INSPO – informační a komunikační technologie pro osoby se specifickými potřebami
12. března 2016, Kongresové centrum Praha

ŘEČOVÁ SYNTÉZA PRO HLASOVĚ POSTIŽENÉ

Daniel TIHELKA[1], Markéta JŮZOVÁ[1], Jindřich MATOUŠEK[1], Barbora ŘEPOVÁ[2]

[1] Západočeská univerzita v Plzni, {dtihelka,juzova,jmatouse}@kky.zcu.cz

[2] Fakultní nemocnice v Motole, barbora.repova@fnmotol.cz

Anotace: *Příspěvek popisuje společnou aktivitu dvou specializovaných pracovišť, výzkumného centra „NTIS" (Nové technologie pro informační společnost) Fakulty aplikovaných věd Západočeské univerzity v Plzni (ZČU) a Kliniky otorinolaryngologie a chirurgie hlavy a krku 1. Lékařské fakulty Univerzity Karlovy a FN v Motole, v oblasti výzkumu možností náhrady hlasu pro hlasově postižené.*
Text rekapituluje současné možnosti náhrady hlasu a představuje nový způsob – „řečovou protézu", která využívá technologii syntézy řeči. Zvláštní pozornost je věnována možnostem personalizace syntézy řeči, tj. nastavení řečového syntetizéru tak, aby uměl mluvit hlasem konkrétního člověka.

Úvod

Lidská řeč patří mezi dovednosti, které denně využíváme, a považujeme ji proto za nedílnou součást našeho života. Jedná se o zvukový projev člověka, který slouží ke komunikaci s okolím a současně tvoří důležitou součást osobnosti každého z nás. Bohužel se může stát, že vlivem onemocnění dojde k výraznému zhoršení schopnosti hlas používat, či dokonce k úplné ztrátě hlasu. Jedním z důvodů takové ztráty je i diagnóza rakoviny hrtanu a následná operace – *totální laryngektomie* (*LET*), během níž je pacientovi odstraněna část hrtanu, a to včetně hlasivek. Dýchací cesty jsou pak vyvedeny na krk v podobě tzv. stomie – „slavíka", pacient tedy následně nemůže normálně mluvit. Tato radikální operace sice zachraňuje životy pacientů a jako léčba je velmi efektivní, výrazně ale snižuje kvalitu jejich života.

Tento článek se bude kromě představení současných možností náhrady hlasu věnovat podrobnějšímu popisu tvorby a použití technologie *řečové syntézy*, která se může stát další, běžně používanou alternativou pro ty, kteří ztratili schopnost mluvit. Protože je často pro pacienty velmi těžké, že nemohou se svými blízkými hovořit jejich vlastním hlasem, snažíme se jim tuto možnost nabídnout formou *personalizované řečové syntézy*, tedy syntetického hlasu, který by zněl stejně, jako jejich původní vlastní hlas (Hanzlíček, Romportl, Matoušek, 2013). Výzkum tvorby personalizované řečové syntézy spadá pod projekt HCENAT (*Naturalness in Human Cognitive Enhancement*).

Možnosti náhrady funkce hlasu

Protože provedením totální laryngektomie ztrácí pacient definitivně schopnost vytvářet hlas, existuje řada pomůcek a metod, díky nimž pacienti alespoň částečně nahrazují svoji ztracenou řeč a které jim pomáhají komunikovat.

Jícnový hlas

Jedná se o metodu využívající zbytky tkáně (slizniční řasy), které po výkonu nad hrtanem zůstaly. Pacient polkne vzduch a poté ho zpětně vypouští – říhá, čímž rozvibruje slizniční řasy a ty vydávají zvuk. Slova jsou pak artikulována rty a jazykem jako při normální řeči. Udává se, že jícnovým hlasem se však naučí mluvit maximálně 10-20 % pacientů po laryngektomii, neboť jeho efektivní používání je poměrně složité a vyžaduje dlouhodobou rehabilitaci. Řada pacientů se jícnový hlas není schopna naučit tak, aby s ním mohla běžně komunikovat. Dalším omezením je, že najednou lze polknout maximálně 150 ml vzduchu, který pak stačí asi na pět slov, pacient tedy nemůže hovořit v dlouhých větách. Na druhou stranu nevyžaduje tento způsob tvorby hlasu žádné pomůcky a obě ruce tak zůstávají volné.

