

## Syntéza reči v IVR systémoch: návrh R&D prostredia

Miloš Cerňak

### Abstrakt

Príspevok opisuje použitie systémov syntézy reči z textu (TTS) v IVR systémoch na úrovni výskumu a vývoja. Na túto úlohu navrhuje softwarové R&D (Research & Development) prostredie, umožňujúce skúmať vzájomnú interakciu prostriedkov dialogového manažéra a prostriedkov na syntézu reči. Pre syntézu reči navrhuje použiť syntézu spájaním v limitovanej doméne.

### Kľúčové slová

syntéza reči, telekomunikácie, IVR systémy, TTS systémy

### Úvod

Syntéza reči je v súčasnosti zrozumiteľná, niektoré systémy reči (napr. NextGen od AT&T) produkujú reč takmer nerozoznateľnú od ľudskej. Najznámejšie sú systémy syntézy z textu TTS (text-to-speech), zahŕňajúce techniky spracovania prirodzeného jazyka. S rastúcou kvalitou syntézy rastú aj jej možnosti použitia v iných aplikáciách, ako napríklad v IVR (Interactive Voice Response) systémoch. Bežné súčasné IVR systémy „produktujú“ reč prehrávaním prednahratých súborov, čo ale nemá zmysel pri častej výmene hlasových promptov alebo ak systémy pracujú v neobmedzenej doméne (napr. nemôžu sa nahráť všetky mená a priezviská z telefónneho zoznamu). Preto moderné IVR systémy začínajú využívať TTS systémy.

### 1 TTS systémy

Od začiatku 90-tých rokov, keď páni Moulines E. a Charpentier F. zverejnili techniku PSOLA (*pitch-synchronous overlap-add*) [1], syntéza spájaním (*concatenative synthesis*) sa stala základom mnohých komerčných a akademických TTS systémov. Tieto systémy sa vyznačujú veľmi dobrou zrozumiteľnosťou a pri kvalitnom spracovaní prirodzeného jazyka, aj vysokou prirodzenosťou.

Na druhej strane, projekt COST 258 poukázal na to, že ďalší vývoj bude skôr v zlepšení spektrálnych techník a modelovaní štýlov a hlasov [2]. V poslednom čase sa do popredia dostáva aj konverzia (transformácia) hlasov, čo môže výrazne ovplyvniť návrh nových systémov.

S rastúcou kvalitou hlasu, rastú aj možnosti jej využitia v telekomunikačných systémoch. Výskum syntézy reči ovplyvňuje hlavne oblasti:

- kódovania signálov (reči), napr. známe vokóderové systémy
- a rozhrania človek – stroj (počítač)

Vyvinúť univerzálny TTS systém je nesmierne náročná úloha, lebo takýto systém by musel zahŕňať mnoho štýlov (čítanie e-mail správ, novin, knih, Internetu atď.) a dostatočný počet hlasov. Preto je použitie TTS systému skôr obmedzené na užšiu doménu určitého štýlu, hlasu, a dokonca aj použitého slovníka [3].

### 2 Hlasové dialógové systémy

Hlasový dialógový systém je počítačový systém, ktorý umožňuje striedavú komunikáciu užívateľa a počítača hlasom. Súčasné IVR systémy obmedzujú užívateľa na to čo má povedať a ako to má povedať. Zlepšiť to môže nie len robustejšie rozpoznávanie reči, ale aj dômyselná implementácia dialógového systému.

Hlasové dialógové systémy môžeme rozdeliť do troch hlavných kategórií [4]:

- systémy s konečným stavom
- rámcové dialógové systémy
- agentské dialógové systémy

Systémy s konečným stavom definujú konečný počet dialógových stavov a pravidiel, čo sa má vykonávať v každom stave. Dialóg sa dá reprezentovať orientovaným grafom.

Rámcové systémy definujú vzor (šablónu), ktorý sa dá prirovnáť k postupnosti slotov, postupne plnených z rozoznávania hlasového vstupu. Tok dialógu už nie je predurčený, lebo systém je závislý od naplnenia všetkých slotov.

