ÚSTAV MERANIA
SLOVENSKÁ AKADÉMIA VIED
Dúbravská cesta 9, 841 04 Bratislava
Tel.: 02/ 5477 4033, Fax: 02/ 5477 5943
Email: umersekr@savba.sk, Web: http://www.um.sav.sk

Oponentský posudek disertační práce Ing. Martin Grůbera

Syntéza expresivní reci s využitím dialogových aktu k popisu expresivity

Posuzovaná disertační práce představuje rozsáhlou interdisciplinární studii čítající včetně příloh více jak 190 stran, z nichž zhruba jednu třetinu tvoří teoretická část. Zde autor provádí rozbor základních metod syntézy řeči, dále se podrobněji věnuje psychologicko-fyziologickému analýze emočních projevů v řečovém dialogu včetně funkčního popisu systémů, které tyto přístupy využívají. V závěru teoretické části se pak autor zabývá metodami pro vyjádření expresivity v syntetickém řečovém signálu. Tento výklad zaměřen především na metody implementaci v TTS systémech se zdůrazněním základních požadavků na strukturu a parametry řečového korpusu pro expresivní syntézu. V praktické části je nejprve proveden podrobný rozbor návrhu dialogového systému včetně použitých dialogových aktů, aplikované metodiky pro vytvoření řečového korpusu a následného statistického zpracování získaných reálných dat. Jádro práce tvoří vlastní experimentální část zahrnující detailní popis realizace nahrávek pro řečový korpus, následná anotace vět pomocí poslechových testů a statistických metod, dále provedené akustické analýzy hodnot F0, doby trvání, energie (representované hodnotou RMS) a pozicemi prvních tří formantů. Dále je v této části práce řešena problematika modifikace metody výběru řečových jednotek pro TTS systém včetně popisu výpočtu penalizační matice, ceny cíle a váhovacích koeficientů pro příznak expresivity. V 9.kapitole následuje vyhodnocení vlastností navrženého dialogového systému provedené formou porovnání lidského vnímání expresivity v přirozené a syntetické řeči produkované TTS systémem. Přílohy A-E obsahují základní popis a definice použitých metod pro statistické zpracování dat řečového korpusu. Práce obsahuje i doprovodné CD s ukázkami nahrávek pro dialog, syntetické věty s expresivitou generovanou TTS systémem a výslednou podobu dialogu člověka s avatarem.

Hodnocení práce:

Z formálního pohledu je provedení práce na vysoké úrovni – po grafické a lexikální stránce prakticky bez chyb. Citační odkazy jsou používány správně, výběr použitých zdrojů literatury je vhodný pro tento typ práce. V teoretické části práce autor prokázal že se dobře orientuje ve stavu řešené problematiky na světové i domácí úrovni. Z experimentální části práce vyplývá že disponuje širokými znalostmi z oblasti psychoakustiky a fyziologie vnímání řeči, dobře se rovněž dokáže orientovat v oblasti zpracování signálů. Navíc ovládá potřebný matematický aparát k pochopení a realizaci jednotlivých analytických a statistických úloh pro řešení stanovených cílů práce. Autor v práci použil celkem 134 pramenů literatury pokrývající velmi dobře celou řešenou problematiku. Disertant se dále může pochlubit celkem 17

vlastními pracemi v angličtině – příspěvky na renomovaných zahraničních i domácích konferencích a workshopech – což je plně dostačující, přesto bych preferoval zastoupení rovněž časopiseckých publikací. Z popisu nastíněného průběhu řešení je jasné, že se jedná o velmi rozsáhlou, komplexní problematiku, kdy bylo zapotřebí provést velké množství prací, analýz, výpočtů včetně vytvoření přípravných a pomocných programových nástrojů. To vše jako celek přesahuje možnosti jednoho člověka a samozřejmě i rámec této disertační práce.