Elektrolarynx

Elektrolarynx je elektrický přístroj s vibrační membránou, která se přiloží zvenku na kůži spodiny dutiny ústní, rozvibruje svaly jazyka a jazyk samotný a tím se vzduch v dutině rozkmitá se stejnou periodou, s jakou kmitá membrána. Vzniklý zvuk je pak artikulačními svaly dále modelován do jednotlivých slov. Nácvik použití elektrolaryngu není jednoduchý a při přiložení přístroje na nevhodné místo není výsledek srozumitelný, nicméně je tato metoda komunikace jistě jednodušší než jícnový hlas. Nevýhodou elektrolaryngu je to, že musí být přiložen ke krku, čili je tím zaměstnána jedna končetina. Existují i „hands-free" elektrolaryngy, ty jsou ale

v ČR omezeně dostupné. Vydávaný zvuk všech těchto zařízení zní velmi elektronicky a nepřirozeně. Zřejmě i to přispělo k faktu, že celosvětově zájem o elektrolaryngy upadá.

Hlasová protéza

Pojmem hlasová protéza se označuje membrána, která je umístěna do uměle vytvořené tracheooesophageální píštěle (spojení původního hrtanu s hltanem), tedy do horní části stomatu u pacienta po laryngektomii. Membrána je při normálním dýchání uzavřená, ale při zakrytí stomatu a usilovném výdechu jde vzduch právě skrz tuto membránu a vytváří tak zvuk. Stejně jako kmity elektrolaryngu je pak modelován v řeč svaly jazyka, rty a rezonančními prostory dutin. Hlas zní poměrně přirozeně, původnímu hlasu pacienta se ale podobá pouze částečně. K jeho vytvoření je opět zapotřebí použít jednu končetinu k uzavření stomatu. Existují i „hands-free" filtry, pojišťovny v ČR je ale v plné výši neproplácejí a jejich používání je nákladné. Zavedení hlasové protézy se provádí přímo při operačním výkonu, nebo po zhojení po radioterapii. Rozhodnutí o možnosti zavedení je vždy na operatérovi, v posledních letech je ale její používání stále častější a v současné době je standardem léčby. Nevýhodou metody je nutnost každodenního čištění protézy kartáčkem a každoroční pravidelná výměna hlasové protézy při ambulantní kontrole.

Syntéza řeči a možnosti jejího využití

Další alternativou pro pacienty po totální laryngektomii se může stát již zmíněná řečová syntéza, tedy automatický převod psaného textu na mluvenou řeč. Tu zde nazýváme pojmem „řečová protéza", aby nedošlo k záměně s „hlasovou protézou".

V současné době lze systémy syntézy řeči (systémy TTS, angl. te*xt-to-speech*) používat na běžných mobilních zařízeních, jako jsou mobilní telefony, tablety a notebooky. Syntetická řeč již také zdaleka nezní roboticky jako v počátcích jejího výzkumu, ale svou kvalitou se velmi přiblížila skutečné řeči člověka; i přes občasné nespojitosti lze někdy jen stěží rozeznat, zda se jedná o přirozený hlas nebo hlas syntetický. Nabízí se tak široké možnosti jejího uplatnění, např. v dialogových systémech telefonních automatů, ke čtení emailů či sms zpráv v mobilu během jízdy autem či pro automatické hlášení informací na nádražích.