#### Dialóg 1:

IVR: Kedy cestujete?  
Užívateľ: V piatok.  
IVR: Kam cestujete?  
Užívateľ: Do Bratislavy

Ing. Miloš Cerňak,  
Katedra telekomunikácií FEI STU,  
Ilkovičova 3, 812 19 Bratislava 1,  
t.: 02/68279 409,  
e-mail: mcernak@kti.elf.stuba.sk

## Dialóg 2:

IVR: Kedy cestujete?  
 Užívateľ: V piatok, idem priamo do Bratislavy.  
 IVR: Našiel som pre Vás tieto spojenia ...

Prvý príklad je typický dialógový systém s konečným stavom a druhý príklad predstavuje rámcový dialógový systém, ktorý po vyplnení rámca „<destinácia> <čas>“, pokračuje v dialógu.

Agentské systémy predstavujú komplexnú komunikáciu systému a užívateľa. Definujú koncept „agenta“, ktorý zodpovedne rieši danú úlohu s prihliadaním na predchádzajúci kontext a s určitým predpokladaním správania užívateľa.

## 3 IVR systémy

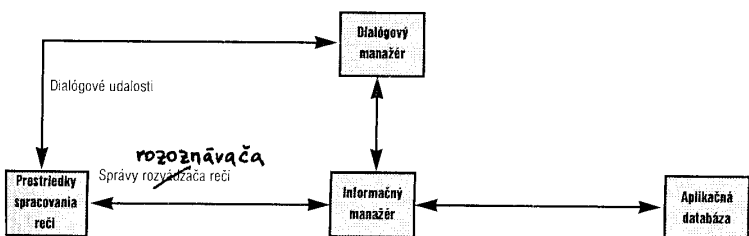
Každý IVR systém zahŕňa tri základné bloky:

- prostriedky spracovania reči
- dialógového manažera
- aplikačnú databázu

Architektúra bežného IVR systému je zobrazená na obr. 1. IVR systém podľa obr. 1 musí zabezpečovať implementáciu algoritmov rozpoznávania reči, detekciu DTMF signálov, prostriedky na nahrávanie a prehrávanie zvukových záznamov a samotné telefónne rozhranie. Súčasne



Obr. 1 Architektúra IVR systému



Obr. 2 Architektúra IVR systému s veľkým slovníkom

1 CSLU Toolkit je nástroj na výučbu a výskum „človek - stroj“ interakcie, zdroj: <http://cslu.cse.ogi.edu/toolkit/>

2 FESTIVAL, systém syntézy reči, zdroj <http://www.cstr.ed.ac.uk/projects/festival/>

3 EPOS, systém syntézy reči vyvinutý na Českej Akadémii Vied IREE na oddelení digitálneho spracovania signálov, zdroj: <http://epos.ure.cas.cz/>

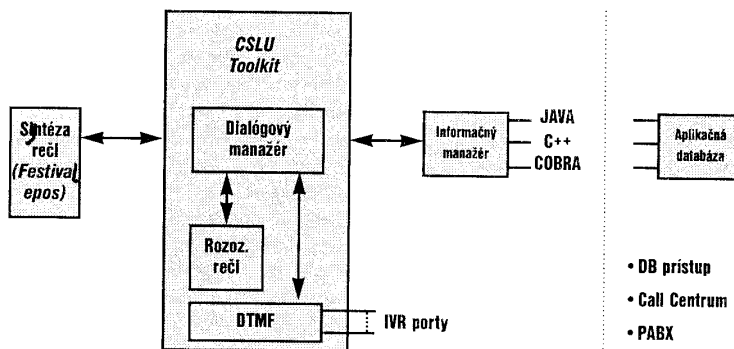
IVR servre ako telefónne rozhranie využívajú PC karty s telefónnymi portami. Pre systémy pracujúce v širokej doméne (napr. rozoznávajúce mien firmy s niekoľko tisíc zamestnancami) sa architektúra rozrastá o informačného manažera, ako je to zobrazené na obr. 2.

Informačný manažer má za úlohu odľahčiť dialógového manažera od komunikácie s aplikačnou databázou kóli

Medzi úspešné IVR s veľkým slovníkom patrí napr. produkt British Telecom CallMinder™ [5].

