Hangkorverzió alkalmazása dysarthriás betegek beszédminőségének javítására

Terbe Dániel[1], Tóth László[1], Ivaskó Lívia[2]

[1]Szegedi Tudományegyetem, Informatikai Intézet
[2]Szegedi Tudományegyetem, Informatikai Intézet
[2]Szegedi Tudományegyetem, Általános Nyelvészeti Tanszék
tothl@inf.u-szeged.hu

Kivonat A dysarthria egy gyűjtőfogalom az artikulációs nehezítettségből eredő beszédzavarra, amelynek hátterében számos betegség állhat. A dysarthriás személyek beszédének minősége, érthetősége leromlik, ami az érintettek szociális kapcsolataira és így életminőségére is rossz hatással lehet. A hangkonverziós technológia fejlődésével felvetődik az ötlet, hogy vajon lehetséges-e ezen betegek hangfelvételeinek minőségét, érthetőségét gépi eszközökkel feljavítani, és beszédkommunikációjukat egy ilyen elven működő eszközzel támogatni. Cikkünkben áttekintjük a (neuronhálós) hangkonverziós algoritmusok fő változatait, majd bemutatjuk a dysarthriás betegek felvételein végzett kísérleteink tapasztalatait, és ezek alapján megvitatjuk az egyes módszerek legfontosabb előnyeit és hátrányait.
Kulcsszavak: dysarthria, hangkonverzió, mély neuronhálók

1. Bevezetés

A beszéd tervezésének és kivitelezésének zavarai komoly kommunikációs akadályt jelentenek azon személyek számára, akik szerzett vagy fejlődési eredetű beszédzavarral küzdő kommunikációs partnerként szeretnének interakcióba lépni társaikkal. A különböző neurodegeneratív betegségek, illetve egyéb eredetű idegi károsodások eredményeképpen kialakuló fonációs és artikulációs zavarok úgy befolyásol(hat)ják átmenetileg vagy tartósan, egyes esetekben pedig egyre fokozottabban megjelenő (progrediáló) formában a beszélők beszédprodukcióját, hogy az egyébként tartalmilag és nyelvtanilag jól formált, grammatikus beszéd az interakciós partnerek számára nagyon nehezen, vagy alig érthető hangokként jelenik meg (Horváth és Hirshberg, 2013). Ez a nehéz érthetőség nagy mértékben csökkenti az érintett személyek önálló életvitelre való képességét, súlyos életminőség-romlást tud előidézni. A jobb életminőség eléréséhez és a megfelelő érthetőség szempontjából is fontosnak tartjuk azon lehetőségek számbavételét, melyek az érintettek hangminőségbeli javítását célzó törekvéseket kívánnak a számítógépes nyelvészeti és a mesterséges intelligencia kutatás eszköztárával elősegíteni. Fontosnak tartjuk az érintett populáció társadalmi reintegrálhatósága

szempontjából a beszédteljesítményük javítását célzó kutatásban való aktív részvételt (Tóth és mtsai, 2018).