Syntézu řeči lze ale také uplatnit v různých aplikacích a pomůckách pro hendikepované. Lidé se zrakovým postižením ji mohou využívat ke čtení obrazovky (tzv. *screen readery*), což jim umožňuje používat chytré mobilní telefony či pracovat na počítači. Snadným a rychlým způsobem lze převést na zvukovou stopu knihy a časopisy a rozšířit tak knihovnu pro nevidomé. Sluchově postiženým je nabízena možnost poslechu neutrální, srozumitelné a akusticky čisté syntetické řeči před zašumělou, dynamickou přirozenou řečí ve filmech a seriálech. A konečně pacienti s poškozeným hlasem i ti, kteří svůj hlas již nenávratně ztratili, mohou komunikovat s okolím pomocí syntézy textu napsaného na mobilním zařízení.

Používané metody syntézy řeči

Pro pochopení dalších odstavců je třeba objasnit pojem *řečová jednotka*. Tímto pojmem se označuje část řečového signálu nesoucí jednoznačnou, mezi různými jednotkami vzájemně nezaměnitelnou informaci. V systémech syntézy řeči se používají krátké řečové jednotky přibližně o velikosti hlásky, což umožňuje syntetizovat jakýkoliv (i třeba nesmyslný) text. Příkladem řečové jednotky je *foném* (přibližně odpovídá hlásce), nejmenší část řeči označující určitý zvuk. Často používanou jednotkou je také *difón*, který začíná uprostřed jednoho fonému a končí uprostřed fonému následujícího.

Uveďme nyní stručně tři běžně používané metody syntézy řeči, které jsou vyvíjeny i v centru NTIS a které jsou také využívány pro tvorbu již zmíněné personalizované syntézy. Jedná se o dvě metody založené na signálovém přístupu (tzv. *konkatenační syntéze řeči*), jejichž základním principem je práce přímo se zvukovými signály použitých jednotek, a jednu metodu s modelovým přístupem. (Matoušek, Tihelka, Romportl, 2006)

Konkatenační syntéza řeči s jedním reprezentantem

V případě konkatenační syntézy řeči s jedním reprezentantem je pro každou jednotku v daném kontextu z nahraných vět vybrán „průměrný" reprezentant. Při samotné syntéze se pak tito reprezentanti řetězí za sebe, čímž vznikne syntetizovaná věta. Poté je třeba dalšími postupy a metodami ještě upravit zvukový signál posloupnosti jednotek tak, aby se v něm objevil např. pokles intonace na konci oznamovací věty, stoupání na konci otázky apod.

Výhodou této metody je velmi rychlá odezva; hojně se využívala v minulém století, protože potřebuje minimum zdrojů použitého zařízení. Kvůli modifikacím je však kvalita výsledné řeči horší v porovnání s následující metodou, zní více roboticky.

Syntéza řeči výběrem jednotek („unit selection")

Jedná se také o konkatenanční metodu syntézy řeči, ale v tomto přístupu se při syntéze využívají všichni reprezentanti všech nahraných jednotek. Při syntéze konkrétní věty se pak hledá optimální posloupnost těchto

reprezentantů tak, aby na sebe dobře navazovali a aby vybraná posloupnost nejlépe reprezentovala požadovaný zvuk. Ten pak vznikne pouhým zřetězením řečových signálů vybraných reprezentantů jednotek. Pokud je posloupnost jednotek dobře vybrána, není třeba dělat žádné modifikace, které by výslednou syntetickou řeč už jen zbytečně zhoršili, podobně jak tomu je u konkatenančí metody s jedním reprezentantem.

Metoda syntézy řeči výběrem jednotek je v dnešní době považována za nejlepší z hlediska kvality syntetizované řeči a její podobnosti se skutečnou řečí člověka. Její kvalita ale není konstantní, závisí vždy na vhodnosti vybrané posloupnosti reprezentantů a jejich zřetězení. Může se stát, že určité spojení nezní úplně přirozeně (přeskočení hlasu, zazpívání apod.). Tyto lokální problémy v syntetických větách se nazývají *řečové artefakty*. Pro minimalizaci jejich výskytu je vhodné mít k dispozici velké množství reprezentantů (tzn. velké množství nahraných vět) a dobře nastavená kritéria jejich výběru, což je úloha, která stále není celosvětově vyřešena a je tedy předmětem intenzivního výzkumu.

