Bírálat Németh Géza: “Célorientált beszédkeltés interakciós rendszerekben” c. MTA Doktora disszertációjáról

1. Általános megállapítások a dolgozat témaválasztásával és struktúrájával kapcsolatosan

A disszertáció a beszédgenerálás foglalkozik. A szerzőnek a PhD fokozata megszerzése óta a tématerületen végzett kutatásait és ezek eredményét foglalja össze. Új módszereket és algoritmusokat ad a beszédkeltésre, amelyek jóságát „meghallgatásos” tesztekkel minősíti. A fókuszban hangsúlyosan a magyar nyelvre vonatkozó fejlesztések állnak, elsősorban látáskorlátozottak számára elérhető alkalmazások kidolgozásával. Az életminőség javítása szempontjából ez nagy értéket ad a fejlesztéseknek. A dolgozat tézisei IV téziscsoportban vannak összefoglalva, amelyek a

- célorientált gépi szövegfelolvasó rendszerek kutatásához, kialakításához és továbbfejlesztéséhez (I., II. és III. Téziscsoport), illetve
- a multimodális információs rendszerek hatékony megoldásainak kutatásához (IV: Téziscsoport)

kapcsolódnak.

A problémakör fontosságát rengeteg alkalmazási terület támasztja alá, amely az sms felolvasástól a nyilvános információs rendszerekig terjed.

A fejlesztési eredmények impresszívek, azonban ezeket sok esetben nem kíséri semmiféle analitikus tárgyalás. Ez egyrészt a témakör sajátossága is, azonban ezen tárgyalás hiánya olyan esetekben is előfordul, ahol lehetséges lett volna egzakt megközelítéseket ismertetni és azokon demonstrálni a módszerek működését.

Az értekezés szinte kizárólag szöveg alapú, némely esetben elnagyolt és nem önmagyarázó, az eredmények feldolgozáshoz a háttérirodalom tanulmányozására is szükség volt. Olyan esetekben hivatkozik új „elmélet” megalkotására, ahol inkább egy empirikus ötlet megvalósítása kerül részletezésre.

Az eredményeket általában szubjektív tesztek minősítik, ami ezen a területen természetes. Azonban bizonyos esetekben nincsenek részletezve a kísérlet jellemzői (mekkora populáció minősített, mi volt az értékelések statisztikája …etc.), ami azokat az állításokat, hogy egy bizonyos módszer „jobban működik mint egy másik” nehezen teszi ellenőrizhetővé.

A szerző is jelzi, hogy a beszédtechnológia korlátait a tárhely az operatív memória és a számítási kapacitás jelenti, azonban a dolgozat nem tartalmaz ezekhez kapcsolódó részletes komplexitás analízist. Néhány esetben a programsorok száma és futási idők annak feltüntetve, de műveletszámról és a hardverről nincs sinformáció. Még olyan estben sem, amikor egy módszernél (pl. HMM alapú beszédszintézis) ez előnyös az „elemválogatáshoz” képest.

Valóban nagyon fontos, hogy a magyar nyelvű beszédkeltésnek hatékony módszerei legyenek. Ugyanakkor az értekezésben főleg a már bevezetett módszerek magyar nyelvre történő adaptálása kapott részletezést. Ez nagy munkát és innovatív megközelítéseket igényelt, amelyek fontos - az életminőséget is javító - alkalmazásokkal bírnak. Azonban a munka jellege inkább fejlesztési természetű. Nem csoda, hogy a dolgozat majdnem fele az alkalmazásokat és ezek technikai paramétereinek a leírását tartalmazza.

A disszertáció inkább tekinthető egy értékes szakmai életmű összefoglalásának, ahol a szerző sok évtizedes eredményes munkája van összefoglalva. Ezek során olyan implementációkra is utalás kerül, ahol a mobil operációs rendszerbeli háttér már idejét múlt (pl. Symbian), ezért az eredmények egy része nem biztos, hogy a kurrens beszédtechnológiák részét képezik.