Már Aronson 1981-es összefoglaló munkája (Aronson, 1981) is rámutat arra, hogy dysarthria több okból is eredhet, és attól függően, hogy milyen területek sérülése, illetve diszfunkciója áll a dysarthriás beszéd hátterében, eltérő hangminőségbeli tulajdonságok mentén lesznek az egyes előfordulások csoportosíthatóak. Annak függvényében, hogy a beszédhangok és a velük együtt realizálódó szupraszegmentális elemek milyen mértékben tudnak megtartott képességek alapján képződni, el lehet különíteni az egyes dysarthria-típusokat, azonban hazánkban ez az elkülönítés jelenleg a diagnosztikus kritériumok tekintetében elsősorban az oki tényezőket, valamint az érintett traktusnak a beszédre gyakorolt hatását figyelembe vevő módon, a hallási élményt szubjektíve értékelő skálán osztályozza (Horváth és Hirshberg, 2013). Értelmezésükben a kortikális sérülésből eredő dysarthriás beszédfolyamatok például elsősorban úgy jellemezhetők, mint a beszéd primer motoros tervezésének és kivitelezésének nem megfelelő működéséből eredő specifikus mintázatok. A stroke eredetű dysarthria a motoros funkciók érintettségéből eredő beszédzavarnak tekinthető, mely (a stroke kiterjedésétől függően) nem kell, hogy érintse a nyelvi tervezési folyamatokat. A klinikai differenciáldiagnosztika során az ilyen területek érintettségéből eredő hangzókülönbségek auditív úton is elkülöníthetők a hangképzőszerveket érintő más atípusos formáktól Aronson (1981) csoportosítása szerint. Akiknél más (nem stroke eredetű) neurodegeneratív megbetegedés vagy traumatikus agysérülés okoz beszédzavart, a beszéd más összetevőinek, például a monoton beszédnek, vagy hiperkinetikus beszédnek a jegyeit produkálják verbális megnyilvánulásaik során. A hangerő, a hangmagasság és a ritmus is fontos összetevői a beszédnek, melyek a sérülés, illetve betegség eredetétől és helyétől függően mutathatnak változatosságot az eltérésben. Azt mondhatjuk tehát, hogy a dysarthria egy olyan összetett klinikai kép, mely a beszéd egyes összetevőit nem azonos mértékben és minőségben érinti az egyes kórképek esetében (pl. ALS, sclerosis multiplex, stroke, traumatikus agysérülés, kisagyi érintettség, egyes idegbénulások, Parkinson-kór, illetve egyéb idegrendszeri zavarok esetében), de a személy számára jelenthetnek ezek az eltérő formák olyan akadályt, mely miatt nehezen érhető beszéde nem teszi őt képessé a megfelelő verbális kommunikációra. Az emberi beszéd bonyolult folyamatában a beszédjel a megbetegedés alapvető sajátosságaitól, a neuroanatómiai eltérés helyétől és kiterjedésétől függően eltérő módokon torzulhat. A motoros funkciók érintettségéből eredő beszédzavarok közül az artikulációs szervek vezérlésének zavara a hangképzési folyamatot befolyásolja, például hibás formánsszerkezetű magánhangzókat eredményezhet. Ha az artikulációs szervek összehangolása sérül, akkor időben elkent, torzult hangzókat kapunk. A hangok adott ideig és hangmagasságon való kitartásának nehézsége a beszéd szupraszegmentális szintjének, a prozódiának a torzulásaként jelentkezik. Végül, a hangszalagok vezérlésének zavara a hangminőség romlását okozza, ez az ún. diszfónia (Markó és mtsai, 2007) gyakran van jelen a dysarthriával egyidejűleg (Camillo és Ortiz, 2007). Különösen igaz ez az életkor előrehaladtával kialakuló természetes változásokat figyelembe véve.

2. Mély neuronhálós hangkonverziós algoritmusok

A hangkonverzió (voice conversion) egy beszédtechnológiai eljárás, melynek célja egy adott (forrás) beszélő hangfelvételének átalakítása oly módon, mintha azt egy másik beszélő (a célbeszélő) mondta volna (Mohammadi és Kain, 2017). A hangkonverziót gyakran a jóval általánosabban értelmezhető hangtranszformáció (voice transformation) speciális esetének tekintik.

A hangkonverzió fő alkalmazója a szórakoztatóipar (Turk és Arslan, 2002): segítségével utólagosan javíthatóvá, manipulálhatóvá válnak a filmek hangsávjai (pl. pár szó betoldásához nincs szükség az eredeti színészre) vagy egy pontatlanul felénekelt éneksáv, de akár régen elhunyt színészek szájába is új szöveget adhatunk. A szórakoztatóipar mellett a másik fő alkalmazást a telekommunikáció beszédszintézisre épülő ágai jelentik. A hangkonverzió segíthet a személyre szabott beszédszintézisben, például egy telekonferencia-alkalmazásban, akár valós idejű gépi fordítással egybekötve. A személyre szabott beszédszintézis speciális esete az orvosi alkalmazás, amikor a beteg eredeti beszédét próbáljuk visszaállítani (voice reconstruction), például gégeműtét után, vagy esetünkben dysarthria fennállásakor.

Hangkonverzióval már régóta próbálkoznak (Moulines és Sagisaka, 1995), de az igazi fellendülést a területen a mély neuronhálós technológiák megjelenése hozta. A legkorábbi, legegyszerűbb algoritmusok párhuzamos hangfelvételeket igényelnek, azaz a forrás- és a célbeszélőnek ugyanazt a szöveget kell beolvasnia. Jóval későbbiek a párhuzamos korpuszt nem igénylő 'non-parallel' algoritmusok (Kaneko és mtsai, 2021), amelyek a bemenő adatokra nézve jóval nagyobb szabadsági fokot biztosítanak, de ez esetben a gépi tanulási feladat is jóval nehezebb. A hagyományos, egy forrás- és egy célbeszélőt feltételező módszerek mellett próbálkoznak már sokbeszélős konverzióval is ('many-to-many', 'many-to-one') (Kaneko és mtsai, 2019). Nem célunk az összes létező szempont és technológiai megoldás áttekintése, ezért csak azokat a módszereket ismertetjük kicsit részletesebben (a 3. fejezetben), amelyekkel személyes tapasztalatot szereztünk.

