■ Z výzkumných záměrů

Stvoření virtuálního dvojníka

Myšlenkou stvoření "virtuálního dvojníka" se tým na katedře kybernetiky (KKY) Fakulty aplikovaných věd ZČU zabývá již poměrně dlouhou dobu. Katedra kybernetiky je známa jako pracoviště, které se zpracování zvukové a vizuální informace věnuje komplexně a zejména v oblasti počítačového zpracování přirozené lidské řeči má bohaté zkušenosti. Po uspokojivém vyřešení základních úloh z této oblasti, jakými jsou např. automatické vytváření řeči (tzv. syntéza řeči), rozpoznávání řeči či identifikace a verifikace mluvčího, se tak dalším logickým krokem stalo jejich další využití v aplikovaných úlohách automatizace komunikace člověk-stroj. Jednou z aplikací zmíněných technologií je právě "virtuální dvojník" – počítačová kopie reálného člověka.

"Pro vytvoření mluvícího virtuálního dvojníka je potřeba kamera, systém zrcadel, mikrofon a hlavně dobrý počítačový program." Těmito slovy začal redaktor České televize Vladimír Kořen jeden ze série pořadů České hlavy, odvysílaný v září tohoto roku. Pojďme se nyní podívat, co se v tomto případě skrývá pod poněkud zjednodušujícím pojmem počítačový program. Takový program, nebo spíše soubor programů, realizuje takzvanou audiovizuální syntézu řeči. Ta se skládá ze dvou základních komponent: vizuální a akustické složky. Vizuální složku tvoří trojrozměrný model lidské hlavy s parametricky animovatelnými pohyby úst i dalších částí obličeje (oči, obočí, tváře, atd.). Úkolem akustické složky je vytvářet vlastní řeč. Obě složky jsou pak synchronizovány tak, že pohyby modelu hlavy vytvářejí dojem vyslovování právě slyšených hlásek virtuální hlavou.

Jako trojrozměrný model hlavy je samozřejmě možné použít nějaký obecně použitelný model hlavy, který je například dodáván s programovým vybavením pro trojrozměrné modelování. Snahou ovšem je přizpůsobit model tak, aby co nejvíce připomínal skutečnou hlavu, nejlépe nějakého konkrétního člověka. Prvním krokem k vytvoření počítačového virtuálního dvojníka je tedy vytvoření trojrozměrného modelu na základě snímání skutečné hlavy. Pro získávání trojrozměrných dat je zapotřebí pohledu alespoň ze dvou míst. Tento problém se často řeší použitím dvou kamer. Přitom je třeba řešit problém synchronizace těchto dvou kamer tak, aby oba stereoskopické snímky byly sejmuty opravdu v jednom okamžiku. Na katedře kybernetiky byl vyvinut sytém využívající pro získání dvou pohledů pouze jednu kameru a systém čtyř zrcadel. Dalším problémem je korespondence bodů v obou stereoskopických obrazech při vypočítávání trojrozměrných souřadnic. Ten je v našem systému vyřešen použitím strukturovaného osvětlení. Obraz je snímán v zatemněné místnosti, kdy na statický obličej je promítán svislý proužek vytvářený datovým projektorem. Výsledkem asi půlminutového snímání touto unikátní technologií je realistický trojrozměrný model odpovídající tvarem a texturou skutečné hlavě.

Pro zachycení pohybu úst pro animaci je zapotřebí zaznamenat pohyby úst konkrétního řečníka. Text pro nahrávání dynamických dat pohybu úst je třeba navrhnout tak, aby obsahoval pokud možno všechny vizuální projevy hlásek v daném jazyce, nejlépe ještě v různých kombinacích okolních hlásek, nebo koartikulace (vzájemné ovlivňování sousedních hlásek) výrazně ovlivňuje vizuální projev vzhledem k velké setrvačnosti artikulačních orgánů. Data jsou pak snímána podobným způsobem jako v předchozím případě, tj. pomocí kamery a systému čtyř zrcadel. Nicméně vzhledem k dynamice dat nelze použít systém proužkového osvětlení. Proto jsou na obličej snímané osoby rozmístěny reflexivní značky a celé snímání je navíc prováděno v infračervené oblasti spektra s využitím speciálního osvětlení. Tato data jsou poté zpracována, jsou zaznamenány časové průběhy trojrozměrných souřadnic řídicích bodů, a tvoří tak databázi vizuálních projevů základních řečových jednotek. Syntéza je pak prováděna tak, že se skládají průběhy po sobě jdoucích řečových jednotek podle zadaného textu, počítají souřadnice řídicích bodů pro každý snímek sekvence a na základě nich se vykresluje trojrozměrný model. Vzhled trojrozměrného modelu je pak dán těmito řídicími body, jejichž změna způsobí například otevření úst do správné polohy, změnu pozice brady nebo pozvednutí obočí. Vizuální část je pak synchronizována s akustickou tak, aby pohyby úst odpovídaly právě vyslovovaným hláskám v akustickém signálu.

