formátovaná ako množina foném/difón uložených vo zvukovom WAV súbore.

Foném je hláska s rozlišovacou platnosťou. Je to abstraktná jednotka reči, ktorou dokážeme rozlíšiť slová, t.j. zmenou jednej fonémy dokážeme vytvoriť druhé slovo. Difóna je postupnosť dvoch foném, v slovenčine je to zvyčajne postupnosť samohláska – spoluhláska. Difóne zodpovedajúci úsek reči sa rozlišuje zo stredu jednej fonémy do stredu druhej. Využívajú sa hlavne z dôvodu, že veľká časť akustickej informácie, ktorá je potrebná k rozlíšeniu spoluhlások, leží v prechodoch medzi spoluhláskou a samohláskou [1].

Pri syntéze sa potrebné PCM vzorky pre danú fonému/difónu vygenerujú z parametrickej databázy a uložia za sebou do nového WAV súboru, ktorý sa následne prehrá. V prípade, že potrebná difóna nie je v databáze obsiahnutá, tak je vytvorená spojením z dvoch nezávislých foném a to tak, že sa uložia za sebou vzorky prvej fonémy zo stredu až do jej konca a vzorky druhej fonémy od jej začiatku až do stredu. Stred fonémy nemusí byť automaticky v polovici vzoriek [2]. Difónová databáza obsahuje približne 1400 difón a všetkých 53 foném slovenského jazyka.

3. HNM model

HNM model predpokladá, že reč je zložená z harmonickej a šumovej časti. Harmonická časť odpovedá kvázi-periodickým zložkám reči a šumová časť odpovedá neperiodickým zložkám reči. Tieto dve zložky sú vo frekvenčnom spektre oddelené časovo premenlivou medznou frekvenciou F_m. Pásmo po F_m je reprezentované harmonickými sínusoidami a pásmo od F_m je reprezentované modulovanou šumovou zložkou. Neznelé časti reči sú reprezentované iba šumovou časťou. Rečový signál potom získame ako sumu harmonickej a šumovej časti $s(t) = h(t) + n(t)$.

Harmonická časť obsahuje iba harmonické násobky základnej hlasivkovej frekvencie F_0. Signál je reprezentovaný sumou sínusoíd s príslušnými frekvenciami, amplitúdami a fázami:

$$h(n) = \sum_{k=-L(t)}^{L(t)} a_k(t).e^{jk\omega_0(t)t} \quad (1)$$

kde L(t) je počet harmonických, $\omega_0(t)$ je základná hlasivková frekvencia a $a_k(t)$ je amplitúda k-tej harmonickej.

Šumová časť je modelovaná použitím energií Barkových pásiem. Tento spôsob spočíva v použití rovnakej metódy ako v harmonickej časti. Keďže šumová časť neobsahuje žiadnu základnú hlasivkovú frekvenciu, F_0 je nastavené na 100 Hz. Fázy sínusoíd sú potom náhodné pretože šum je náhodný proces.

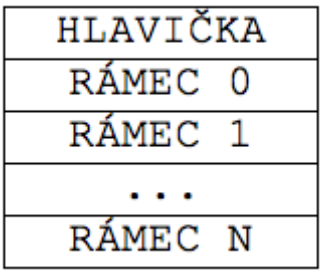
Barkove pásma sú psychoakustickou stupnicou, ktorú navrhol v roku 1961 Edmund Zwicker. Sú pomenované po Heinrichovi Barkhausen, ktorý prvý navrhol subjektívne merania hlasitosti. Je to 25 kritických pásiem počutia. Hranice týchto pásiem sú (v Hz): 0, 20, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500 [3].

Kritické pásmo je frekvenčné pásmo, v ktorom dochádza k výrazným zmenám pri subjektívnom vnímaní zvuku. Ak sú dva tóny k sebe frekvenčne blízko, výsledný tón je zlúčením týchto dvoch tónov a obe frekvencie aktivujú tú istú časť bazilárnej membrány. Ak sa tóny od seba vzdialia mimo kritického pásma, čiže dostatočne na to, aby aktivovali rozdielne časti bazilárnej membrány, počujeme každý tón samostatne.

