Válasz Dr. Gósy Mária professor asszony bírálatára

Köszönöm Dr. Gósy Mária professzor asszony alapos bírálatát, javaslatait és támogató összegzését. Válaszaimat a bírálat szerkezetéhez illeszkedve fogalmazom meg. A bíráló megállapításait *félkövér betűtípussal* idézem meg. A saját válaszaimat pedig normál betűtípussal adom meg.

1. Általános megállapítások

1.1 „Németh Géza a gépi beszédelőállítással annak indulásától folyamatosan foglalkozik. Az értekezés adekvátan mutatja be azokat az elemzéseket, fejlesztéseket, a megvalósított elképzeléseket és gyakorlati vonatkozásokat, amelyeknek a részese volt. Első ízben történik meg, hogy ennek a témának az összegzésére vállalkozzon valaki, ráadásul nem külső szemlélőként, hanem aktív részeseként a több évtizedes kutatómunkának. Személyes meggyőződésem, hogy a múlt, azaz a „ma" sikereihez vezető megtett út eredményeinek és kudarcainak ismerete nélkül nem értékelhető megfelelően a jelen, és csak bizonytalanul tervezhető a jövő. Éppen ezért nagyra értékelem a jelen disszertáció szakmailag hiteles, pontos áttekintését, mivel a mesterséges beszédelőállítás jelene az elmúlt időszak kritikáján, visszajelzésein és minősítésén kell, hogy alapuljon. Ezt a szemléletet a jelölt maradéktalanul érvényesíti, és a feladatot kitűnően teljesíti."

Köszönöm professzor asszony elismerő sorait. Valóban arra törekedtem, hogy bemutassam, hogy az alkalmazási céltól és a rendelkezésre álló műszaki feltételektől függően milyen kutatási célkitűzések merültek fel, azokat hogyan lehetett elérni és milyen gyakorlati alkalmazásokba lehetett bevezetni. Ennek megfelelően a már megállapodottnak tekinthető kutatási eredményekre összpontosítottam.

1.2 *„ A szakirodalomban szereplő tételek relevánsak, jók, bár néhány magyar szerző fonetikai publikációját szívesen olvastam volna. Ha már szó esett a formáns kialakulásáról, akkor Gunnar Fant munkája megkerülhetetlen lett volna."*

Egyetértek bírálóm hiányérzetével. Kutatásaink jelentős részben a fonetikai vizsgálatok eredményein alapultak (pl. a korpusz alapú rendszerek költségfüggvényei súlyainak meghatározása). Ezt jelzi az is, hogy az MTA Nyelvtudományi Intézet Kempelen Farkas Beszédkutató Laboratóriuma (ma ELKH Nyelvtudományi Kutatóközpont Fonetikai Osztály) által szervezett Beszédkutatás konferencia és kiadvány sorozatban rendszeresen megjelentek kutatócsoportunk eredményei. Viszont a terjedelmi korlátok miatt a hivatkozásokban igyekeztem a szűkebb témakörhöz szorosan kapcsolódó irodalmakra hagyatkozni. Ezért szerepel csak Gunnar Fant egyetlen többszerzős publikációja[1].

1.3 *„. Nem véletlen, ahogyan a disszertáció bibliográfiájából is látszik, hogy a jelölt szerzőtársakkal dolgozott egy-egy (rész)téma megoldásában. … Az azonban problémát jelent, hogy nem derül ki a jelölt saját kutatásának pontos beazonosíthatósága. Jó lett volna, ha a hivatkozott két vagy többszerzős munkálatok tárgyalásakor egyértelműen kijelöli a saját feladatkörét, meghatározza a saját eredményeit."*

Valóban, a kezdetektől fogva csapatmunkában végeztem kutatásaimat. Köszönöm bírálómnak, hogy felhívta a figyelmemet arra, hogy nem eléggé egyértelmű saját eredményeim elhatárolása. Ennek a célnak az eléréséhez arra törekedtem, hogy a (rész)témavezetésemmel készült PhD disszertációk

[1] [2] K. N. Stevens, S. Kasowski és C. G. M. Fant, „An electrical analog of the vocal tract," Journal of the Acoustical Society of America vol. 24. issue 2, p. 734–742, 1953.

téziseitől jól elhatárolható téziseket fogalmazzak meg, melyek így saját eredménynek tekinthetők. A közreműködő kutatótársakat az egyes fejezetek vége felé valamint a Köszönetnyilvánítás szakaszban említettem meg.