Mám však několik připomínek týkajících se struktury a členění práce:

- Vzhledem k značnému množství použitých symbolů, matematických vztahů a definic je podle mého názoru autorův seznam „symbolů a zápisů" obsahující pouze 9 položek nepostačující. V této souvislosti bych považoval za užitečné do práce zařadit alespoň přehled typografické konvence včetně symboliky popisovaných matematických operací, který by pomohl lepší srozumitelnosti a orientaci v textu.
- V práci se několikrát vyskytují poznámky typu „tento jev/postup by stál za prozkoumání" (za všechny např. str. 61), v návrhu dalšího vývoje (str. 124-5) se však nevyskytují – působí tak spíše dojmem řečnické otázky.
- V práci je několikrát zmiňováno použití „testu významnosti ANOVA" (např. str. 75) a na rozdíl od jiných případů (definice obecně známých základních statistických veličin typu střední hodnoty, rozptyl atd.) zcela bez vysvětlení funkce a základního principu.
- Celá práce je napsána takovým způsobem, že není možné přesně zjistit co bylo vlastním přínosem disertanta a co vytvořil někdo jiný (v rámci daného pracovního kolektivu). Jedná se především o časté používání formulací typu: „my předpokládáme", „v naší práci" (str. 55, 57) apod. Rovněž nelze poznat zda autor používá již hotové SW nástroje, případně kdy a kdo byl jejich tvůrcem (např. aplikace pro nahrávaní metodou WoZ – str. 49 a pro nahrávání expresivního korpusu – str. 58-59.

Ukázky syntézy na přiloženém CD ve formátu *.mp3 souborů se vyznačují vysokou kvalitou syntetické řeči, expresivita je ve všech ukázkách velmi dobře rozeznatelná. Obecně lze říci, že lepší výsledky byly dosaženy s použitím metody Unit Selection. U expresivní syntézy metodou HMM jsou v několika případech slyšitelné lokální změny energie resp. zákl. tónu – např. v nahr. *show-interest_01.mp3*, rovněž ne vždy zcela odpovídá deklarovaný styl expresivity danému emotivnímu vjemu při poslechu (např. u souboru *happy-empathy_02.mp3* kdy syntetická řeč působí spíše smutně).

Dotazy do diskuse v průběhu obhajoby:

- V průběhu nahrávání reálných dat pro dialog se spolu s řečí zaznamenával i glotální signál. Dále je v práci stručně zmíněno jeho využití pro určování pozic pitch-pulsů a následně hodnot základního tónu F0. Tomuto účelu lze s úspěchem využít i jiné nástroje (např. zmiňovaný program PRAAT [9]) pracující pouze s řečovým signálem. Chtěl bych proto vědět co vedlo autora k tomu, že zvolit právě tuto metodu, zda přineslo její použití očekávané výsledky (kromě zvýšení složitosti fáze vlastního nahrávání).
- Při vyhodnocovaní-posuzování vlivu hodnot F0 ve větách se do výpočtu zahrnovaly pouze znělé úseky nebo jen oblasti fonémů. V praxi se často při analýze řeči vytváří obálka –

virtuální průběh F0 pro celý signál pokrývající i oblasti neznělých úseků (např. pomocí interpolace kubickými splajny). Tímto způsobem lze získat více hodnot které při statistickém zpracování mohou přinést reálnější pohled na analýzu vlivu F0. Rád bych znal názor disertanta na tuto metodu, případně důvod proč ji nebylo v práci možné/vhodné použít.

- Co znamenají (vyjadřují) záporné hodnoty koeficientů šikmosti a špičatosti uvedené v tabulkách 7.6-7.11 ?
- U uvedených vzájemných pozic formantů F_1 x F_2 pro základní samohlásky (str. 160-4) chybí porovnání (posouzení shody případně diskuse rozdílů) s dosaženými výsledky s jiných autorů v češtině, nebo platí také obecně i pro jiné jazyky (např. němčina) ? Využily se nějak také hodnoty třetího formantu F_3 ?

Závěr:

Přes dílčí připomínky předložená disertační práce zcela vyhovuje podmínkám kladeným na tvůrčí vědeckou práci. Stanovené cíle práce byly splněny, do budoucna je třeba ještě dořešit některé dílčí úlohy k dosažení vyšší přirozenosti v expresivní řeči generované TTS systémem. Proto práci **doporučuji k obhajobě** a zároveň **navrhuji** udělit Ing. Martinu Grůberovi akademicko-vědecký titul **doktor** v oboru KYBERNETIKA.