V rámci Barkových pásiem ucho nie je citlivé na zmeny energie pre stacionárne kvázišumové signály. Za predpokladu, že rezíduum reči je podobné šumu, možno ho modelovať pomocou výpočtu krátkodobých energií v každom pásme [3]. Hodnoty energií v jednotlivých Barkových pásmach sú ďalej označené skratkou BBE (Bark band energy).

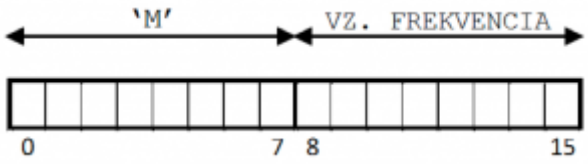
4. Komprimovaná parametrická databáza

Vzorky každej fonémy/difóny sú generované z komprimovanej parametrickej databázy. Základná štruktúra komprimovanej parametrickej HNM databázy je na Obr. 2 [4].



Obr. 2. Základná štruktúra komprimovanej databázy.

HLAVIČKA pozostáva z dvoch bajtov. Prvý bajt slúži ako identifikácia HNM databázy, druhý bajt určuje vzorkovaciu frekvenciu databázy.



Obr. 3. Hlavička databázy.

Podporované vzorkovacie frekvencie a ich identifikačné čísla v hlavičke databázy sú v Tab. 1.:

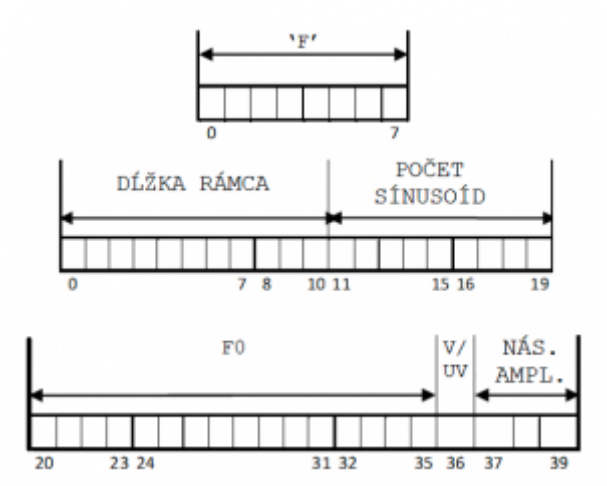
f [Hz]	ID
8 000 Hz	0
16 000 Hz	1

22 050 Hz	2
44 100 Hz	3
96 000 Hz	4

Podstatná informácia o sínusoidách je uložená v bloku rámcov (RÁMEC 1, RÁMEC 2, …). Každý rámec má dve zložky – harmonickú a šumovú.

4.1. Harmonická zložka

Hlavička harmonickej časti rámca je zároveň identifikátorom začiatku rámca a má štruktúru zobrazenú na Obr. 4 [4].



Obr. 4. Hlavička rámca.

Prvý bajt, v ktorom je uložená ASCII hodnota písmena F (0x46), identifikuje začiatok rámca. Nasledujúcich 5 bajtov pozostáva z hodnôt DĹŽKA RÁMCA – počet generovaných vzoriek, POČET SÍNUSOÍD v danom rámci, F0 základná hlasivková frekvencia, rozhodnutie o znelosti/neznelosti rámca (V/UV), a násobič amplitúdy (NÁS. AMPL.). Je potrebné poznamenať, že prirodzené čísla sú v databáze uložené vo formáte little endian.

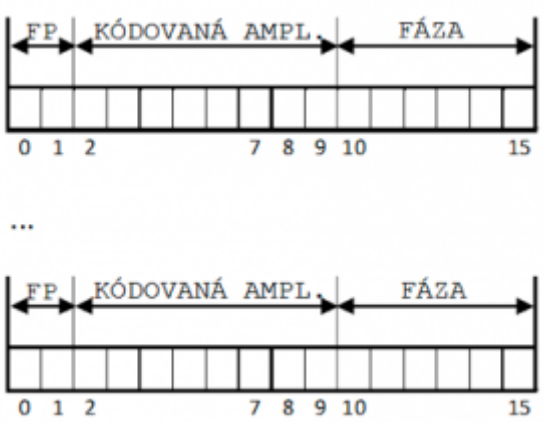
Dĺžka rámca je binárne kódovaná v 11 bitoch. Dĺžka rámca je vlastne počet vzoriek, ktoré potrebujeme generovať z každej sínusoidy a šumovej zložky v danom rámci. Počet vzoriek rámca je vždy párny, a preto je zakódovaná hodnota DĹŽKA RÁMCA polovica skutočnej dĺžky rámca. Pri dekódovaní je táto hodnota jednoducho vynásobená dvomi. Toto umožňuje použiť rámce o dĺžke až 4094 vzoriek, čo je výhodné najmä pri analýze signálov s vyššou vzorkovacou frekvenciou. V ďalších 9 bitoch je uložený počet sínusoíd daného rámca. Maximálna podporovaná hodnota je 511 sínusoíd.