1.4 *„A gépi beszédkeltés értelemszerűen a magyar nyelv hangzó változatáról szól, mégis jó lett volna, ha képet kap az olvasó a mindenkori nemzetközi eredményekről, technológiákról, fejlesztésekről. Egy ilyen (akár csak néhány mondatos) összehasonlítás még jobban kiemelte volna a magyar beszédelőállítás teljesítményszintjét. A dolgozat tartalmaz ugyan 3,5 oldalt (2. fejezet, 5–8), ami egyfajta történeti visszatekintésnek is felfogható, itt azonban inkább a technológia bemutatása történik (l. táblázat), az eredményeké nem."*

Köszönöm bírálóm felvetését. Szerencsés lett volna áttekinteni a nemzetközi eredményeket is. Bele is kezdtem ennek megfogalmazásába, de két ok miatt maradt ezt ki a dolgozatból. Egyrészt nehéz nyelveken átívelő minőségi összehasonlításokat tenni, mert a nyelvek eltérő struktúrája, prozódiai jellemzői, a szubjektív értékelés, az éppen összehasonlított minták és a kérdések megfogalmazása mind befolyásolják az eredményeket. Általánosan annyit meg lehet jegyezni, hogy ha az 1 - 5-ös MOS minősítő skálán egy rendszer meghaladja a 3.5-ös értéket, akkor jól érhető, ha pedig az eredmény jobb, mint 4, akkor az emberi minőséghez közelítőnek tekinthető. Az eredeti emberi bemondások jellemzően 4.6 fölötti értéket szoktak kapni.

Másrészt egy érdemi nemzetközi kitekintés további terjedelmet vont volna el az amúgy is szűkös keretekből. Kompromisszumként arra törekedtem, hogy az egyes technológiákhoz kapcsolódóan jellemző irodalmi hivatkozásokat adjak meg. Az I. téziscsoport szerinti megoldásunk egy német nyelvű robotikai projektet is segíteni tudott [50].

2. Részletes megjegyzések és kérdések

2.1 *„Az 5. oldalon látható 3. ábra számomra nem értelmezhető ebben a „történelmi" (inkább történeti) áttekintésben. Az ábra jól ismert (bár hivatkozást nem találtam), a formánsok kialakulásának magyarázó szemléltetése. Az ábraaláírás szerint viszont „A gépi beszédkeltés formáns modelljének alapelve". A formánsok valóban a kezdeti beszédszintetizáló rendszerek alapjául szolgáltak, de csak ennyi a kapcsolat az ábra információja és az aláírás között. Érdemes lenne majd korrigálni, avagy magyarázni."*

Elnézést kérek azért, hogy lemaradt a hivatkozás a 3. ábráról[2], amit egy áttekintő cikkemből vettem át. Az ábrát követő néhány bekezdésben próbáltam a formáns-alapú forrás-szűrő beszédszintézis modellt röviden ismertetni. Sajnálom, hogy ez csak korlátozottan sikerült.