2.1. Hangkonverzió dysarthriás beszéd feljavítására

A hangkonverzió szokványos alkalmazása esetén azt várjuk, hogy a konvertált hang lehetőleg minden szempontból hasonlítson a célbeszélő hangjához. Ezt a hasonlóságot azonban nehéz objektív, tudományos módon megfogalmazni, és még nehezebb egzakt mérőszámokkal számszerűsíteni. Mindenesetre a hasonlóság két fő tényezője a hangszín és a hangmagasság, és e két tulajdonságot viszonylag egyszerű módon lokálisan, azaz a jel időbeli lefutásának bolygatása nélkül is lehet módosítani. A dysarthriás beszéd feljavítása esetén azonban kicsit más a cél, mint a szokványos hangkonverziónál. Kiindulási hangként a beteg jelenkori felvételei állnak rendelkezésünkre, amelynek minőségén, érthetőségén szeretnénk javítani. Minden más tulajdonságát azonban lehetőleg szeretnénk változatlanul hagyni, hogy megőrizzük a beteg személyiségét tükröző vonásokat. A optimális megoldáshoz célhangként hangminták lennének szükségesek ugyanazon alany betegséget megelőző állapotáról, ilyen azonban legtöbbször nem áll rendelkezésre

(de nyugaton léteznek már cégek, amelyek ilyen hangarchívum készítését kínálják lassan progrediáló betegségben szenvedőknek). Ilyenkor valamilyen donor hangot kell használnunk célhangként. Itt ismét jó lenne, ha léteznének hatalmas donor hangadatbázisok, melyekben meg lehetne találni az alany hangjához leginkább illeszkedő mintát – de ilyen magyar nyelvű adatbázisól sem tudunk egyelőre. Mivel a célunk a minőség és az érthetőség javítása, a konverzió során alapvetően az artikulációs finommozgást szeretnénk átültetni a dysarthriás felvételre, nem célunk viszont a személyre jellemző hangszín átvitele. Ezzel összhangban alapvetően a hangmagasság átvitelére sincs szükségünk. A helyzet azonban nem ilyen egyszerű, mivel a dysarthria gyakran együtt jár a hangkeltés zavarával is, ilyen esetekben viszont a hangszalagok működését leíró beszédtechnikai komponens, a gerjesztőjel módosítása is szükséges lehet a beszédminőség javításához. Továbbmenve, dysarthriás beszéd esetén legtöbbször a prozódia is sérül, ami egyszerűbb esetekben csak lelassulásként jelentkezik, de súlyosabb szinten jelentősen hozzájárulhat a beszéd érthetőségének romlásához. Az itt bemutatott, kezdeti kísérleteinkben olyan pácienseket választottunk, akiknél a fő problémát az artikuláció elkentsége okozza, de a fentiek érzékeltetik, hogy általános esetben a probléma milyen sokrétű lehet. Az általunk kipróbált módszereket eredetileg elsősorban egészséges beszélők hangszínének felcserélésére fejlesztették ki (lásd előző fejezet), nem pedig leromlott minőségű beszéd feljavítására. Munkánk során részben azt vizsgáltuk, hogy ezek – az eddig egészséges beszéden alkalmazott eljárások – mennyire alkalmazhatók dysarthriás beszéd konverziójára. A szakirodalomban megoldásként javasolt algoritmusokat részletesen a 3. fejezetben tárgyaljuk.

3. Kísérleti konfigurációk és tapasztalatok

3.1. Adatbázisok

A dysarthriás beszéd kutatását nagyban megnehezíti, hogy nemzetközi szinten is kevés a megfelelő méretű dysarthriás korpusz, mivel a felvételek publikussá tételét a szigorú betegjogi/adatvédelmi szabályozás megnehezíti. Mi a kísérleteinkhez főképp az UASpeech nevű, angol nyelvű publikus adatbázist használtuk, amely 15 dysarthriás és 13 kontroll beszélőtől tartalmaz felvételeket, izoláltan kiejtett szavak formájában (Kim és mtsai, 2008). Kísérleteinkben egy-egy típusú leképezést igyekeztünk megvalósítani, amihez kiválasztottunk egy férfi beteget és egy férfi kontrollszemélyt. A rendelkezésre álló tanítóanyag kb. 45 perc volt a beteg részéről.