Akustická složka audiovizuální syntézy řeči má za úkol vytvářet řeč, a to v takové formě a kvalitě, aby co nejvěrněji kopírovala řečové charakteristiky konkrétního člověka; tedy nejen samotný hlas a jeho kvalitu, ale i styl mluvení atd. Jde o časově nejnáročnější část tvorby virtuálního dvojníka. K automatickému vytváření řeči se využívá technologie syntézy řeči z textu (z anglického text-to-speech, TTS) – nejobecnější a také nejtěžší úloha syntézy řeči, jejímž úkolem je převést libovolný text na odpovídající řeč. Díky technologii TTS "může" náš virtuální dvojník "ozvučit" libovolný text - tj. může vyslovit libovolnou promluvu. Jde o sadu speciálních modulů a algoritmů, které zajišťují automatický převod psaného českého textu na mluvenou řeč. Zahrnují tak zpracování textu (např. analýza a normalizace), převod textu do výslovnostní podoby (tj. fonetickou transkripci a generování průběhů prozodických vlastností řeči), tvorbu inventáře akustických jednotek a vlastní metodu pro vytváření řeči.

Pro potřeby českého TTS systému jsme vyvinuli unikátní metodiku vysoce kvalitní syntézy řeči. Systém je založen na tzv. konkatenační syntéze řeči, v současné době celosvětově nejúspěšnějším a nejpoužívanějším přístupu k syntéze řeči. Stručně řečeno, základním principem tohoto přístupu je reprezentace důležitých akustických událostí lidské řeči pomocí tzv. řečových jednotek či segmentů řeči. Výsledná řeč pak vzniká konkatenací, tj. řetězením těchto řečových jednotek. Vhodnými řečovými jednotkami jsou přitom jednotky subslovní, např. hlásky (nejčastěji posazené do kontextu okolních hlásek – tzv. trifóny) nebo difóny (zjednodušeně řečeno jde o jednotky začínající v polovině jedné hlásky a končící v polovině hlásky následující). Klíčem k úspěšné syntéze řeči je pečlivá příprava inventáře řečových jednotek – tj. segmentů řeči, s kterými syntetizér řeči pracuje. Protože kvalita výsledné syntetické řeči do značné míry závisí na bohatosti řečových segmentů obsažených v inventáři a přesnosti, s jakou jsou tyto segmenty extrahovány z referenčních promluv, navrhli jsme novou metodiku automatické konstrukce inventáře na základě velkého množství reálných řečových promluv. Automatizace je důležitým aspektem našeho systému, nebo umožňuje v krátkém časovém horizontu (řádově dny) vytvořit velice precizní a akusticky a lingvisticky "bohaté" (je možné použít obrovské řečové korpusy - desítky hodin řeči) inventáře akustických jednotek, které pak do značné míry přispívají k vysoké kvalitě vytvářené řeči. Jde o tzv. korpusově orientovanou konkatenační syntézu řeči, nebo právě řečový korpus (tj. sada reálných řečových promluv vyslovených jedním řečníkem, jehož hlasem pak syntetizér řeči mluví, a jejich reprezentace v ortografické, fonetické, spektrální či prozodické oblasti) je základním materiálem pro vytvoření inventáře řečových jednotek. Náš systém je jediným syntetizérem řeči v ČR, který tuto technologii nové generace využívá. Významným kritériem kvality je přirozenost vytvářené syntetické řeči. Přirozenost řeči přitom do značné míry závisí na kvalitě modelování tzv. prozodických charakteristik řeči (zjednodušeně řečeno popisují vývoj melodie promluvy a hlasitost a trvání jednotlivých segmentů řeči). Pro náš systém jsme navrhli unikátní metodu modelování a výběru přirozených průběhů prozodických vlastností řeči opět extrahovaných z reálných řečových promluv.

Obrázek hlavy reálné osoby a vytvořeného virtuálního dvojníka.