2.2 *„A 17. oldalon látható 6. ábra igen hasonló a 25. oldalon látható 8. ábrához. A különbségek bizonyára lényegesek, de mivel a 6. ábra magyarázatát nem találtam meg a szövegben (a 8. ábráét igen), ezért a vonatkozások homályosak. Feltétlenül szükséges annak kissé részletezőbb levezetése, hogy az elemkiválasztásos módszer hogyan működik az elem-összefűzéses megoldáshoz képest."*

Köszönöm bírálómnak, hogy felhívta a figyelmemet erre a hiányosságra. A 6. ábra a hullámforma elemösszefűzéses megoldást mutatja be. Ennek lényege, hogy minden hullámforma elemből csak egyetlen prototípust tárolunk az adatbázisban. A mintegy 1600 magyar diád adatbázis elem összesen 2,5 perc időtartamnak felel meg (ld. 2. táblázat DIAD adatbázisok). A megfelelő elemek összefűzésével tetszőleges magyar szövegnek megfelelő prozódia nélküli (lebegő jellegű) beszéd

[2] [20] G. Németh, „Kempelentől a WaveNet-ig: a gépi beszédkeltés tudományának fejlődése," in A humán tudományok és a gépi intelligencia, G. Tocsvai Nagy, Szerk., Budapest, Gondolat Kiadó, 2018, pp. 127-155.

előállítható. A prozódiát megfelelő jelfeldolgozó algoritmusok segítségével állítjuk elő. Az adatbázis mérete az n ∗ Mbyte nagyságrendbe esik.

A 8. ábra szerinti hullámforma elemkiválasztásos megoldás esetében változó méretű adatbázis elemeket alkalmazunk és jellemzően egy-egy elemtípusból többet is tartalmaz az adatbázis, összesen akár több mint 10 óra időtartamban. A költségfüggvények azt a célt szolgálják, hogy olyan elemeket válasszunk ki, melyeknek összefűzése után a megfelelő prozódiai jellemzők eléréséhez nem szükséges további jelfeldolgozás. Ennek megfelelően az adatbázis jellemzően n ∗ Gbyte méretű.

2.3 „*Nem látom indokoltnak a 4. táblázatot, ami a német nyelvű változatra vonatkozik. Minthogy a disszertáció középpontjában a magyar beszéd áll, jobb lett volna magyar példákat közölni a német nyelvűek helyett.”*

Köszönöm bírálóm megjegyzését. Valóban igaz, hogy kutatásaink és eredményeink magyar nyelven egyediek és hiánypótlóak, azonban megoldásaink sok esetben más nyelvekre is könnyen átvihetőek. Ezt kívántam ebben az esetben bemutatni, amikor hatékony megoldásunk egy német nyelvű robotikai projektet is segíteni tudott.

2.4 „*A beszédészlelési (lehallgatásos) tesztek nagyon fontosak a beszéd minőségének (szubjektív) jellemzésére. A 32. oldalon lévő 11. ábra, a 33. oldalon a 12. ábra, avagy a 13., a 14. és a 18. ábra ilyen kísérletek eredményeit mutatják be az átlagértékekkel. (Az alkalmazott módszertan teljes mértékben elfogadható, ti. az internetes részvétel a kísérletben, mégis fontosnak tartom a személyes aggályaimat kifejezni az ilyen jellegű kísérletekkel kapcsolatban. Nem tudjuk kontrollálni a résztvevőket /adatok önbevallása/, a kísérleti helyzetet és a technikai apparátust sem, pl. a fülhallgató minőségét. Mindez pedig jelentős befolyással lehet a kapott eredményekre. Ez a megjegyzés a jövő kutatásainak szól.) Az eredmények bemutatásakor sokkal informatívabb lett volna, ha a szóródásról is látunk adatokat, illetve szemléltetést (pl. boxplot ábrák). Mindenképpen szükséges lett volna, ha a jelölt közli a statisztikai eredményeket, például egy ANOVA-vizsgálatét. Az sem derül ki, hogy ilyen jellegű elemzést folytattak-e. Ezért különösen zavaró, hogy a „szignifikáns” szó sokszor megjelenik a disszertációban minden alátámasztás, statisztikai felírások, adatok nélkül.”*

Köszönöm bírálóm értékes elvi megjegyzéseit. Felhasználói szempontból a gépi beszédkeltésnek napjainkig nincs jobb minősítési eljárása, mint a szubjektív meghallgatásos teszt. Az internetes megoldás valóban számos módosító tényezőt tartalmaz (PC, vagy mobil hangszóró minősége, fül- és fejhallgatók, stb.), azonban az is igaz, hogy ez közel áll a valós felhasználási körülményekhez. Az időbeli és anyagi korlátok is ritkán teszik lehetővé megfelelő létszámban a stúdió körülmények közötti tesztelést.