2. A dolgozat tartalma és struktúrája

A dolgozat tíz fejezetet tartalmaz, amelyek tartalmi elemei a következők:

- A bevezető a beszédgenerálás modelljeit ismerteti röviden, irodalmi hivatkozásokkal és a probléma rövid történeti áttekintésével.
- A 2. fejezet „A gépi beszédkeltés különböző megközelítései, történelmi áttekintés” címmel ad egy részletesebb összefoglalót a problémakör fejlődéséről
- A „Kutatási célkitűzések” című harmadik fejezet, több a munka minőségét értékelő állításokat tartalmaz, pl. a Jaws for Windows esetén egy amerikai fejlesztő elismerő e-mail-jére utal, nem ártott volna ezt az e-mailt a függelékben csatolni az illető affiliációjával. Hasonlóan, mikor a szerző arról nyilatkozik, hogy a II: téziscsoport eredményeihez kapcsolódó előadásuk „felkeltette a hasonló témán francia nyelven dolgozó kutatók figyelmét és érdeklődtek a részletek iránt” szintén dokumentálni. Ugyanakkor fontos elismerését jelenti a munkának a két támogatott H2020-as projekt (PAELIFE és VUK AAL). A rövidítések mellett itt a címük és a konzorcium megadása is informatív lett volna.
- Az „Eszközök és módszerek” fejezet a kutatáshoz használt adatbázisokat, valamint a kutatás módszertanát részletezi.
- A szerző eredményeinek ismertetése az 5. Fejezettől („A diád és triád elemek összefűzésén alapuló gépi szövegfelolvasás”) kezdődik, amely az I. Téziscsoporthoz kapcsolódik. Ebben a szerző kidolgozta az első magyar nyelvű hullámforma elemösszefűzéses gépi szövegfelolvasó rendszerét és megtervezte a diád és triád hullámforma elemek megvalósításához szükséges akusztikus adatbázis szerkezetét
- A „Célorientált, korpusz alapú gépi felolvasó rendszerek” (6. Fejezet) a szerző a korpusz alapú elemkiválasztásos módszert ismerteti, három célterületen: (i) időjárás-jelentés, (ii) árlista , (iii) pályaudvari utastájékoztatók. Ezek során a szerző kidolgozta a korpusz alapú hangnyomás-idő függvények automatikus válogatásán alapuló beszédkeltő modelljét. Egyúttal többszintű modellt fejlesztett ki az elemválogatáson alapuló magyar

nyelvű szövegfelolvasó különböző területein való alkalmazásaihoz. Új módszert adott a prozódiai frázisok hasonlósága alapján képzett prozódiai csoportok létrehozásához.

- A „Statisztikus parametrikus gépi szövegfelolvasó rendszerek” c. fejezetben a szerző a rejtett Markov modellt magyar nyelvű gépi felolvasásra adaptálta és meghatározta a tanításhoz szükséges beszédadatbázis szerkezetét. Új módszer került kidolgozásra a rövid és kérdő mondatok jobb minőségű szintéziséhez. A maradékjelre alkalmazott elemkiválasztásos eljáráson alapuló beszédkódolót hozott létre.
- A „Multimodális beszédinformációs rendszerek” c. fejezetben a grafikus és beszédfelhasználói területeket integrálja a szerző. Új megoldást dolgozott ki a modalitások szinkronizált kezelésére. Új akusztikus üzenetforma – a spemoticon – került kidolgozásra.
- „Az eredmények alkalmazása, műszaki alkotások” c. fejezetben a szerző által kifejlesztett alkalmazások kerülnek hosszas ismertetésre: (i) e-mail felolvasó rendszer távközlési szolgáltatásként, (ii) sms felolvasó rendszer, (iii) árlista bemondó szolgáltatás, (iv) MÁV állomások utastájékoztató rendszere, (v) magyarul beszélő robot alkalmazása kórházi környezetben, (vi) gyógyszer betegtájékoztató információs rendszer.
- „A tézisek összefoglalása egységes szerkezetben” c. fejezet a téziseket ismerteti újra.

3. Részletes megjegyzések és kérdések a dolgozat állításaival kapcsolatosan

5. Fejezet 16. old.

„Kidolgoztam a magyar nyelv sajátosságainak megfelelő első diád és triád hullámforma elemösszefűzéses gépi szövegfelolvasó eljárás rendszertervét (ld. 6. ábra), amely diád és triád méretű magyar hangkapcsolódások felhasználásával készít gépi beszédet, és igazoltam, hogy az ezek felhasználásával létrehozott rendszer MOS (Mean Opinion Score) szubjektív értékelés szerint jobb hangminőséget ad, mint a korábbi, más elven működő megoldások (például Hungarovox [28], Brailab [44], PC talker [45]) Az eljárást kiterjesztettem német nyelvre is”

A rendszerterv és a korpuszkészítés nagyon fontos feladat, de nem triviális, hogy a blokkvázlaton látható rendszerterv miért tudományos eredmény. Mennyire kellett ennek megalkotásához tudományos eszközöket használni és mennyiben nagy az újdonságtartalma.