## 4 Návrh R&D softvérového prostredia

Navrhované riešenie spája vývojové prostredia pre syntézu reči, rozoznávajúce



Obr. 3: Architektúra systému IVR systému

narastajúcej výmene alternatívnych hypotéz medzi rozoznávateľom reči a DB.

Oproti tradičným IVR systémom, IVR s veľkým slovníkom zahŕňa:

- rozpoznávač reči s veľkým slovníkom
- dynamickú modifikáciu slovníka
- TTS systém

reči a modelovanie dialógu. V podstate, to všetko zahŕňa samotný CSLU Toolkit<sup>1</sup>, lebo je plne integrovaný s FESTIVAL-om<sup>2</sup>. Tento koncept je tu rozšírený o externé využitie iných platforiem na syntézu reči ako napr. EPOS<sup>3</sup> a vytvorenie interpretátora príkazov na komunikáciu s externými aplikáciami (komerčnou databázou, call centrom atď.)

Na obr. 3 je zobrazená architektúra systému, postavená nad Windows platformou.

Navrhované prostredie sa skladá zo štyroch hlavných blokov:

1. Prostredia pre syntézu reči (kapitola 2.1)
2. Dialógového manažera (kapitola 2.2)
3. Interpretera príkazov (kapitola 2.3)
4. Externej aplikácie

CSLU Toolkit používa DNA (Dialogic's native interface) 3.x pod Windows NT/2000 pre komunikáciu s doskou IVR portov (napr. D/21x, D/41x, Proline/2V, D/41ESC).

### 4.1 Prostredie na syntézu reči

Najjednoduchší spôsob integrácie vysokokvalitnej syntetickej reči je implementovať syntézu reči v limitovanej doméne [3] systémom FESTIVAL. Pretože dialógový manažer bol vyvinutý pre Windows plat-

formu a verzia FESTIVAL-u pod Windowsom nemá všetky potrebné vlastnosti, na zvládnutie úlohy sa musí najprv vyvinúť hlas v limitovanej doméne pod OS UNIX alebo Linux a následne sa môže „exportovať“ do Windows-u. Exportovanie hlasu je v tomto prípade veľmi jednoduché, ide iba o skopírovanie adresárovej štruktúry vyvinutého hlasu. Jediný rozdiel v jeho inicializácii, a to príkazmi

```
cd root_of_voice
festival festvox\scheme_of_voice.sch (voice)
```

Samozrejme, môže sa použiť akýkoľvek iný syntetizátor, v tom prípade bude volaný priamo z Tcl kódu (napr. príkazom `exec`) dialógového manažéra ako externý program. Podrobnejšie bude rozobraná syntéza reči v IVR systémoch v kapitole 3.

#### 4.2 Dialógový manažér a rozpoznávanie reči

Ako dialógový manažér sa tu s výhodou používa CSLU Toolkit. Je ponúkaný spolu so zdrojovým kódom, no s určitými komerčnými obmedzeniami. Ponúka grafické rozhranie vytvárania dialógu RAD (*Rapid Application Development*) a zároveň zabezpečuje rozpoznávanie hlasu a detekciu DTMF signálov. Na rozpoznávanie hlasu ponúka hybridné HMM-NN (*Hidden Markov Model - Neural Network*) riešenie, spolu s prostriedkami na vytvorenie a tréning vlastného rozpoznávača reči. Pre fonetickú transkripciu používa Worldbet [6], čo je ASCII prepis IPA (*International Phonetic Alphabet*) fonetickej znakov sady, preto je možné vytvoriť aj slovenský rozpoznávač reči.

#### 4.3 Interpreter príkazov

Pod interpreterom príkazov sa tu zahŕňa akékoľvek rozhranie smerom k externým aplikáciám. Predstavuje „most“ medzi dvoma aplikáciami - IVR a napr. databázou. Bežný príklad interpretera je Tcl rozšírenie pre databázový prístup, pre volanie JAVA metód alebo vytvorenie Tcl TCP servera pre soketovú komunikáciu s inou aplikáciou.

