Az MI által generált beszéd jellemzői

DOI: 10.46522/S.2024.02.6

GYÉRESI Júlia PhD
University of Arts, Târgu Mureș
juliagyere@gmail.com

Abstract: Characteristics of AI-Generated Speech
Success in life is not primarily attributed to cognitive skills. Rather, it is the level of development in social skills that creates real connections, forms relationships, and keeps attention engaged. There are a few human abilities that, for the time being, can still compete with artificial intelligence: our capacity for critical thinking, and our intelligence for dealing with human emotions and relationships. Machines are better at active attention, as they are able to pay attention continuously, without interruption. In our attention-deficient society, this is a fact of prime importance. Speech generated by AI can vocalize written text and segment it according to punctuation marks, but it has not yet mastered the vocal exploration of deeper interpretive connections. Will it be capable of doing so in the future? Let's examine the vocal characteristics of human speech and draw parallels with the opportunities offered by AI-generated speech synthesis applications! With the advancement of human robots, personal and technological developments will make speech modulations, the creation of individual tone, and the feeding of memories a part of the algorithm in the near future.

Key words: *AI; speech; speech synthesis applications.*

Minden kétséget kizáróan a 21. század első negyedének határmezsgyéjén a technofuturisták, a tudósok és a közgazdászok körében az egyik legsarkalatosabb kérdés az, hogy miként fejlődik és változik a társadalmunk az egyén és a közösség szintjén az MI egyre hangsúlyosabb térnyerésének következtében. Vajon társevolúcióra és együttműködésre kell-e számítanunk az emberek és a mesterséges intelligencia által működtett humanoidok között, vagy inkább összecsapásokra és lehetséges konfliktusokra? Természetesen mindez elsősorban attól függ, hogy az MI eléri-e, és főként milyen módon az autonóm gondolkodás és viselkedés szintjét, képes lesz-e önrendelkező módon függetlenítenie önmagát az adatfeldolgozó rendszerek algoritmusaitól.

Az emberi intelligencia utánzására és fejlesztésére történő törekvés egészen a számításokat végző gépek történetének kezdetéig, az abakuszig vezethető vissza. A modern számítógépek közel egy évszázada jelentek meg, de azóta szédítő üteművé

vált a fejlődésük. Ha arra gondolunk, hogy amikor 1960-ban feltalálták az integrált áramköröket, akkor egy átlagos integrált áramkör két tranzisztort tartalmazott, ma már a legfejlettebb integrált áramkörben több mint 20 milliárd található (Tilesch és Hatamleh 2021, 21).

A legelső szövegfelolvasó alkalmazás a VODER (Voice Operating Demonstrator) volt, amelyet a Bell Labs fejlesztett ki az 1930-as évek végén. Az alkalmazás feltalálója Homer Walter Dudley amerikai elektronikai és akusztikai mérnök, aki az első elektronikus hangszintetizátort 1939-ben a New York-i világkiállításon mutatta be. Szimulálta az emberi beszédet elektronikus áramkörök és kézi vezérlés segítségével. A VODER nem volt teljesen autonóm, magasan képzett kezelőre volt szükség a gép kezeléséhez. A kezelő tízbillentyűs billentyűzettel és lábpedállal vezérelte a VODER-t, amely létrehozta a hangot, és folyamatosan meghatározta, hogy milyen beszédhangok keletkezzenek. A kezelő a billentyűzetet használta a különböző beszédkomponensek, például magánhangzók, mássalhangzók és a hangartikuláció egyéb aspektusainak kiválasztására. A billentyűzet minden billentyűje meghatározott beszédhangoknak (fonémáknak) felelt meg, és a kezelőnek úgy kellett játszania a gépen, mint egy hangszeren, hogy koherens mondatokat alkothasson. A lábpedál szabályozta a hang magasságát, hasonlóan ahhoz, ahogyan az emberi hangszálak szabályozzák a hangmagasságot beszéd közben. A hangmagasság beállításával a kezelő olyan intonációt és dallamot tudott előállítani, amely elengedhetetlen a beszéd természetes hangzásához. A hang- vagy zajgenerátor által keltett hangot ezután a kezelő úgy alakította, hogy utánozza a természetes beszéd minőségét. Különböző magánhangzók, mássalhangzók és intonációk kombinálásával a VODER érthető szavakat és mondatokat tudott előállítani. A magánhangzókhoz hasonló zöngés hangokat a hanggenerátor segítségével hozták létre, miközben a kezelő úgy manipulálta a hangmagasságot, hogy az megfeleljen az emberi intonációnak. A zöngétlen hangokat, például az sz vagy s hangzót a zajforrás generálta, és a kezelő a szűrőket használta a hang finomításához. Mivel a VODER teljes mértékben kézi vezérlésű volt, a kezelőknek alapos képzésen kellett részt venniük, hogy elsajátítsák a gép működési elvét. Minden egyes hang és szó a kezelőszervek precíz összerendezettségét követelte meg, így az nehéz és munkaigényes folyamat volt. Jól működtetve azonban a VODER meglepően tiszta és érthető beszédet tudott produkálni, jelentős áttörést jelentett a beszédszintézisben, és megalapozta az elektronikus beszédgenerálás későbbi fejlesztéseit, ami olyan teljesen automatizált rendszerekhez vezetett, mint a későbbi VOCODER, és végül eljutottak a modern szövegfelolvasó rendszerek szintjére. Habár mai mércével mérve a VODER kezdetlegesnek számított, mégis mérföldkő a mesterséges beszéd generálása felé vezető úton.