A szubjektív értékelésnek ennél az eredménynél nincsenek megadva a paraméterei (hány ember hallgatta, milyen kor/nem eloszlásban, mik a statisztikai eredmények)

„20. oldal A rendszert a MAILMONDÓ szolgáltatás [48] és [49] fejlesztése és alkalmazása során széles körben teszteltük és

megállapítottuk, hogy jobb minőséget nyújt, mint a korábbi magyar nyelvű gépi szövegfelolvasó megoldások (ld. 11. ábra, 32.o). A német nyelvű változatot kutatási együttműködés keretében a TU Kaiserslautern és a Fraunhofer IESE anyanyelvű munkatársaival validáltuk [50].

Ennek az állításnak validációja nheéz, a jelzett cikk, illetve MATÁV 2000-ben kiadott jelentése (21 éves) nem letölthető.

6. Fejezet

24. oldal A szerző megközelítésében a hosszabb elemek kiválasztása növeli a minőséget. Ideális esetben lehetséges/szöveg, vagy mondat szereplejen az adatbázisban. Nem kerül konkrét kifejtései az adatbázis komplexitása és a hosszú elemekből építkező beszédszintézis minősége közötti összefüggés ? Erre vonatkozóan semmilyen becslés nincs a dolgozatban. „Természetesen ez a gyakorlatban kivitelezhetetlen, ezért olyan egységeket rögzítenek az adatbázisba, hogy a szintetizálandó mondat nagy valószínűséggel hosszú elemekből legyen összefűzhető." Milyen statisztikai módszer alapján lettek ezek kiválasztva, hogy a fenti kritériumot teljesítsék ?

A vonatkozó tézisben is említett 26. oldal 9. ábrája magától értetődőnek tűnik, szükséges lenne annak alátámasztása, hogy ezen ábra megalkotásához milyen tudományos kihívások megoldására volt szükség.

A prozódiai cimkézés utáni, a címke illeszkedést kereső „válogató függvényt" nem részletezi a szerző. Nincs vizsgálat arra vonatkozólag, hogy például hogyan lehet a legyorsabb keresést alkalmazni.

A 27. és 28. oldalon két költségfüggvényt C(n) és P(n) definiál a szerző, majd ezek összegét optimalizálja. A C(n) függvény súlyértékeit „500 mondat többszöri szintézisével határoztuk meg". Konkrétan hogyan ?

Másrészt miért állítja elő a teljes célfüggvényt additív formában C(n)-ből és P(n)-ből, hiszen az összegfüggvény optimuma távol eshet mind C(n) és P(n) optimumától és nem biztos, hogy jó kompromisszumos megoldást szolgáltat. Miért nem kényszeres optimalizálásként tekint a problémára, ahol az egyik függvény adott minőségű kényszere mellett keresi a másik optimumát.

Az sem kerül részletezésre, hogy milyen statisztika alapján generált korpuszban találhatunk nagy valószínűséggel szófüzéreket (nemcsak szavakat) ?

A Viterbi algoritmus használatát illetően a szerző az eredeti 1967-es cikkre hivatkozik, holott azóta vannak továbbfejlesztett verziók, amelyek az algoritmus gyorsabb végrehajtását teszik lehetővé. Ezért felvetődik a kérdés nem lenne érdemes egy továbbfejlesztett algoritmust használni, illetve mennyire fontos a sebesség?

7. Fejezet

A HMM beszédkeltésre való használatánál nem kerül kifejtésre a tanulási algoritmusok kérdésköre (pl. Baum-Welch algoritmus), hogy ezek milyen gyorsan érik el az optimumot.

4. Összefoglaló és javaslat a nyilvános vitára bocsátásról

A disszertáció alapján elmondható, hogy a jelölt fontos kutatásokat végzett és a mesterséges beszédkeltés területén, amelyek egy része főleg fejlesztés jellegű. Az eredmények hasznosságát számos életminőséget is javító alkalmazás mutatja, amelyek a disszertáció utolsó fejezetében kerültek összefoglalásra. Ezek az eredmények a tézisek műszaki hasznosulásáról is jó képet adnak. A tézisek publikáltsága is tükrözi a kutatások főleg magyar nyelvű orientációját.

A dolgozat lehetett volna pontosabb és részletesebb, a módszertant és az algoritmusokat illetően, mert a jelen formájában ez néha megnehezítette a eredmények tudományos jellegének áttekintését és értékelését. Ugyanakkor az eredmények minősége megfelel az MTA doktora címmel szemben támasztott követelményeknek és a disszertációnak a nyilvános vitára bocsátását mindenképpen javaslom.

Budapest 2021. május 10.

Dr. Levendovszky János