Emellett magyar nyelvű adatgyűjtésbe is fogtunk, a magyar nyelvű korpusz összegyűjtése jelenleg is folyamatban van[1]. Tapasztalataink között a magyar nyelvű felvételeken kapott kezdeti eredmények konklúziói is megjelennek.

[1] Kutatásetikai engedély: ETIKAI/IV-11043-1 TUKEB Határozat (Dysarthriával élők beszédminőségének vizsgálata), kutatásvezető orvos: Dr. Sandi Dániel (SZTE Neurológiai Klinika)

3.2. Vokóderek

A hangkonverzió bemenete és kimenete is egy hangfelvétel. Habár léteznek már közvetlen hullámforma inputtal és outputtal dolgozó neuronhálók is (Kim és mtsai, 2020), egyelőre szokványosabb a hangfelvételeket valamilyen spektrális reprezentációra konvertálni, ez ugyanis tömörebb tárolást és könnyebb manipulálhatóságot is biztosít. A konverzió tehát alapvetően három fő lépésből áll: a hangot először analizáljuk, "szétszedjük", az analízis eredményének bizonyos paramétereit módosítjuk, végül a hangot szintetizáljuk, újra "összerakjuk". Elsőre technológiai részletkérdésnek tűnhet, de fontos döntés, hogy az algoritmus milyen belső reprezentációt használ, azaz milyen módon analizálja a hangot. A régebbi megoldások valamilyen hagyományos, jelfeldolgozáson alapuló vokódert (voice encoder-decoder) használtak erre a célra (pl. WORLD (Morise és mtsai, 2016)), melyek tipikusan szétszedik a beszédet gerjesztőjelre és spektrális burkológörbére, és ezeket erősen tömörítve reprezentálják. Ennek ára, hogy a rekonstrukció nem tökéletes, azaz a vokóder használata már magában kisebb minőségromlással járhat. Újabban azonban megjelentek a a gépi tanuláson alapuló ún. neurális vokóderek, amelyek sokkal egyszerűbb reprezentációkból – pl. egyszerű mel-spektrogramból – képesek megdöbbentően jó minőségű beszédet szintetizálni (Luong és Tran, 2021; Kumar és mtsai, 2019; Prenger és mtsai, 2019). Mi a MelGAN vokódert alkalmaztuk, amely 80 sávos mel-spektrogramot használ a jel reprezentációjára és abból szintetizál hullámformát. Fontos megemlíteni, hogy tapasztalatunk szerint ezek a vokóderek nyelvfüggetlenek, azaz angol nyelven betanított vokóder képes lesz magyar nyelvű beszéd szintetizálására is.