De ne feledkezzünk meg a gépi beszéd-előállítás előfutáráról sem, a magyar származású Kempelen Farkas feltalálóról, aki a tudományos alapokat az 1791-ben megjelent tanulmánykötetében már lefektette. Találmányát huszonkét éven át tökéletesítette, és az egyik legkorábbi kísérlet volt az emberi beszéd mechanikus reprodukálására, az emberi beszédhang utánzására. Kempelen mechanikai találmányairól elhíresült polihisztor volt, világszerte ismert leghíresebb alkotása

mindmáig egy sakkozó automata. De az általa létrehozott artikulációs beszédszintetizátor sokkal jelentősebb: a modern fonetika megalapozójává vált általa, szerkezete lényegében az emberi hangrendszer mechanikus modellje volt. A következő kulcselemekből állt: 1. Fújtató (tüdőt jelképező) – a légzőrendszerként funkcionáló fújtató a levegőt átnyomta a „szélláda" belsejébe, egy légmentesen lezárt fadobozba. 2. Zöngeképző rezgőnyelv (hangszálak) – szabályozta a rezgéseket, utánozta a hangszálakat, és hangot generált. 3. A garatnak megfelelő cső, amelyből két kis cső indult felfelé az orrjáratokat modellezve. 4. Rezonátorok és üregek (a szájüreget formázó gumitölcsér és az orrjáratok) – ezek a részek beállíthatóak voltak különböző hangok formálására, hasonlóan ahhoz, ahogyan az emberi száj, nyelv és ajkak magán- és mássalhangzókat alkotnak. 5. Karok és szelepek – Kempelen gépe lehetővé tette a légáramlás beállítását, elősegítve a különböző beszédhangok kialakítását. A gépet két kézzel lehetett működtetni, és rendkívül sok gyakorlást igényelt az összerendezett mozgássorok begyakorlása a kezelő karok és kapcsolók sorozatának irányításával. A gép a magánhangzók és a mássalhangzók korlátozott tartományát tudta generálni, szótagokat, szavakat, rövid mondatokat tudott előállítani, de messze nem volt tökéletes a teljes emberi beszéd reprodukálására. Kempelen az eredményeit a *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine (Az emberi beszéd mechanizmusa beszélőgépének leírásával)* címmel publikálta. A beszédkeltő gépet 2001-ben építették meg újra Nikléczy Péter és Olaszy Gábor fonetikusok és beszédkutatók eredeti nagyságában és működőképesen (Nikléczy és Olaszy 2016). Kempelen Farkas munkája megalapozta a beszédszintézis és a fonetika későbbi fejlesztéseit, úttörő erőfeszítései a mesterséges beszéd későbbi technológiáit inspirálták, beleértve a modern beszédszintézis- és hangfelismerő rendszereket.