Toto prostredie predstavuje IVR systém, v ktorom sa TTS nie len využíva, ale umožňuje aj výskum nových TTS prístupov a algoritmov priamo v návaznosti na IVR systém. Týmto spojením sa okamžite získa prehľad o možnom použití určitej TTS aplikácie v reálnom čase s IVR systémom, celkové aplikačné pamätové ná-

roky a neposlednom rade možnosť využívať prostriedky pre syntézu reči v procese rozpoznávania reči (napr. využitie TTS pravidiel pre fonetickú transkripciu, ktorú rozpoznávač reči potrebuje pre časti slovníka, ktoré takúto transkripciu nemajú priradenú).

Architektúra navrhovaného prostredia je porovnateľná s architektúrou zobrazenou na obr.1. Aj napriek tomu, že nezahŕňa inf. manažéra, narastajúci tok dát medzi dialógovým manažérom a DB pri veľkoslovníkovom systéme sa dá zvládnuť vytvorením paralelných threadov v Tcl skriptoch. Tcl od verzie 8.2 sa radí medzi „thread - safe“ aplikácie.

Podľa rozdelenia hlasových dialógových systémov (1.2 kapitola), navrhovaný systém sa zaraďuje medzi systémy s konečným stavom, s určitými prvkami rámcových dialógových systémov.

#### 5 Syntéza reči v IVR

Syntéza reči v IVR systémoch sa môže vyskytovať od úplne jednoduchého prehrávania promptov až po všeobecné TTS systémy. Keďže väčšina IVR systémov pracuje v určitej obmedzenej oblasti (doméne), je typické použiť taký TTS systém, ktorý má v danej doméne najlepšiu kvalitu. Samozrejme, môže sa použiť aj všeobecný TTS systém, ktorý dokáže prispôbiť meniť štýl, produkovať vysokokvalitný výstup pre text; takéto systémy žiaľ v súčasnosti neexistujú. Najvyššia kvalita sa dosahuje syntézou spájaním, aplikovanou na určitú doménu.

Systém FESTIVAL má implementovanú, voľne dostupnú syntézu reči v limitovanej doméne [3], založenú na CATR (*Classification and Regression Trees*) algoritme, ktorým sa predspracovaním vytvoria zhluky jednotlivých typov segmentov, napr. foném. Medzi výhody tejto implementácie patrí krátka doba vývoja novej domény a skutočne vysoká kvalita reči. Nevýhoda je ale veľmi zlá kvalita syntézy mimo domény; problémom syntetizátorov pracujúcich v limitovanej doméne je spôsob spracovania nedoménových vstupov. Vyvinutý doménový hlas by mal mať vyvinutý aj záložný hlas (napr. všeobecný ditónový), ktorý by ho nahradzoval pre nedoménové vstupy.

Ani doménové syntetizátory však nie sú konečným riešením pre IVR systémy. Stačí aby bolo potrebné syntetizovať vlastné mená. Tu sa musia použiť iba všeobecné TTS systémy.

#### Záver

V príspevku bol predložený návrh na R&D prostredie pre IVR systémy s dôrazom na možnosti výskumu prostriedkov syntézy reči. Cieľom bolo poukázať na spôsob výskumu syntézy reči v návaznosti na aplikáciu v IVR systémoch. Navrhnuté riešenie je vhodné pre akademické využitie, hlavne pre komerčné využitie použitého dialógového manažéra. Vyvinutý syntetizátor reči však môže byť komerčne využitý v spojení s iným komerčným IVR systémom.

#### Rereferencie

- [1] CHARPENTIER, F., STELLA M. G.: Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones, Proceedings of Eurospeech, Paris, 1989, č. 2, s. 13-19
- [2] KELLER, E.: Towards Greater Naturalness: Future Directions of Research in Speech Synthesis, Improvements in Speech Synthesis (COST 258), 2002, ISBN 0471-49985-4, s. 3-17
- [3] BLACK, A., LENZO, K.: Limited Domain Synthesis, ICSLP2000, Beijing, China, 2000
- [4] MCTEAR, Michael F.: Spoken dialogue technology: Enabling the conversational user interface, Speech Technology Expert eZine, december 2001
- [5] WESTALL, F. A., JOHNSTON, R. D., LEWIS, A. V.: Speech Technology for Telecommunication, Chapman & Hall, London, BT telecommunication series 11, 1998
- [6] HIERONYMOUS, James L.: ASCII Phonetic Symbols for the World's Languages: Worldbet. Technical report, Bell Labs, 1993