V dnešní době, charakteristické rychle se rozvíjejícími multimediálními technologiemi, se nabízí celá řada aplikací zde popisované audiovizuální syntézy řeči. Možné je i využití jejích komponent odděleně – akustická složka například v podobě obecného TTS systému nachází široké uplatnění v hlasových dialogových systémech od různých telefonních služeb typu automatického čtení SMS zpráv až po čtení e-mailů, elektronických dokumentů nebo knih apod. Na ZČU je například provozován hlasový dialogový telefonní systém (http://voice.zcu.cz) poskytující informace o výsledcích přijímacího řízení na Západočeskou univerzitu v Plzni. O kvalitě syntetické řeči vytvářené naším systémem svědčí řada aplikací, vytvářených ve spolupráci s průmyslovými partnery a společenskými organizacemi. Za všechny jmenujme např. službu Voice SMS, zajišťující čtení SMS zpráv zaslaných ze všech GSM sítí na pevnou telefonní linku. Připravuje se také nová služba Voice MMS, zajišťující zaslání textové zprávy ve formě videosouboru se záznamem promluvy zasílaného textu vybraným virtuálním dvojníkem. Neocenitelné služby může náš syntetizér řeči poskytnout handicapovaným lidem: němí lidé nebo lidé s poruchami hlasu mohou využívat svůj "osobní" TTS systém pro generování řeči, lidé, kteří ztratili řeč například po mozkové mrtvici, mohou využít technologii založenou na virtuálním dvojníkovi pro výuku řeči, lidem s poruchami sluchu může virtuální dvojník pomoci porozumět slyšenému textu, nevidomým lidem zase technologie TTS předčítá text, a pomáhá jim tak přistupovat k textovým informacím. Ačkoliv byl náš syntetizér řeči původně samozřejmě vyvíjen pro syntézu české řeči (v současné době náš syntetizér mluví ženským a mužským českým hlasem), postupy navržené pro syntézu řeči se ukázaly být natolik obecné, že jsme je použili i pro vytvoření syntetizéru dalších jazyků – konkrétně slovenštiny a němčiny. Plánujeme i syntézu dalších jazyků. V budoucnosti se chceme mj. zabývat i syntézou expresívní či spontánní řeči. Půjde tak o další krok směrem k vytváření naprosto přirozené řeči, která bere v potaz i např. emocionální stav řečníka.

Projektem vytvoření virtuálního dvojníka se zabývá na katedře kybernetiky poměrně velký tým. Vizuální složkou a souvisejícími úlohami (audiovizuální rozpoznávání řeči, počítačové vidění v průmyslu i lékařství apod.) se zabývá tým pod vedením Miloše Železného, Ph.D., složený ze tří doktorandů (Ing. Petr Císař, Ing. Zdeněk Krňoul, Ing. Pavel Campr) a řady studentů magisterského studia. Akustickou syntézou řeči a s ní spojenými úlohami (např. modelováním prozodických vlastností řeči, transformacemi hlasu apod.) se zabývá tým pod vedením Jindřicha Matouška, Ph.D., čítající vědeckovýzkumné pracovníky (Daniel Tihelka, Ph.D.), doktorandy (Ing. Jan Romportl, Ing. Zdeněk Hanzlíček) i několik studentů magisterského studia. Výsledky výzkumu jsou publikovány v odborných časopisech (například prestižní impaktovaný časopis Signal Processing) a prezentovány na prestižních odborných konferencích (jako např. EUROSPEECH, ICSLP, AVSP apod.), kde se setkávají s kladným ohlasem od odborníků z celého světa. Tyto úspěchy vyústily i v navázání řady mezinárodních kontaktů. Od prosince 2004 je katedra kybernetiky členem evropské sítě excelence v oboru multimodální komunikace člověk-stroj SIMILAR NoE. V rámci tohoto projektu došlo k výměnným stážím pracovníků katedry kybernetiky s kolegy z Institutu informatiky a automatizace Ruské akademie věd v Sankt-Petěrburgu. K dalším výměnným stážím došlo mezi katedrou kybernetiky ZČU a Centrem výzkumu řečových technologií na univerzitě v Edinburghu. K dalším významným kontaktům lze zařadit kontakty na pracoviště zabývající se podobnými technologiemi na univerzitách ve Vancouveru (Kanada), Santa Cruz (USA) nebo Louvain (Belgie). Jak ukázal zájem reportérů České televize o zařazení virtuálního dvojníka do série pořadů České hlavy, je tato oblast výzkumu zajímavá i pro laickou veřejnost a určitě se ještě s virtuálním dvojníkem při popularizaci české vědy setkáme. Více informací o projektech oddělení umělé inteligence na katedře kybernetiky FAV lze nalézt na stránkách http://ui.zcu.cz.

Miloš Železný, Ph.D.,
Jindřich Matoušek, Ph.D.,
KKY FAV