A ma használatos szöveg-beszéd (TTS, Text-to-Speech) generátorok olyan technológiai rendszerek, amelyek írott szöveget alakítanak beszéddé. Ez a technológia különösen hasznos olyan területeken, mint a kommunikációs akadályokkal küzdők segítése, a mobilasszisztensek (pl. Siri, Google Assistant), automatikus hívásrendszerek vagy éppen e-könyvek hangos felolvasása. A szöveg-beszéd generátorok fejlődése az elmúlt években jelentősen felgyorsult a mesterséges intelligenciának és a gépi tanulásnak köszönhetően. Vegyük sorra, hogy miként működik a szöveg-beszéd generátor: 1. A szöveg feldolgozása: az első lépésben a rendszer elemzi a bemenetként adott szöveget, figyelembe veszi a mondatok szerkezetét, a szavak jelentését, és különös figyelmet fordít a kiejtésre, az írásjelekre, valamint a hangsúlyokra. 2. Fonémák és prozódia: a rendszer a szöveget fonémákra, azaz beszédhangokra bontja, majd kiválasztja azokat a hangmintákat, amelyek a legjobban megfelelnek az adott nyelv szabályainak és az adott mondat kontextusának. 3. Szintetizálás: a rendszer a feldolgozott információ alapján létrehozza a beszédet, amely során figyelembe veszi a természetes intonációkat, a ritmust, és a mondatok érzelmi tartalmát.

Két fő technológia létezik a TTS generátoroknál: 1. Kézileg szintetizált modellek: ezek olyan korábbi rendszerek, amelyek előre rögzített beszédhangokból építkeznek, és a hangminta adatbázis kombinálásával hoznak létre mondatokat. Az ilyen rendszerek

esetenként monotonnak tűnhetnek, mivel a hang kifejezési lehetőségei korlátozottak. 2. Neurális hálózatok által vezérelt modellek: az újabb fejlesztésű TTS rendszerek mesterséges intelligenciát, főként neurális hálózatokat használnak. Az olyan modellek, mint a Tacotron vagy WaveNet, képesek sokkal természetesebben és kifejezőbben beszédet generálni, mivel nagyobb adatbázisokkal és gépi tanulási algoritmusokkal dolgoznak.

Az alkalmazások és a felhasználási területek közül néhány:

Oktatás: segíthetnek a vizuálisan sérült vagy olvasási nehézségekkel küzdő emberek számára hozzáférhetőbbé tenni az írott anyagokat.

Ügyfélszolgálat: automatikus telefonos rendszerekben és chatbotokban is alkalmazzák, hogy természetesebb, emberibb beszédélményt nyújtsanak.

Navigációs rendszerek: GPS-alapú útvonaltervezőkben gyakran használják a TTS rendszereket, hogy valós időben mondják el az útvonal-instrukciókat.

A TTS technológia fejlődése révén egyre természetesebb, kifejezőbb és pontosabb beszédgenerálás érhető el. A szövegek dekódolása során a beszéd szegmentális tartománya mint lineáris szegmensek sorozata (hangok, hangsorok, hangkapcsolatok) kerül meghangosításra. Erre tevődik rá az értelmezés során a szupraszegmentális elemek sora, és ezek az elemek befolyásolják a beszéd ritmusát, dallamát és kifejezőerejét, gyakran hordoznak fontos jelentésárnyalatokat, közvetítik a szövegek többletjelentését. A legfontosabb szupraszegmentális elemek: hangsúly, hanglejtés, időtartam, ritmus, hangszín, szünettartás, hangerő, hangmagasság, beszédsebesség. És ez az a területium, amely felelős azért, hogy a beszéd árnyalt, természetes, érthető és sokrétű jelentéstartalommal bíró legyen. A szöveg-beszéd generátorok nem rendelkeznek érzelmi intelligenciával, mert gépek, nem tudnak oly módon érzelmesek és kifejezőek lenni a hangi megnyilatkozások során, mint egy ember. Tettem egy próbát Ady Endre *Akik mindig elkésnek* című versével: Latinovits Zoltán hangját klónozta a Vocloner alkalmazás. Első hallásra nehéz volt megállapítani, hogy ki mondja a verset: a színész vagy pedig a gép.

A nyelv megértése nem egy passzív folyamat, értelem és lélek is kell hozzá. Az anyanyelv elsajátítása a született nyelvelsajátítási képességeinken múlik. A megértés és a nyelvképzés során számtalan folyamat játszódik le az ember agyában és lelkében. Ha elmondunk egy verset, akkor az előadásmódunk a temperamentumunktól, az értelmi, az érzelmi, a kreatív intelligenciánktól függ, az addigi élettapasztalatainktól, a pillanatnyi pszichés állapotunktól és a tehetségünktől.