3.3. Vektorszintű párosítást igénylő módszerek

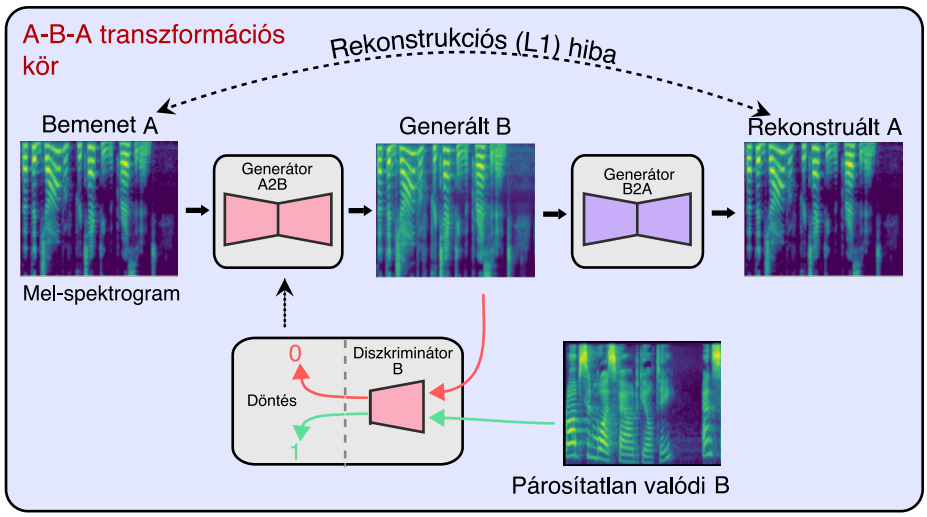
A vokóderre épülő módszerek esetében a forrás- és a célfelvételt az algoritmus egy-egy spektrális vektorsorozat formájában kapja meg. A legegyszerűbb algoritmusok a minták egymásba való transzformálását a vektorok szintjén oldják meg, azaz vektor-vektor leképezést végeznek, vagy legfeljebb rövidebb (maximum szótagnyi hosszúságú) részsorozatokon dolgoznak. Ez csak akkor megvalósítható, ha a forrás- és a célbeszélő hangfelvételei párosítottak, azaz mindketten ugyanazt a szöveget mondják. A tartalmi egyezés azonban nem garantálja a felvételek belsejében a tökéletes szinkronitást. Ezért ezek az algoritmusok előfeldolgozásként igénylik a vektorsorozatok egymáshoz illesztését, amire olyan klasszikus algoritmusokat használhatunk, mint pl. a dinamikus idővetemítés. Cserében viszont – mivel a célértékek vektorszinten vannak definiálva – a vektor-vektor transzformáció nagyon egyszerűen, akár tradícionális neuronhálós megoldásokkal is megvalósítható. A transzformáció hatékonyságát olyan egyszerű hibafüggvényekkel mérhetjük, mint pl. az átlagos négyzetes eltérés (MSE hiba). Az egész terület egyik kulcsproblémája azonban, hogy az elvégzendő átalakítást lehetetlen egyszerű matematikai eszközökkel definiálni. Ezért a transzformációs hiba mérésére bevethetünk egy második neuronhálót, amelyet a tiszta és a dysarthriás artikuláció megkülönböztetésére tanítunk be. Ez vezet el az ún. generatív ellenséges neuronhálók (GAN) konstrukcióhoz, ami a képfeldolgozásban nagyon sikeresnek

bizonyult. Az általunk kipróbált módszer neve MMSE DiscoGAN (Purohit és mtsai, 2020). It a neurális architektúra egy egyszerű négyrétegű teljesen kapcsolt hálózat, amelynek a bemenete az eredeti cikk alapján 40 darab MFCC együtthható (Mel Cepstral Coefficient), de kipróbáltuk 80 sávos mel-spektrogrammal is. Ennél a módszernél a hibafüggvény két elemből tevődik össze: (1) a céltól való átlagos négyzetes eltérés (MSE); (2) egy összetettebb adversarial hiba. Az előbbi hibafüggvény használja ki a párosított tulajdonságot, hiszen ott a bemenethez társított célvektor elérésére törekszünk a hibatag minimalizálása során. Az utóbbi pedig a szakirodalomban DiscoGAN (Kim és mtsai, 2017) néven fellelhető GAN alapú tanítási módszer. Ez egymástól függetlenül és egyszerre jelent meg a CycleGAN technikával (Zhu és mtsai, 2017), amely később aztán jobban elterjedt. Kísérleteink során kipróbáltuk a második tag elhagyását is, tehát egyszerűen négyzetes eltérés alapján tanítottunk, ami valóban kicsit rosszabb eredményekhez vezetett – ez azt mutatja, hogy az adversarial hibatag beépítése a párosított tanításba ténylegesen hasznos lehet.

3.4. Párosítást nem igénylő módszerek

A képfeldolgozásban nagy sikert aratott az ún. neurális stílustranszfer, amellyel például fotóinkat Von Gogh-stílusú festményekké alakíthatjuk. Ilyenkor a képen viszonylag csekély módosítást kell végezni, hiszen maga a tartalom megőrzendő. A hangkonverzió célja is hasonló, csak hangfelvételekkel: a felvétel hangszínének adott beszélőhöz való igazítása a nyelvi tartalom megtartásával. A képfeldolgozásban erre a célra az ún. feltételes GAN-okat (conditional GAN, cGAN) alkalmazzák, ahol a 'feltétel' tulajdonképpen maga az input kép. A GAN technológia nagy előnye, hogy a tanításához nem feltétlenül kellenek input-output párok. Ennek megfelelően a képi stílustranszferben bevált GAN-okat természetesen hangkonverzióra is megpróbálták alkalmazni (Yang és Chung, 2020), akár párosított felvételek nélkül, ami a tanítás során jóval nagyobb szabadsági fokot ígér. így elméleti szinten például könnyedén megvalósíthatóvá teszi a 'many-to-many' leképezést is. Ennek a nagy szabadságnak azonban megvan a hátulütője: mivel nincsenek vektorok szintjén definiált célértékek, rendkívül nehéz tudtára adni a hálózatnak, hogy milyen jellegű módosításokat szeretnénk elérni, és mely módosítások nem kívánatosak. Technikailag különféle jellegű megszorítások megadásával szokták a hálózatot a jó irányba 'terelgetni'. A nagy szabadsági fok miatt ráadásul ezeknek a módszereknek a stabil betanításához jóval több tanítópélda szükséges és maga a tanítási folyamat is hírhedten nehéz, érzékeny a paraméterbeállításokra, ráadásul sokkal nagyobb a számításigényük is (egy tanítás napokig is eltarthat).