A tesztelési eljárásokban az adott időszakban szokásos módszereket alkalmaztuk. A 11. ábra 2006-os publikációnkban jelent meg először ([9] Beszédkutatás), Ebben még nem alkalmaztunk statisztikai elemzést, de az átlagok között is olyan jelentős az eltérés, hogy egyértelmű a szignifikáns különbség. A 12., 13, 14, és 15. ábra 2009 – 2013 között statisztikai elemzés alapján készült, a szórást a jobb olvashatóság és áttekinthetőség miatt nem ábrázoltuk. A frissebb publikációinkban alkalmazunk újabb megjelenítési módszereket (pl. boxplot) is.

2.5 *„Mit ért a jelölt ’prozódiai frázison’? A fonetikai szakirodalomban többféle felfogás és definíció is létezik, ezért a terminus alkalmazása nem egyértelmű. Nem találtam definíciót a „prozódiai egységekre” vonatkozóan sem (pl. 26-27.oldal), javaslom a pótlásukat.”*

A prozódiai frázisra a következő definíciót fogalmaztam meg: elemi gondolati egység, célszerűen két akusztikai szünet közti hullámforma elem (disszertáció 39.o). A prozódiai frázis és a prozódiai egység fogalmát ekvivalensként alkalmazom. Felolvasás esetén a prozódiai frázis jellemzően egybeesik a két tagoló írásjel közötti szöveghez tartozó hullámformával.

2.6 „*A 16. oldalon ez olvasható: „Megterveztem a diád és triád hullámforma elemek megvalósításához felhasználható akusztikus adatbázis szerkezetét….” Jó lett volna látni a szerkezet – feltételezem mátrixos alapú – kialakításának mérnöki tervét.”*

A diád adatbázist még lehetett volna egy kétdimenziós mátrixszal ábrázolni, viszont a triádnál már három dimenziós megjelenítés kellett volna. Önmagában ez a szerkezet nem lenne új tudományos eredmény, hiszen az alapelvből egyenesen következik. A tervezés része volt azonban a .címkézés részletessége (hanghatár és alapfrekvencia), bizonyos esetekben (pl. zárhangok) a hanghatár célszerű megválasztása (pl. zárhangoknál a zár után és a felpattanás előtt). A zöngés hangoknál a periódus határok célszerű megválasztása (a maximum előtti nullátmenet). Az egyenletes alapfrekvenciát biztosító szövegkorpusz kidolgozása is ebbe a körbe tartozik.

2.7 *„17. oldalon: „A hosszú mássalhangzókat csupán időtartam módosítással tudjuk előállítani.” Mi ennek a műszaki megoldása? Minden mássalhangzóra ugyanazt az elvet, illetve megoldást kell/lehet alkalmazni? Ha nem, miért nem?”*

A zöngétlen gerjesztésű, hosszan kitartható hangok hosszú változatát lehet előállítani a hang stabil szakaszán (célszerűen a közepéből) kivágott szakaszok többszörözésével illetve időben fordított sorrendben lejátszott változatával. A zöngés gerjesztésű mássalhangzóknál figyelni kell a stabil szakaszon (célszerűen a hang közepén) levő periódusok pontos kivágására és többszörözésére. A zárhangoknál viszont a felpattanást nem szabad módosítani, hanem a zár szakaszt kell meghosszabbítani.