Základná hlasivková frekvencia F0 je kódovaná v nasledujúcich 16 bitoch nasledovne. Predpokladáme, že detegovaná F0 je vždy v rozsahu 0 až 600 Hz, čo je pre ľudskú reč rozumné ohraničenie. Kvantizačný krok je potom $600/2^{16} = 0{,}0091553$, čo je viac než dosť pre kódovanie F0. Predpokladáme, že pre znelú reč je možné frekvencie sínusoíd získať ako harmonické frekvencie F0. To znamená, že k-ta sínusoida má frekvenciu k.F0. Pre prípad neznelého signálu opisujeme jeho spektrálnu obálku. Keďže šum nemá žiadnu základnú hlasivkovú frekvenciu tak F0 je nastavená na hodnotu 100 Hz a

frekvencie ďalších sínusoíd sú získané rovnakým spôsobom ako pre znelé signály.

Nasledujúci 1 bit označuje rámec za znelý alebo neznelý. Posledné 3 bity sú použité pre násobič amplitúdy, ktorý nám pomáha binárne kódovať desatinnú čiarku hodnôt amplitúdy. Zvyčajne sú v každom rámci desiatky sínusoíd a maximálna hodnota všetkých sínusoíd v rámci je detegovaná. Násobič amplitúdy je číslo 10^n, kde n závisí na desatinnom mieste amplitúdy, napr. ak maximálna amplitúda je 0,08, n = 1 a násobič amplitúdy je 10. Potom všetky amplitúdy v rámci sú vydelené touto hodnotou a potom zakódované.

Štruktúra týchto 6 bajtov sa objaví v každom rámci len raz, na jeho začiatku. Za touto hlavičkou nasleduje kódovanie amplitúdy a fázy jednotlivých sínusoíd v 2 bajtoch. Nasledujúca štruktúra 2 bajtov sa opakuje v každom rámci toľko krát, koľko sínusoíd je v rámci, čo je dané hodnotou POČET SÍNUSOÍD v hlavičke rámca.



Obr. 5. Štruktúra kódovania harmonickej zložky.

Prvé 2 bity označené ako FP sú určené na binárne kódovanie desatinného miesta amplitúdy. V ďalších 8 bitoch je kódovaná amplitúda. Desatinné miesto amplitúdy je upravené tak, aby žiadna amplitúda nebola väčšia ako 1, ale bola ku 1 čo najbližšie, t.j. pre hodnotu amplitúdy 0,004 upravíme desatinnú čiarku na hodnotu 0,4. Keďže amplitúda nemôže byť väčšia ako 1 je kódovaná nasledovne:

$$A = tan\left(\frac{\frac{\pi}{4}}{2^8}v\right) \tag{2}$$

kde je prirodzené číslo kódované vo 8 bitoch KÓDOVANÁ AMPL. Kvantizačný krok pri tomto prístupe je 0,0030680, čo je lepšie ako pri štandardnom kódovaní pohyblivej rádovej čiarky a dostatočne presné pre naše potreby. Fáza môže nadobúdať hodnoty 0 až 2π a je kódovaná nasledovne:

$$\varphi = \frac{2\pi}{2^6}v \tag{3}$$

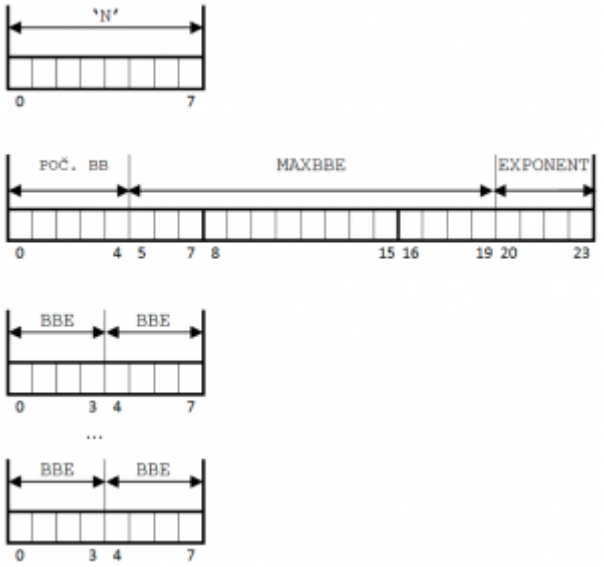
kde v je prirodzené číslo kódované v 6 bitoch FÁZA. Hlavným dôvodom kódovania fázy je lepšia kvalita syntetizovanej reči.

4.2. Šumová zložka

Hlavička šumovej zložky rámca má štruktúru uvedenú na Obr. 6 [5]. Šumová zložka

pozostáva z energie Barkových pásiem, ktoré sú počítané osobitne pre každý rámec. Podľa [6] je kódovanie energie Barkových pásiem v komprimovanej databáze nasledovné. Frekvenčné spektrum 0 – 20 000 Hz je rozdelené na 25 pásiem, v ktorých je počítaná BBE. Z vypočítaných hodnôt energií Barkových pásiem je nájdená maximálna hodnota maxBBE v danom rámci. Následne sú všetky energie Barkových pásiem normované týmto maximom. Potom je určený dekadický exponent maxBBE.

Hlavička šumovej časti rámca (‘N’) s počtom parkových pásiem (POČ. BB), maxBBE a jeho exponentom (EXPONENT) je uložená v 4 bajtoch. Za hlavičkou sú uložené všetky energie Barkových pásiem, každá v 4 bitoch. Kvantizácia energie Barkových pásiem je vcelku nepresná, avšak nie je to veľký problém, keďže ľudské ucho nemá veľké rozlíšenie vo vnímaní amplitúdy šumového signálu. Hlavným dôvodom je použiť čo najmenej bitov kvôli kompresii. Použitie 4 bitov pre každé BBE nám dáva 16 možností pre BBE, ktoré sú rozdelené do 4 skupín. Každá skupina používa svoj vlastný násobič. Algoritmus kvantizácie a spätného výpočtu hodnoty BBE je v Tab. 2.



Obr. 6. Štruktúra kódovania šumovej zložky.

Tab. 2. Kódovanie energie Barkových pásiem.

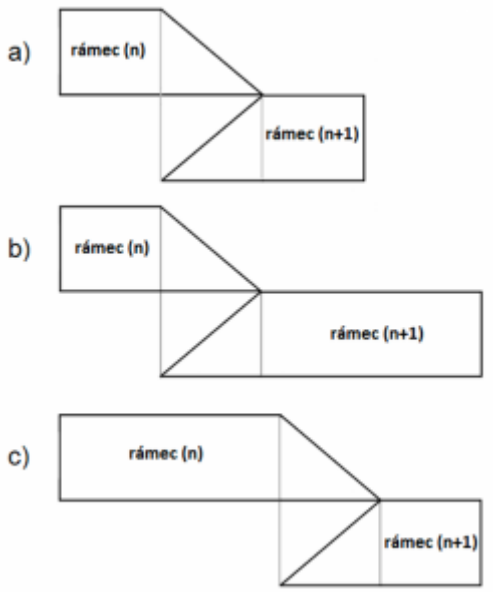
Násobič skupiny	Sekvencia	Sekvencia [v bitoch]	Výpočet BBE
$10^0 = 1$	0	0000	10^0.3/4.maxBBE
	1	0001	10^0.3/4.maxBBE
	2	0010	10^0.2/4.maxBBE
	4	0011	10^0.1/4.maxBBE
$10^{-1} = 0.1$	4	0100	10^{-1}.4/4.maxBBE
	5	0101	10^{-1}.3/4.maxBBE
	6	0110	10^{-1}.2/4.maxBBE
	7	0111	10^{-1}.1/4.maxBBE
$10^{-2} = 0.01$	8	1000	10^{-2}.4/4.maxBBE
	9	1001	10^{-2}.3/4.maxBBE
	10	1010	10^{-2}.2/4.maxBBE
	11	1011	10^{-2}.1/4.maxBBE