A legelterjedtebb párosított adathalmazt nem igénylő tanítási módszer az ún. CycleGAN (Zhu és mtsai, 2017), amelyet az 1. ábrán szemléltetünk. A modellben két generátorhálózat van, melyek a két (forrás és cél-) tartomány közötti oda-vissza transzformációért felelősek. Emellett mindkét tartományhoz van egy-egy diszkriminátor-hálózat, melynek feladata annak eldöntése, hogy a beadott minta az adott térhez tartozik-e (a spektrogram az A vagy B beszélőhöz tartozik-e). A diszkriminátor hálózatok tanítása úgy történik, hogy mutatunk neki valódi



1. ábra: A CycleGAN tanítási módszer illusztrációja. Az ábrán csak az A-B-A irányt mutatjuk, de ugyanez (párhuzamosan) a B-A-B irányban is megtörténik.

(ténylegesen a tartományból vett) és hamis (a másik tartományból átalakított) mintákat. A diszkriminátor célja ezek megkülönböztetése, míg a generátor hálózatok arra vannak tanítva, hogy képesek legyenek becsapni a diszkriminátor hálózatot (egyre jobb és jobb minőségű, élethűbb minták generálásával). Tehát a generátor és diszkriminátor hálózatok egymás ellenében vannak tanítva és versengenek egymással, innen a "generatív ellenséges neuronháló" elnevezés.

A feltételes GAN esetén kulcskérdés, hogy párosítatlan tanítópéldák mellett hogyan tudjuk garantálni az eredeti tartalom megőrzését – vegyük észre, hogy a diszkriminátor ezt nem oldja meg. Erre szolgál a modell ciklikussá tétele: a generátor hálózatoknak teljesíteniük kell azt a megszorítást, hogy az oda-vissza transzformáció után vissza kell kapnunk a kiindulási mintát (rekonstrukciós hiba minimalizálása). Ez a feltétel hozza létre a domainek közötti egy az egyhez való leképezést, illetve teszi lehetővé, hogy ne véletlenszerű mintagenerálás történjen, hanem az adott bemenethez tartozó másik térbeli párt kapjuk.

Tesztjeink során két párosítást nem igénylő eljárást próbáltunk ki. Az első cikk szerzői kifejezetten a dysarthriás beszéd javítását célozták meg, és állításuk szerint módszerük felülmúlja a CycleGAN technikát (Chen és mtsai, 2018). A metódus három neuronhálót alkalmaz: egy generátort, egy diszkriminátort és egy kontrollert. A kontroller egy tömör kódban reprezentálja a bemenő beszédet, amelyből aztán a generátor ismét beszédet készít (ez lényegében egy autoenkóder), miközben a diszkriminátor arra sarkalja a generátort, hogy minél élethűbb mintákat gyártson. Először nagy mennyiségű (kb. egy napnyi) egészséges adaton tanítják a rendszert (ilyen adatból sokkal könnyebb nagy mennyiséget gyűjteni) és miután a generátor megtanult jó minőségű egészséges beszédet produkálni,

kisebb méretű dysarthriás adatbázison tanítják tovább már csak a kontroller részt.

A másik, MaskCycleGAN-VC elnevezésű módszert sima hangkonverzióra alkalmazták a kiindulási cikkben (Kaneko és mtsai, 2021). A modell a CycleGAN-VC legfrissebb, negyedik generációs változata, ami a CycleGAN (eredetileg képekre kifejlesztett) módszer hangkonverziós feladatra szabott változata.