2.8 *„19. oldalon: „A technológia fejlődésével kiderült, hogy a kis erőforrás igényű diád alapú beszédszintetizáló rendszerek alkalmasak voltak a 2000-es évek elején megjelenő okostelefonokon valós idejű működésre.” Hiányolom a műszaki okfejtést és a konkrét adatokat. Mit jelent, hogy kis erőforrás igényű?”*

A „kis erőforrás igényű” megjelölés az adott időszak műszaki szintjére vonatkozik. Például 2004-ben egy átlagos PC paraméterei a következők voltak[3]: 1GHz órajelű Intel CPU, 128Mbyte operatív memória, 30Gbyte merevlemez tár. Egy ilyen PC-n ügyfélszolgálat automatizálási alkalmazásban, több mint 100 csatornát tudott a szintézis rendszer kiszolgálni. A szintén 2004-ben megjelent, akkor korszerű Nokia 6630-as okostelefon[4] 10Mbyte operatív memóriával, 64Mbyte-os memóriakártyával, 220MHz órajelű egymagos processzorral rendelkezett, 150.000 Ft induló ár mellett. 22kHz mintavételi frekvencia, 16 bites lineáris PCM kódolás mellett egy magyar diádos adatbázis mérete átlagosan 6,3Mbyte (disszertáció 22.o), ami túl sokat foglalt volna el a telefon memóriájából. Viszont a mintavételi frekvenciát a telefóniában szokásos 8kHz-re csökkentve, és 8 bites logaritmikus PCM kódolást alkalmazva az adatbázis mérete mintegy 1,2Mbyte-ra volt csökkenthető, ami már reálissá tette azt, hogy az operatív memóriában tároljuk. A jelfeldolgozási algoritmusokat sebességre optimalizálva a két szempont együttesen tette lehetővé a valós idejű működést. Ez akkor világszinten is újdonság volt. A Nokia meg is hívta a MIT-Systems Kft-vel közösen fejlesztett SMS-felolvasó

[3] http://answers.google.com/answers/threadview?id=414355

[4] https://phonesworlds.com/phone/nokia/6630-733/Nokia-6630-492/nokia-6630/

alkalmazásunkat az egész világból meghívott 30 külső fejlesztő közé a Nizzában rendezett Mobile Minds kiállításra.

Érdemes megemlíteni, hogy egy mai átlagos okostelefon (pl. Motorola One Fusion+[5]) mintegy fele akkora áron 6Gbyte operatív memóriát és 128Gbyte beépített tárhelyet kínál, 8magos processzoral, legalább 1,8GHz-es órajellel.

2.9 „*20. oldalon: „A rendszert a MAILMONDÓ szolgáltatás (G. Németh, et al. 2000) és (Straub 2000) fejlesztése és alkalmazása során széles körben teszteltük és megállapítottuk, hogy jobb minőséget nyújt, mint a korábbi magyar nyelvű gépi szövegfelolvasó megoldások (ld. 9. ábra). A német nyelvű változatot kutatási együttműködés keretében a TU Kaiserslautern és a Fraunhofer IESE anyanyelvű munkatársaival validáltuk (Koch, és mtsai. 2008).” Itt hiányolom a validálás számszerű értékelését.”*

Sajnos a német változat esetén nem volt lehetőségünk nagyobb létszámú formális tesztelésre. Az előző (2.8) válaszomban említettek szerint a „kis erőforrás igényű” megoldás lehetővé tette, hogy egy robot beágyazott számítógépén fusson a szintetizátorunk német nyelvű változata. Annyit tudtunk megállapítani, hogy a német kollégák elfogadták és használhatónak tartották kutatásaik szempontjából a mi rendszerünk minőségét.

