Opponensi bírálat

Németh Géza: Célorientált gépi beszédkeltés interakciós rendszerekben
című doktori értekezéséről

Németh Géza értekezésének témája nagyívű, összekapcsolja a gép beszédkeltés fejlesztési eredményeit, valamint azt az utat, illetve egyes állomásait, amelyek a kezdetektől napjainkig jellemzik ezt a diszciplínát. Egyfelől elemzi a mesterséges beszédelőállítás különféle aspektusait, módozatait, másfelől pedig az idő függvényében mutatja be a műszaki, technológiai, számítástechnikai fejlődés azon mérföldköveit, amelyek hatással voltak a gépi beszédkeltésre, és amelyeket felhasználtak a fejlesztések során. Ez az elismerésre méltó precíz, értő áttekintés a maga nemében újdonság, és fontos tudományos értéket képvisel.

Az egyes fejlesztési szakaszok bemutatása, az adott időszakok módszertanának leírása és az eredmények értékelése mintegy egymásra épülve bontakoztatja ki a gépi beszédkeltés rendszereit, jellemzőit és alkalmazásait, valamint annak a folyamatos kreatív munkának a részleteit, amely lehetővé tette a jelenleg legkorszerűbbnek tekinthető, nemzetközi mércével is elismerendő eredményeket. Különösen hangsúlyozandók a disszertáció utolsó fejezetében leírtak, amelyek a gépi beszédelőállítás eredményeinek alkalmazásait tárgyalják (9. fejezet, 63–89 oldal). Az itt ismertetett gyakorlati alkalmazások szemléletesen láttatják, hogy a korábban elemzett rendszerek hogyan funkcionálnak a mindennapi életben, és milyen meghatározó szerepet kapnak különféle területeken.

Mindezek alapján megállapítható, hogy a disszertáció témája időszerű, a kísérletektől, a rendszerfejlesztéseken át az alkalmazásokig egységes keretben tárgyalja a mesterséges beszédelőállítás legfontosabb periódusait és azok eredményeit.

A mesterséges beszédelőállítás ciklusai voltaképpen tükrözik a beszédkutatás három nagy forradalmát. Az első a beszéd láthatóvá tétele volt, a második a számítógépek megjelenése, illetve az azokon alapuló algoritmusok, módszerek fejlesztése a fonetikai vizsgálatokhoz, a harmadik a tárkapacitás rendkívüli mértékű megnövekedése, amely mind a kutatási irányokat, mind az alkalmazásokat alapvetően befolyásolta. Németh Géza értekezése kimondatlanul is alátámasztja ennek a három nagy jelentőségű ténynek a megjelenését a gépi beszédkeltésben. Ahhoz, hogy a beszéd mesterségesen előállítható legyen, a Kempelen beszélőgépe óta eltelt több mint 200 évben nagy változásoknak kellett bekövetkezniük. Az igazán hasznos felismerések az elmúlt mintegy 70 évben történtek. A kutatóknak meg kellett ismerniük a beszéd akusztikai reprezentációját, a beszédhangok, hangkapcsolatok, szavak, majd a folyamatos beszéd akusztikai-fonetikai jellemzőit, hiszen ezek nélkül a mesterséges előállítás fel sem merülhetett. A magyar nyelv hangzó változatának ilyen jellegű megközelítése mintegy hatvan évvel ezelőtt kezdődött, majd a 20. század nyolcvanas éveitől rendkívüli fejlődés következett be. A múlt század utolsó évtizedeiben használt számítógépek kapacitása, illetve hardver, szoftver jellemzői egyfelől lehetővé tették a mesterséges beszédelőállítás elindulását, de éppen a módszertani (hardver, szoftver, tárkapacitás) korlátok eléggé meg is nehezítették a kutatók munkáját. A komputerek, a számítástechnika, a kapcsolódó más technológiák igen gyors fejlődése az utóbbi fél évszázadban mind-mind újabb lehetőséget adott a gépi beszédkeltés egyre jobb minőségű megvalósítására. Ennek a fejlődési ívnek és benne a saját kutatómunkájának az egzakt bemutatása a jelölt disszertációjának tartalma.

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.