3.5. Mondatszintű párosítással dolgozó módszerek

A párosítatlan felvételekkel dolgozó GAN-okhoz képest első pillantásra visszalépést jelentenek a sorozatból sorozatba leképező ún. sequence-to-sequence (seq2seq) neuronhálók, ezek betanításához ugyanis párosított felvételek kellenek. Azonban a párosítás csak a felvételek szintjén szükséges, a vektorok illesztését már elvégzi az algoritmus. A seq2seq hálók ötlete a gépi fordítás területéről ered, de jelenleg rendkívül népszerűek például a beszédfelismerésben (Novitasari és mtsai, 2020) és a beszédszintézisben is (Wang és mtsai, 2017). A párosított mintákon tanítás előnye az lehet, hogy konkrétabban tudjuk definiálni az input-output párokat, viszont a hibalehetőséget rejtő vektorszintű illesztésre nincs szükség, azt már elvégzi a technológia. Újabban az LSTM alapú seq2seq modelleket elkezdték leváltani az ún. transzformer alapú hálózatok (Li és mtsai, 2019), melyek ugyan tipikusan nagyobb modellek, viszont a párhuzamosíthatóságuk miatt jóval gyorsabban lehet tanítani őket és teljesítményben is felülmúlják az elődjüket. Az utóbbi időben ezeket a módszereket elkezdték alkalmazni a hangkonverzióban (Zhang és mtsai, 2019; Huang és mtsai, 2019; Tanaka és mtsai, 2019; Jia és mtsai, 2019), sőt beszédjavításra is (Huang és mtsai, 2021b; Biadsy és mtsai, 2019). Saját kísérleteket ezzel a módszercsaláddal még nem végeztünk, ez a jövőbeni terveink közt szerepel.

3.6. Tapasztalatok

A három kipróbált algoritmus közül tapasztalataink szerint a párosított mintákkal dolgozó DiscoGAN igényli a legkevesebb tanítóadatot, viszonylag könnyen és gyorsan tanítható. Ennek oka, hogy a vektorszintű illesztés miatt minden egyes spektrális vektorhoz jól definiált célvektor tartozik. Ugyanez a tulajdonság eredményezi azonban a módszer gyenge pontját is: azt tapasztaltuk, hogy a dinamikus idővetemítésen alapuló illesztés maga is egy hibaforrás, mivel gyakran nem sikerül megfelelően összeillesztenie a mintákat (különösen elkent artikulációjú és/vagy akadozó beszéd esetén). Konkrétan, sok esetben a vetemítés melléktermékeként fellépő minőségromlást nagyobb fokúnak éreztük, mint a dysarthria által okozottat. Összességében tehát ezt a módszer több szempontból is limitáltnak találtuk, a kimenetként kapott hangmintákat gépiesnek, nem természetes hangzásúnak érzékeltük. A négyzetes hibafüggvényt a GAN-hibával kiegészítve sikerült ugyan javulást elérni, de így sem értük el az általunk vágyott minőséget.

A tisztán GAN-alapú, ezért párosított tanítómintákat nem igénylő módszerek elvileg teljesen természetes hangzású beszéd előállítására is képesek, a gyakorlati

tapasztalataink azonban messze nem voltak ilyen jók. Az első kipróbált módszer szerzői azt állítják, hogy megoldásuk a nyelvi tartalmat és a dysarthriás beszélő hangszínét megtartja. Ezzel szemben nekünk az volt a benyomásunk, hogy ez már a szerzők által demonstrációként feltöltött hangmintákra sem feltételül teljesül, és az angol adatbázison való újratanítás után kapott eredményeinkben sem ezt tapasztaltuk. A másik, MaskCycleGAN-VC nevű algoritmussal meggyőzőbb eredményeket kaptunk, hosszas paraméterhangolás után sikerült minőségi javulást elérnünk a dysarthriás felvételeken. Enyhén leromlott beszéd esetén, azaz kisebb korrekcióra tehát alkalmas tűnt ez a módszer, nagyobb léptékű transzformáció esetén azonban nehézségekbe ütköztünk. Nagyobb mértékű változtatás megengedésekor ugyanis az algoritmus olyan tulajdonságokat is elkezd megváltoztatni, amelyeket mi nem szeretnénk: a célbeszélő hangmagasságát is átveszi, rosszabb esetben pedig a nyelvi tartalmat is megváltoztatja - gyakorlatilag véletlenszerű halandzsát generál. Úgy véljük, hogy ezek a nehézségek a jelenlegi GAN-alapú módszertan alapvető gyengeségére tapintanak rá: nem tudjuk pontosan definiálni (és a modell tudomására hozni), hogy az inputnak milyen fokú módosítását engedjük meg, és az a módosítás pontosan mit is módosítson (és mit nem). A párosított módszer esetén vannak ugyan célvektorok, de az olyan egyszerű függvények mint a spektrogram pixeleinek négyzetes eltérése nem képesek szétválasztani a különböző érzékelési dimenziók (hangszín, hangmagasság, artikulációs tisztaság, stb.) mentén fennálló eltéréseket. A párosítatlan módszerek GAN-jainak pedig konkrét célértékek híján még kevésbé tudjuk átadni, hogy mit is várunk tőlük. Ezért mindenképpen szükségesnek látjuk a szokványos GAN-célfüggvények kiegészítését különféle ügyes megszorításokkal, és az irodalomban látunk is ilyen irányú próbálkozásokat (Tanaka és mtsai, 2019) . Úgy gondoljuk továbbá, hogy a cél pontosabb megfogalmazásában az is segítene, ha a szimpla mel-sektrogram reprezentáció helyett olyan vokódereket használnák, amelyek a beszéd részletesebb elemzését adják – elvégzik például a korábban sztendernek számító gerjesztőjel-burkológörbe szétválasztást (Ferro és mtsai, 2020).