Nevýhodou této metody je, že potřebujeme mít na používaném zařízení k dispozici velkou databázi reprezentantů všech jednotek a zařízení musí být současně schopno provádět velké množství výpočetních operací při hledání optimální posloupnosti. S určitými úpravami databáze (vyloučením málo používaných jednotek) a výrazným zpřísněním kritérií výběru při hledání optimální posloupnosti lze však např. používat tuto metodu i na běžně dostupných mobilních telefonech, a to bez dlouhé prodlevy mezi napsáním syntetizovaného textu a vygenerováním příslušné promluvy. (Tihelka, Kala, Matoušek, 2010)

Syntéza řeči metodou statistického modelování (HMM syntéza)

Metoda parametrické stochastické syntézy je založena na popisu vlastností řečového signálu pomocí speciálních statistických modelů, tzv. skrytých Markovových modelů (HMM). Parametry modelů pro jednotlivé řečové jednotky jsou získány analýzou trénovacích řečových nahrávek, stejných jako u předchozích metod. Ve fázi syntézy je vytvořena posloupnost modelů řečových jednotek odpovídajících syntetizovanému textu a podle těchto modelů je generována výsledná řeč. Hlavní výhodou této metody je konstantní (i když nižší) kvalita výsledné syntetické řeči. Mezi další pozitiva patří i možnost vytváření nových hlasů – úpravou (transformací) parametrů statistických modelů, to vše s využitím menšího množství řečových dat, podle kterých jsou nalezeny odpovídající transformační vztahy. (Hanzlíček, 2010)

Výhodou oproti metodě syntézy řeči výběrem jednotek je menší velikost databáze obsahující pouze modely pro daný hlas. Naopak hlavním úskalím metody je řada složitých operací zahrnujících zpracování a analýzu řečového signálu, statistické modelování i rekonstrukci výsledného řečového signálu. Právě důsledkem těchto operací je celková kvalita syntetizované řeči o něco nižší, než v případě syntézy řeči výběrem jednotek.