Mi volt az alapvető különbség a két előadásmód között?
A prozódikus jelzések: hangerősség, hangmagasság, a beszéd gyorsasága, ritmus, hangszín, hanglejtés, intonáció, hangsúly, szünet, időtartam az emberi beszédet utánozták a mesterségesen generált versmondás során is. A tagolás az írásjeleknek megfelelően történt. Mi volt mégis más?
Hibák: erős nyomatéki hangsúly az igén, túl gyors a felsorolás felolvasásának a ritmusa, túl erőtlen a szókezdő hangsúly a névszókon. Hiányok: sóhaj, légvétel, érzelmi

telítettség nem színezi az előadásmódot, nem alkalmazza a különféle nyomatékolási módokat a vers tartalmának megfelelően. Sokkal sterilebb, differenciálatlanabb volt a Vocloner generálta előadásmód, mint a Latinovits Zoltáné.

Természetesen az emberi hangok klónozásának technikai fejlettsége napról napra finomodik, cizellálódik, akár más nyelven is képes az utánzásra a Voice Cloning. A mesterséges intelligencia a betáplált szavak, mondatok, összefüggő szövegminták alapján feltérképezi az adott személy hangjának egyedi jellemzőit. Megtörténik az akusztikus, majd a nyelvi modellezés. A hang elemeire bontása és megértése után képes az újraalkotásra, legyen szó bármilyen nyelvű szövegről, amennyiben elegendő információval rendelkezik az adott nyelv szabályairól. Záró akkordként a szintetizált hangot finomhangolják, hogy hű másává váljon az eredetinek, az egyedi kiejtési és nyelvi sajátosságokat is beleértve.

Raymond Kurzweil amerikai feltaláló, író, az optikai karakterfelismerés, a szöveg-beszéd szintézis, a beszédfelismerés és az elektronikus billentyűs hangszerek úttörője a *Hogyan alkossunk elmét?* című tanulmánykötetében így fogalmaz: „Az emberek gyakran fenyegetésként élik meg azokat a vitákat, amelyek felvetik annak a lehetőségét, hogy egy gép is rendelkezhet tudattal, mivel az ilyenfajta megfontolásokat a tudattal rendelkező lények spirituális értékének a becsmérléseként értékelik. Ám ez a reakció a gép fogalmának a félreértését tükrözi. Az ilyen kritikusok a témát a gépek mai tudása alapján ítélik meg, és azok bármennyire lenyűgözőek legyenek is, egyetértek abban, hogy a technológia jelenlegi példányai még nem méltóak arra, hogy tudatos lényekként tiszteljük őket. Az a jóslatom, hogy nem lesznek megkülönböztethetőek a tudatos lényeknek tekintett biológiai emberektől, éppen ezért a gépek is osztozni fognak abban a spirituális értékben, amelyet a tudatosságnak tulajdonítunk. Ez nem az emberek lebecsülése, inkább a jövő (egyes) gépei iránti megértésünk felértékelődése. Valószínűleg különböző terminológiát kell kidolgoznunk ezekre az entitásokra, mivel másfajta gépek lesznek. (…) Az emberek már most is alkotnak spirituális gépeket. Sőt olyan szinten egybeolvadunk az általunk készített eszközökkel, hogy az ember és gép közötti különbségtétel egyre inkább elmosódik, míg végül meg is szűnik. Ez a folyamat már nagyban zajlik, még akkor is, ha a legtöbb, bennünket kibővítő gép még nincs a testünkben és az agyunkban” (Kurzweil 2022, 210). A közeljövőben már eldől, hogy képes lesz-e kiváltani minket az MI? Amikor pontosan meghatározhatóvá válik az MI széles spektrumot felölelő szerepe – az orvoslás, az oktatás, a szórakoztatás, a szállítás, a közbiztonság, a balesetvédelem, a művészet területén –, akkor talán a szabályozhatósága is lehetségessé válik.

KÖNYVÉSZET

KURZWEIL, R., 2022. *Hogyan alkossunk elmét?* Budapest: Pallas Athéné Books.

NIKLÉCZY, P. és OLASZY, G., 2016. Kempelen Farkas beszélőgépének rekonstrukciója. *Nyelvtudományi Kutatóközpont,* No. 4.

TILESCH, Gy. és HATAMLEH, O., 2021. *Mesterséges intelligencia.* Budapest: Libri Kiadó.