V Bratislavě, 7.1. 2013.

.....

Dr. Ing. Jiří PŘIBIL

Západočeská univerzita v Plzni, fakulta aplikovaných věd, katedra kybernetiky

Posudek na disertační práci k získání titulu doktor v oboru Kybernetika
Ing. Martina Grůbera

Syntéza expresívní řeči s využitím dialogových aktů k popisu expresivity

Ing. Martin Grůber se ve své disertační práci zabývá jednou z úloh zpracování řeči. Práce je věnována příspěvku ke zlepšení přirozenosti syntetické řeči, tj. syntéze expresívní řeči.

Předkládaná práce je členěna do 10 kapitol a je doplněna rozsáhlým seznamem prostudované literatury (obsahuje 134 položek), seznamem vlastních publikací nebo publikací, jejichž je spoluautorem, (17 v angličtině a 2 kvalifikační práce v češtině), dále abstraktem, obsahem, seznamem obrázků a seznamem použitých symbolů a zápisů. Velmi užitečné jsou také přílohy, kterých je sedm. První čtyři kapitoly jsou koncipovány jako stručný úvod a rozbor metod užívaných k syntéze řeči obecně, k popisu a syntéze expresívní řeči včetně vytvoření korpusu. Pátá kapitola je věnována cílům disertační práce. Zbylé kapitoly jsou stěžejní. Ve třech kapitolách je popsán postup vývoje systému pro syntézu expresívní řeči v dialogu a jedna kapitola shrnuje dosažené výsledky. V závěru je provedena rekapitulace předchozích kapitol, dosažené výsledky jsou zhodnoceny a je navržen jejich budoucí vývoj. Přílohy se týkají podrobnějšího popisu použitých metod, výsledků a ukázek.

Téma zvolené pro disertační práci je velmi aktuální. Syntéza řeči, v dnešní době pak především syntéza expresívní řeči, je potřebná v mnoha oblastech lidské činnosti. K tomu, aby syntetická řeč byla využívána v ještě větší míře, než je tomu dosud, je třeba, aby byla nejen srozumitelná (to je dnes již nutný požadavek a většina systémů jej splňuje), ale také přirozená, tedy co nejméně se lišící od přirozené řeči. Z těchto důvodů se touto oblastí zpracování signálu zabývají velmi intenzivně mnohá výzkumná pracoviště na celém světě. O tom svědčí množství příspěvků na nejrůznějších konferencích, seminářích a workshopech. Podílí se na nich i autor předkládané disertační práce. Jeho práce se zaměřila na expresívní syntézu pro potřeby dialogového systému. Za vysoce pozitivní považuji fakt, že se jedná o dialogový systém zaměřený na komunikaci seniorů s počítačem. Tématem dialogu byl popis fotografií z jejich života.

Zpracované téma má velký rozsah. Autor se zabývá vytvořením korpusu expresívní řeči pro danou tématiku (to samo o sobě je velmi náročný úkol), zkoumáním anotací textů obsažených v korpusu a velmi podrobnou analýzou dat v závislosti na různých akustických parametrech. Klade si vysoké cíle, které se mu však díky píli a velkému pracovnímu nasazení podařilo splnit. Navržené řešení je komplexní, použité metody jsou moderní. Jsou založeny na rozsáhlých znalostech v mnoha oborech, jsou podepřeny velmi dobrým matematickým zázemím a zkušenostmi s řešením otázek zpracování řečového signálu, které měl autor možnost získat na předním pracovišti v této oblasti výzkumu v naší republice. Velký význam celé práce je nejen v množství experimentů s ověřováním a porovnáváním výsledků získaných odvozenými metodami, ale také v tom, že práce je součástí mezinárodního projektu Companions, který je sponzorovaný Evropskou komisí a byl součástí projektu GAČR. Popisované výsledky experimentů jednoznačně ukazují na správnost volby tématu, na dobře zvolený postup a náplň prací. Bohužel jsem nemohla posoudit ukázky a poslechové testy, které jsou na přiloženém CD pouze ve formátu MP3, a ten jsem v době psaní posudku neměla možnost si přehrát. Velmi oceňuji, že se jedná právě o konkrétní využití výsledků této práce.