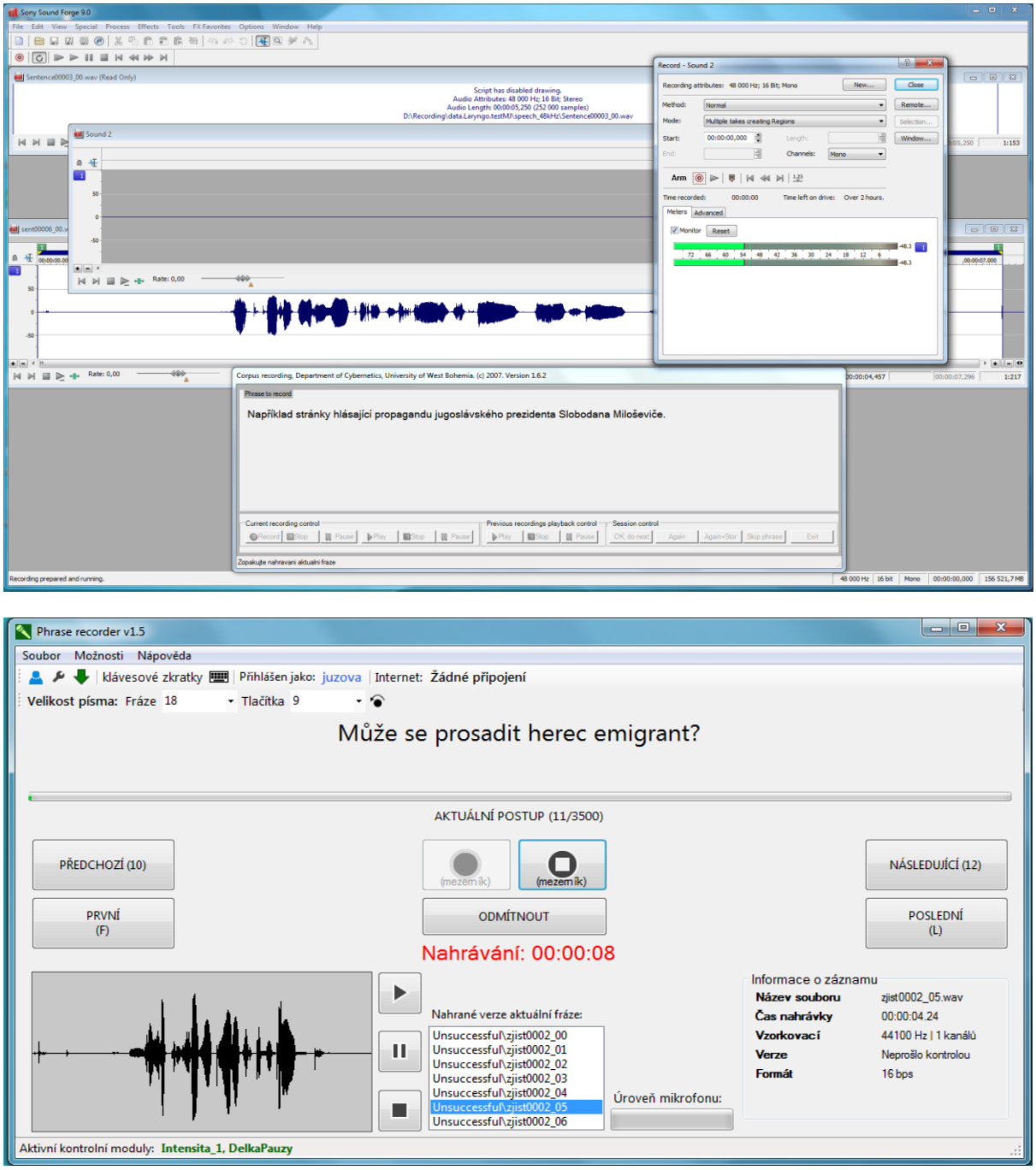
Postup vzniku personalizované syntézy

Stejně jako při tvorbě komerčního profesionálního hlasu je i při tvorbě personalizované syntézy pro pacienty s hlasovým postižením zapotřebí pořídit nahrávky. Věty na nahrávání korpusů pro účely syntézy řeči jsou obecně vybírány tak, aby obsahovaly všechny používané řečové jednotky ve všech prozodických a fonetických kontextech, viz (Matoušek, Romportl, 2006). Běžně se generuje 10-12 tisíc delších vět (často složitých souvětí), aby řečový korpus obsahoval velké množství reprezentantů každé jednotky. Profesionální řečník pak tyto vybrané věty postupně předčítá, k čemuž slouží speciální nahrávací program upravený tak, aby byly věty nahrávajícímu postupně předkládány. Součástí programu jsou i různé kontroly nahrávek, např. kontrola hlasitosti, dodržování krátkých pauz na začátku a konci každé věty, kontroluje se také konzistence nahrávek (rychlost řeči, výška, barva apod.), která je pro syntézu řeči velmi důležitá. Nahrávání vždy probíhá v odhlučněné místnosti, aby byla zaručena co nejlepší kvalita pořízeného řečového záznamu.