2.10 *„43. oldalon: „Kezdeményeztem egy rejtett Markov modell (HMM) alapú magyar nyelvű gépi szövegfelolvasó (TTS) rendszer létrehozását és meghatároztam a modellalkotás lépéseit.” Melyek voltak ezek?”*

A magyar nyelvű HMM rendszer létrehozásához a következő lépéseket határoztam meg:

1. Reprezentatív szövegkorpusz megtervezése és megvalósítása. A szövegkorpusz felolvasásával a magyar hangkapcsolatokat és jellemző prozódiai szerkezeteket tartalmazó hanganyag jön létre.
2. A szövegkorpusz felolvasása több hangon anyanyelvű beszélőkkel (legalább egy férfi és egy nő).
3. A hangfelvételek feldolgozása, címkézése, adatbázisba rendezése
4. A környezet függő címkék meghatározása, az azokat előállítani képes algoritmusok és szoftver komponensek meghatározása
5. A lehető legjobb minőséget az adott számítási kapacitások mellett biztosítani tudó TTS beszédkódoló algoritmus meghatározása, a kapcsolódó szoftver komponensek megvalósítása
6. A fenti pontok szerinti eszközök felhasználásával HMM modellek betanítása.
7. A fenti pontok szerinti eszközök és a betanított HMM modellek segítségével beszédet előállítani képes szoftver komponensek megvalósítása
8. Szabály alapú korrekciós algoritmusok a statisztikai modellek esetleges durva hibáinak korrigálására (pl. alapfrekvencia emelkedés kijelentő mondat végén).

2.11 *„„A számszerű kiértékelés” alcímek esetében többször nem található számszerű adat. Ezeket miért nem közölte?”*

Tudomásom szerint minden altézishez tartozik legalább 1 számszerű eredmény az alábbiak szerint. Ezeket kiegészíti az egyéb, többnyire éles szolgáltatások említése.

I.1 A 11. ábra ad számszerű értékelést.. A német nyelvű változatra kitértem a 2.9 válaszban

[5] https://www.alza.hu/motorola-one-fusion-kek-d5877003.htm

I.2 Szintén a 11. ábra ad számszerű értékelést. A Jaws for Windows és a Robobraille rendszerekben egyaránt több éve működik a ProfiVox diád/triád rendszer érdemi kritika nélkül.

II.1 Számszerű eredményeket a 11. és a 12. ábra mutat.

II.2 Számszerű eredményeket tartalmaz a II.1 tézisnél (a disszertációban a 36.oldalon hibásan hivatkozom az I.1-re). További számszerű eredményeket tartalmaz a 13. és a 14. ábra.

II.3 Számszereű kiértékelést ad a 18. ábra.

III.1 A számszerű kiértékelés megjelenik a 12. ábrán.

III.2 Terjedelmi korlátok miatt a részletes számszerű kiértékelés beillesztése elmaradt, viszont hivatkozik a kapcsolódó publikációkra ([18], [79], [80], [81], [82], [83], [84].)

III.3 A számszerű kiértékelést a 24. ábra tartalmazza.

IV.1 A számszerű kiértékelést a 8. táblázat ismerteti.

IV.2 A számszerű kiértékelés a 29. ábrán található.

IV.3 A számszerű kiértékelés a 32. ábra és a 9. táblázat alapján került ismertetésre.

2.12 *„A 24. ábra (50. oldalon) jelmagyarázata angol nyelvű. Érdemes lenne a megfelelő magyar szavakkal helyettesíteni, még akkor is, ha a szövegben magyarul megjelennek a négyelemű skála pontjai."*

Köszönöm bírálóm megjegyzését. Elmulasztottam az ábrát újra rajzolni, mikor az eredeti angol nyelvű publikációból [86] átvettem. Ezúton pontosítok:

A 24. ábra felső sorában szereplő döntési alternatívák jelentése balról jobbra:

Kijelentő Inkább kijelentő Inkább kérdő Kérdő (mondat)

3. Összefoglaló és javaslat a nyilvános vitára bocsátásról

„II.3. tézist azzal a megjegyzéssel, hogy a 'prozódia' terminus használata, illetve az ehhez kapcsolódó kifejezések definícióinak hiánya bizonytalanná teszi a „szegmentálás" módját."

Köszönöm Dr. Gósy Mária professzor asszony pozitív és támogató véleményét. Remélem, hogy a 2.5 pontban adott válaszom szerinti definíció elfogadható a számára.

Budafok, 2021. 07. 10.

dr. Németh Géza