Po formální a technické stránce je předkládaná práce na velmi dobré úrovni. Je psána přehledně, autor prokázal schopnost pracovat tvůrčím způsobem. Také logická stavba práce je na velmi dobré úrovni. Značný počet prací, jichž je autorem či spoluautorem, většinou

publikovaných na prestižních mezinárodních konferencích, dokazuje nejenom autorovu schopnost vědecky pracovat, ale i schopnost informovat o dosažených výsledcích.

Přesto mám několik připomínek a dotazů. Mezi připomínky patří např. konstatování, že:

- Tabulka 7.8. na str. 79, citovaná na str. 81 se netýká hodnot koeficientů šikmosti a špičatost, ale trvání fonémů.
- Str. 85, poslední věta paragrafu 7.3.6. – domnívám se, že barva hlasu je obsažena až ve vyšších formantech, ne v prvních třech.
- Nelíbí se mi slovo „vysyntetizovat" na str. 119.

Dotazy k práci mám následující:

- Píšete, že dnes se používá především automatická segmentace. S tím souhlasím. Vysvětlete mi ale, jaký vliv na syntetickou řeč mají systematické chyby automatické segmentace a jak se odstraňují.
- Jakým způsobem zabudujete dialogové akty do počítačového systému?
- Na str. 64 v odstavci, který se týká vývoje korpusu pro syntézu expresívní řeči, máte tabulku 7.1 se slovním označením pro hodnoty kappa. V textu vysvětlujícím tuto tabulku píšete, že není žádný obecný mechanismus pro vazbu mezi slovním hodnocením a hodnotami kappa, ale je přijímána ta, kterou uvádí tabulka. Uveďte prosím vztah nebo podrobněji popište způsob výpočtu Fleissovy a Cohenovy kappy. Nemůže třeba hodnocení souviset s rozdílnými schopnostmi anotátorů rozeznat expresivitu? Zjišťujete u anotátorů např. hudební vzdělání, hru na nějaký hudební nástroj?
- Paragraf 7.3.4 Analýza doby trvání - v tomto paragrafu porovnáváte Vaše výsledky s výsledky ze dvou zahraničních publikací. V této souvislosti se chci zeptat, o jaké věkové kategorie mluvčích se jednalo v případě zahraničních publikací a zmiňovaných Vašich předchozích prací (publikace [44], [41]). Nemůže to být tím, že senioři mají celkově pomalejší vyjadřování?
- Str. 88 – Vysvětlete, co jste myslel větou „Právě rozdílů mezi vektory reprezentujícími různé dialogové akty bude využito při výpočtu akustické části penalizační matice..." a poznámkou pod čarou „Vektory reprezentující dialogové akty však budou tvořeny poněkud odlišně a budou více než třírozměrné".
- V tabulce 7.12, str. 89, počítáte korelace mezi velkou skupinou fonémů (jedná se o všechny fonémy a o znělé fonémy v dialogových aktech) a fonémem „e". Na základě poměrně vysoké korelace usuzujete na reprezentativnost výsledků akustických parametrů napříč dialogovými akty. Opravdu budou vysoké korelace i s jinými znělými fonémy? Mezi všemi, resp. znělými fonémy představuje „e" přibližně jen 15%, resp. 17%.
- Při poslechových testech pro přirozenou expresívní řeč a pro syntetickou expresívní řeč byl rozdílný počet posluchačů (o 1 posluchače). Jednalo se o 13 stejných posluchačů v obou případech?

Na závěr konstatuji, že Ing. Martin Grůber projevil schopnost samostatně vědecky pracovat. Vytčené cíle byly splněny. Proto mohu konstatovat, že **disertační práci doporučuji k obhajobě**.

V Třeboni, 30.1.2013

Prof. Ing. Jana Tučková, CSc.
Katedra teorie obvodů, FEL ČVUT v Praze