V případě nahrávání pacientů jsme byli nuceni tento proces mírně modifikovat. Protože se většinou jedná o neprofesionální řečníky, vytvořili jsme novou sadu vět speciálně vybraných tak, aby nebyly moc dlouhé (max. 8 slov) a aby neobsahovaly dlouhá cizí slova. Algoritmus výběru vět je podrobně popsán v (Jůzová, Romportl, Tihelka, 2015). Vzniklá sada obsahuje 3500 vět. Kvůli často velmi krátkému časovému intervalu mezi diagnózou rakoviny hrtanu a radikální operací bývá ale na nahrávání málo času, a proto jsou věty navrženy tak, aby bylo možné hlas vytvořit i v případě, že nebudou k dispozici všechny. Stále samozřejmě platí, že čím více vět řečník přečte a čím vyšší kvalita a konzistence nahrávek je dosažena, tím kvalitnější repliku svého hlasu pak má pacient k dispozici. Proto říkáme pacientům, aby nahráli alespoň několik set vět, i přes časové a zdravotní omezení. Kvůli unavenosti hlasu při mluvení a mnohdy i bolestem děláme během nahrávání časté pauzy, aby si mohli odpočinout a my tak zajistili alespoň částečnou konzistenci získaných řečových dat.

Pro zajištění dobré kvality nahrávek i v tomto případě doporučujeme pořizování záznamu v odhlučněné místnosti (ve zvukové komoře v areálu ZČU v Plzni nebo v Motole v Praze). Teoreticky je možné i nahrávání v domácím prostředí pacienta, s touto možností však teprve experimentujeme. Pro účely tvorby personalizované syntézy řeči byl v rámci projektu HCENAT vytvořen také nový nahrávací program, který je jednoduchý na ovládání a umožňuje snadnou obsluhu nahrávacího procesu i lidem, kteří nejsou příliš zdatní v používání

počítače. Jeho součástí jsou samozřejmě i různé kontroly, které jsme převzali z „velkého“ softwaru. Srovnání náhledů uživatelských rozhraní obou programů můžete vidět na obr. 1.



Obr. 1: Srovnání nahrávacích programů: vlevo - program používaný pro nahrávání profesionálních řečníků, vpravo - nově vytvořený program pro nahrávání pacientů

Pořízené nahrávky je třeba zpracovat. Velmi důležitá je v tomto případě (nahrávající není profesionální řečník) pečlivá ruční anotace – tedy kontrola, zda pacient přečetl předloženou větu správně; v případě rozdílů je třeba text věty upravit tak, aby odpovídal nahrávce. Označují se také přeřeky, zadrhávání, opakování slov a neřečové události jako nádechy, odkašlání apod.

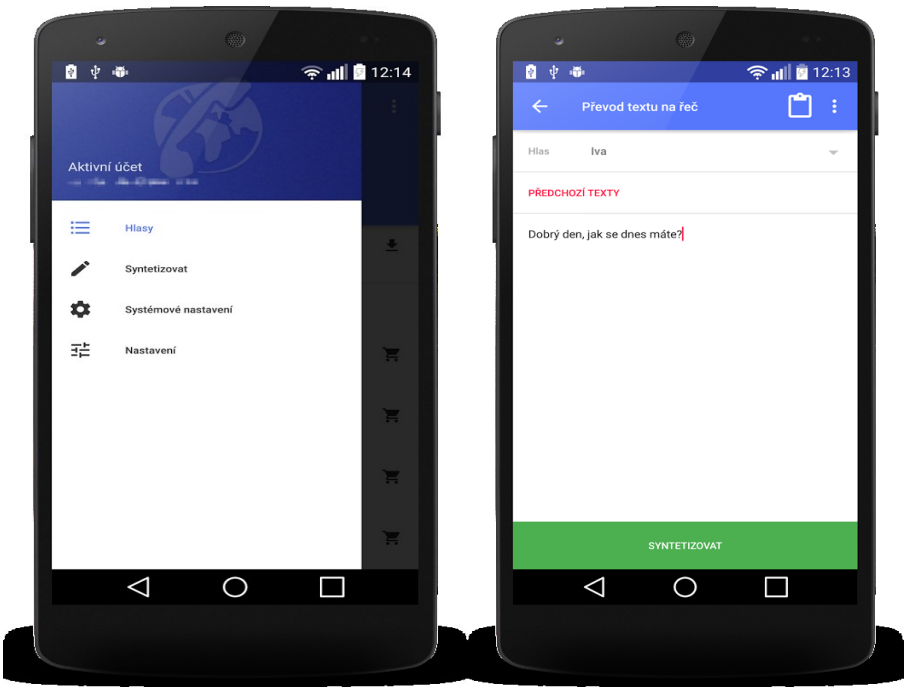
Zkontrolované nahrávky jsou pak dále zpracovávány standardními technikami používanými k tvorbě dat pro systémy TTS, řada dílčích úloh ale stále vyžaduje velké množství ruční práce. Mezi jednotlivé kroky zpracování patří např.