10^{-3} = 0.001	12	1100	10^{-3}.4/4.maxBBE
	13	1101	10^{-3}.3/4.maxBBE
	14	1110	10^{-3}.2/4.maxBBE
	15	1111	10^{-3}.1/4.maxBBE

5. Syntéza

Po zadaní textu na syntézu do aplikácie je tento text prepísaný do SAMPA abecedy a postupne sú generované PCM vzorky pre každú difónu. Vzorky pre jednotlivé difóny sa postupne uložia do výstupného súboru, ktorý je po ukončení syntézy užívateľovi prehratý. Generovanie zvukových vzoriek pre každú difónu pozostáva z niekoľkých krokov.

V prvom rade sú vypočítané hodnoty zo všetkých potrebných rámcov. Priemerne je generovaných z každého rámca okolo 200 vzoriek. Keďže však priemerný počet vzoriek difóny v databáze je okolo 1400, je potrebné generovať vzorky z viacerých rámcov. Následne sú vzorky dvoch susedných rámcov prekryté cez polovicu kratšieho z rámcov spolu s aplikovaním trojuholníkových okien. Trojuholníkové okná zabezpečia, že súčet dvoch hodnôt vzoriek nám nevyjde mimo rozsahu <-1 ; 1>. Princíp prekryvu dvoch susedných rámcov je ukázaný na Obr. 7. Všetky rámce, okrem prvého a posledného rámca každej difóny, sú prekrývané z obidvoch strán.



Obr. 7. Tri rôzne prípady dĺžok susedných rámcov. a) Obidva rámce sú rovnako dlhé, b) Nasledujúci rámec je dlhší, c) Nasledujúci rámec je kratší

Posledným krokom pred prehratím vygenerovaných vzoriek je ich prenásobenie hodnotou $2^{16}/2$ = 32768 , keďže v pôvodnej databáze boli PCM vzorky kódované do 16 bitov.

6. Záver

Cieľom mojej práce bola implementácia syntézy slovenskej reči na mobilný telefón použitím komprimovanej parametrickej databázy. Implementovanie komprimovanej parametrickej databázy rozšíri možnosti ďalšieho vylepšovania aplikácie

multimediálneho čítania SMS správ v možnosti upravovať prozódiu syntetizovanej reči. Hlavným vylepšením bude úprava prozódie úpravou parametrov sínusoíd pred generovaním vzoriek. Spôsob úpravy parametrov sínusoíd je predmetom iného prebiehajúceho výskumu na Katedre telekomunikácií. Inými zlepšeniami syntézy môže byť aj voľba správneho kompresného pomeru databázy, aplikácia iných ako trojuholníkových okien (napr. Hanningove okno) a aplikácia týchto okien aj pri prechode medzi difónami.

7. Odkazy na literatúru

1. PSUTKA, J., „Komunikace s Počítačem Mluvenou Rečí“, Academia, 1995
2. Talafová, R., „Syntéza reči v mobilnom telefóne”, Diplomová práca, Katedra telekomunikácií , FEI STU, Bratislava 2007
3. Zölzer, U., „Digital Audio Signal Processing“, Wiley, 2008, pp. 277-278
4. Rozinaj, G., Rybárová, R., Turi Nagy, M., „Sinusoidal Parametrization for Speech Synthesis in Mobile Phones“
5. Nagy, M.T., Rozinaj, G., „Compression of a Slovak Speech Database using harmonic, noise and transient model“, ELMAR 2010 Proceedings, Zadar 2010, pp. 363-366
6. [6] Turi Nagy, M., „Analýza a syntéza audio signálov pomocou SN (sinusoids + noise modeling) modelu“, Diplomová práca, Katedra telekomunikácií, FEI STU, Bratislava 2004

Spoluautorom článku je Ing. Renáta Rybárová PhD., Katedra telekomunikácií, Fakulta Elektrotechniky a Informatiky, Slovenská Technická Univerzita, Ilkovičova 3, Bratislava 812 19

Práca bola prezentovaná na Študentskej vedeckej a odbornej činnosti (ŠVOČ 2011) v sekcii Aplikovaná mechanika a získala Diplom dekana, ISBN 978-80-227-3508-7