Elismerés illeti a jelöltet a gépi beszédelőállítás eredményeinek tekintetében is; ezek önmagukért beszélnek, a gyakorlati megvalósítások pedig kétségtelen letéteményesei az elvégzett munka hasznának. A tézisek kérdésfelvetései jók, az elemzések a téziscsoportokon belül logikusak, jól egymásra épülnek, átláthatók, az alkalmazott módszerek leírása megfelelő, a szemléltetés kiváló. A célkitűzés jól megfogalmazott, egyértelmű. A disszertáció koherens egész, felépítése arányos, a kitűzött cél eléréséhez vezető utat minden fejezet koherensen összekapcsolva jeleníti meg. 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.

A disszertáció a gépi beszédkeltés különböző megközelítéseivel, rövid történeti áttekintéssel indul. A kutatási célkitűzéseket az eszközök és módszerek, illetve az adatbázisok bemutatása követi. A kutatások módszertana kissé általános modellt használ, illetve röviden kitér a technológiai fejlődés hatásaira a beszédszintézis vonatkozásában. Ezt követik a téziscsoportoknak megfelelő vizsgálatok prezentálásai: a diád és triád elemek összefűzésén alapuló gépi szövegfelolvasás (I. téziscsoport), a célorientált, korpusz alapú gépi felolvasó rendszerek (II. téziscsoport), a statisztikus parametrikus gépi szövegfelolvasó rendszerek (III. téziscsoport) továbbá a multimodális beszédinformációs rendszerek (IV. téziscsoport). Ezeket a fejezeteket követik az eredmények alkalmazásai, műszaki alkotások, közcélú beszédinterakciós rendszerek bemutatásai. A munkát az összegzés és az irodalomjegyzék zárja.

Az elismerés és a pozitív értékelés mellett azonban szeretném felhívni a jelölt figyelmét néhány problémára, amelyek megoldása még jobbá teheti a munkáját. A kritikai megjegyzéseimet nem súlyozottan ismertetem, vannak közöttük lényegesek és talán kevésbé lényegesek, de úgy gondolom, valamennyi fontos annyira, hogy a jelölt mérlegelje őket.

Számos tudományterületre jellemző (különösen napjainkban), hogy a kutatáshoz, a vizsgálatokhoz team-re, azaz kutatócsoportra van szükség. A modern mérnöki tudományok egyébként is fokozottan alkalmaznak más tudományterületi eredményeket. Különösen indokolt a közös munka a nagyon komplex jelenségek, folyamatok, mechanizmusok megismerésében, a fejlesztések megvalósításában. A beszéd rendkívül összetett jelenség, tanulmányozása csak interdiszciplinárisan, több tudományterület képviselőinek együttes, közös munkája révén jöhet létre. Nem véletlen, ahogyan a disszertáció bibliográfiájából is látszik, hogy a jelölt szerzőtársakkal dolgozott egy-egy (rész)téma megoldásában. Az is elfogadható, hogy az értekezésben túlnyomórészt egyes szám első személyben taglalja a munkákat és az eredményeket. (A 7.3. fejezetben többes szám első személyt használ.) 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.

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. 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.

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. 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.

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.

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.

Néhány esetben hiányérzetem maradt az értekezés olvasásakor, a következőkben felsoroltakra kérném a jelölt válaszait.

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.

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?

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?

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.

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 számszerű kiértékelés” alcímek esetében többször nem található számszerű adat. Ezeket miért nem közölte?

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. Itt sem találtam leírást statisztikai vizsgálatra (pl. Likert-skála alkalmazása?).

Összegző vélemény:

Megállapítom, hogy Németh Géza doktori disszertációja nemzetközi tekintetben is új eredményeket tartalmaz, amelyek nagymértékben hozzájárulnak a tudományszak további fejlődéséhez. A vonatkozó mérnöki, illetve a határtudományok területeit tekintve is kiváló tudományos eredményekről ad számot, amelyek a gyakorlatban is felhasználódtak. A disszertáció minden tekintetben megfelel a doktori munkákkal szemben felállított követelményeknek.

Elfogadom az

I. téziscsoportot, a

II.1. és II. 2. téziseket, a

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

Elfogadom a

III. és a IV. téziscsoportokat.

Németh Géza doktori disszertációját nyilvános vitára alkalmasnak tartom.
A jelölt számára az MTA doktora cím odaítélését javaslom.

Prof. Dr. Gósy Mária
tudományos tanácsadó, egyetemi tanár
Nyelvtudományi Intézet, ELKH – ELTE Fonetikai Tanszék

Budapest, 2021. március 26.