- hledání hranic řečových jednotek (tzv. *segmentace*)
- počítání statistických modelů pro HMM syntézu
- počítání příznaků používaných pro výběr jednotek
- tvorba databáze řečových jednotek

Posledním krokem je pak aplikace jedné z metod syntézy řeči a vytvoření syntetického hlasu. Celý proces zpracování a tvorby hlasu trvá zpravidla 2-3 týdny, v závislosti na počtu pořízených nahrávek. Výsledkem je datový soubor obsahující databázi řečových jednotek nebo jejich modelů a další data vyžadovaná pro chod systému TTS. Tento soubor je pak distribuován pacientům a ti ho mohou používat v mobilních zařízeních či na osobním počítači pro účely syntézy řeči svým vlastním hlasem.

Používání „řečové protézy“

Pokud chce pacient využít námi nabízenou alternativu existujících náhrad hlasu, může k tomu využívat svůj mobilní telefon, tablet, notebook nebo počítač. Do zařízení si nahraje svůj syntetický hlas a např. pomocí aplikace vyvinuté společností *SpeechTech, s.r.o.* (obr. 2, odkaz na aplikaci v literatuře), spin-off firmou Západočeské univerzity v Plzni, jej následně může využít ke komunikaci s okolím. V aplikaci pouze napíše požadovaný text a téměř okamžitě zařízení tento text převede do zvukové podoby a „přečte“ jej. Pokud pacient nemá zájem nebo čas nahrávat, nebo se mu jeho/její (třeba již značně poškozený) hlas nelíbí, je mu/jí samozřejmě nabídnuta generická syntéza profesionálního řečníka (v aplikaci je k dispozici ke stažení několik mužských i ženských hlasů).



Obr. 2: Ukázka používání aplikace SpeeTech TTS na mobilním telefonu

Protože se k tvorbě jednotlivých hlasů používají skutečné nahrávky pacientů, zní personalizovaná řečová syntéza velmi podobně jako původní pacientův hlas v době nahrávání. Nabízí se samozřejmě otázka, zda by nebylo možné syntetický hlas nějak „vylepšit“, a to z toho důvodu, že při nahrávání mají pacienti často svůj hlas již poškozený (mají chrapot, někdy spíše šeptají). Pokusy s vylepšením personalizované syntézy děláme, pro tyto účely se většinou používá metoda statistického modelování, kde je možné měnit různé parametry modelů. Lze také využít modely profesionálního řečníka a ty na pacientův hlas na základě nahrávek adaptovat (Hanzlíček, 2011). Problémem je, že úprava hlasu není jednoduchá a navíc takto vylepšený syntetický hlas se může barvou a dalšími charakteristikami výrazně lišit od původního pacientova hlasu.

Aktuální nevýhodou používání řečové syntézy pro komunikaci s okolím je, že je zaměstnána vždy alespoň jedna ruka, podobně jako u hlasové protézy. Výhody jsou ale ty, že elektronický hlas se neunaví, může mluvit libovolně dlouho, je možné s ním např. i telefonovat. Navíc pokud se jedná o vlastní nahraný hlas, pro pacienta i jeho nejbližší je obrovským benefitem skutečnost, že i přes ztrátu hrtanu s hlasivkami stále komunikuje svým hlasem.

Shrnutí a vize do budoucna

Během loňského roku jsme již vytvořili několik syntetických hlasů pro pacienty Fakultní nemocnice v Motole, ukázky syntézy si můžete poslechnout v tabulce (odkaz najdete v literatuře), kde je můžete také porovnat s původními nahrávkami pacientů. Někteří pacienti personalizovanou řečovou syntézu běžně využívají i pro komunikaci související s jejich profesním životem, jiní alespoň občas jako připomenutí svého ztraceného

hlasu v kruhu svých blízkých. Od samotných pacientů i jejich rodinných příslušníků dostáváme na personalizovanou syntézu pozitivní ohlasy.

Nevýhodou „řečové protézy“ zůstává fakt, že je třeba k jejímu použití využívat nějaké další zařízení, které musí mít tito lidé vždy při sobě. Pro některé z nich je navíc používání mobilních telefonů či tabletů velmi komplikované, protože na ně nebyli dříve zvyklí (jedná se často o starší osoby). Z toho důvodu je třeba nahlížet na personalizovanou řečovou syntézu jako na určitou formu protézy. V případě nové protézy totiž pacienti vždy procházejí rehabilitačním procesem, učí se s ní zacházet. Obdobně by se měli školit i v používání syntézy na mobilním zařízení, což se zatím systematicky neděje.

Naší vizí do budoucna je, že by se řečová syntéza mohla používat v nositelné elektronice, tzv. w*earables*, do které by mohla být integrovaná i senzorická čidla pro její přirozenější ovládání.

Poděkování

Výzkum vedoucí k těmto výsledkům byl financován z prostředků Norského finančního mechanismu na období 2009-14 a Ministerstva školství, mládeže a tělovýchovy v rámci projektové smlouvy č. MSMT-28477/2014, projekt č. 7F14236.

Literatura

Hanzlíček, Z., Romportl, J., Matoušek, J.: Voice Conservation: Towards Creating a Speech-Aid System for Total Laryngectomees. *Beyond Artificial Intelligence: Contemplations, Expectations, Applications, vol. 4,* p. 203-212, Springer, Berlin, Heidelberg, 2013.

Matoušek, J., Tihelka, D., Romportl, J.: Current state of Czech text-to-speech system ARTIC. *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, *vol. 4188*, p. 439-446, Springer, Berlin, Heidelberg, 2006.

Tihelka, D., Kala, J., Matoušek, J.: Enhancements of Viterbi Search for Fast Unit Selection Synthesis. *Proceedings of Int. Conf. Interspeech 2010*, p. 174-177, 2010.

Hanzlíček, Z.: Czech HMM-Based Speech Synthesis . *Text, Speech and Dialogue*, *Lecture Notes in Computer Science*, *vol. 6231*, p. 291-298, Springer, Berlin, Heidelberg, 2010.

Matoušek, J., Romportl, J.: On building phonetically and prosodically rich speech corpus for text-to-speech synthesis. *Proceedings of the second IASTED international conference on Computational intelligence*, p. 442-447, ACTA Press, San Francisco, 2006.

Jůzová, M., Romportl, J., Tihelka, D., 2015. Speech Corpus Preparation for Voice Banking of Laryngectomised Patients. T*ext, Speech and Dialogue*, *Lecture Notes in Computer Science, vol. 9302,* Springer, Berlin, Heidelberg, 2015.

Hanzlíček, Z.: Czech HMM-based Speech Synthesis: Experiments with Model Adaptation. *Text, Speech and Dialogue, Proceedings of the 14th International Conference TSD 2011*, *Lecture Notes in Artificial Intelligence*, *vol. 6836*, p. 107-114, Springer, 2011.

Tabulka s ukázkami: https://docs.google.com/presentation/d/1iWeeWFW-jYIO1fMV9CxImC261QoPUUiZKOsT9w8UxXc/present#slide=id.gc411cf050_0_0

SpeechTech TTS, https://play.google.com/store/apps/details?id=com.speechtech.engine