Akár a párosított, akár a párosítatlan módszercsaládot nézzük, az időbeli transzformációra, azaz a kiindulási felvétel ritmusának és sebességének módosítására egyik sem képes, és nem is explicit céljuk. A tökéletes megoldáshoz azonban a szupraszegmentális, prozódiai jegyek átvitele is szükséges lenne (a hangmagasság mellett a hangsúly és az időzítési sajátosságok, a beszéd ritmusa is ide tartoznak). A megoldást a párosított mondatokkal dolgozó seq2seq módszerektől remélhetjük, mivel ezek a bemenetet egy időbeli sorozatként kezelik, és akár időbeli transzformációkat is képesek végezni. A hátrányuk az, hogy rengeteg (több napnyi) adaton kell őket tanítani és nagy méretű, bonyolult (nehezen implementálható) rendszerek. Ezért ezeket a modelleket tipikusan előtanítják beszédfelismerési vagy TTS (text-to-speech) feladaton, mert ilyen célra jóval nagyobb mennyiségű adat érhető el, majd ezután tanítják tovább hangkonverzióra (Huang és mtsai, 2021a).

3.7. A hangkonverzió objektív kiértékelése

Mint láthattuk, a hangkonverzió céljának formalizálása sem könnyű, épp ezért az eredményként kapott beszédjel kiértékelése sem triviális. A standard kiértékelési szempont, hogy a kapott beszédjel mennyire hasonlít a célbeszélő beszédére, esetünkben azonban ennél fontosabb, hogy javult-e a minősége, azaz mennyire tűnik természetes, valódi beszédnek, rosszabb kiindulási felvétel esetén pedig, hogy egyáltalán a beszéd érthetőségén sikerült-e javítani. Egyik szempontra sem könnyű objektív mérőszámot adni, a legmegbízhatóbb kiértékelési mód ezért még mindig a szubjektív, lehallgatásos tesztelés. Emellett sokan próbálkoznak olyan klasszikus, objektív metrikák használatával mint az MCD (mel-cepstral distortion), PESQ (perceptual evaluation of speech quality) vagy a STOI (short-term objective intelligibility). Ezek eredményét azonban fenntartásokkal kell kezelni, mert közismerten gyengén korrelálnak az emberi minősítéssel és preferenciával (Sündermann, 2005). Mi a STOI metrikával kísérleteztünk, de azt tapasztaltuk, hogy csak tökéletesen illesztett minták összehasonlítására alkalmas, így a két párosítatlan módszer eredményét nem tudtuk vele kiértékelni. Az érthetőség mérésére felvetődik a gépi beszédfelismerők használata is (Purohit és mtsai, 2020), azonban ezzel a megoldással kapcsolatban is felmerülnek aggályok (a felismerési hiba nem csak a hangminőségtől függhet). A magyar nyelvű, hasonló jellegű kiértékelések reprodukálhatóságához pedig szükség lenne egy de facto standardként kezelhető, jó hatásfokú és publikusan elérhető magyar nyelvű felismerőre. Végezetül, történtek már kísérletek a konverzió minőségének neuronháló általi megbecslésére az emberi pontozás alapján, speciálisan hangkonverziós feladatra (Lo és mtsai, 2019). Ez az irány ígéretesen hangzott, ezért megpróbálkoztunk a fenti cikkben ismertetett, betanított neuronhálóval is. Kiderült azonban, hogy normál beszédre a hálózat gyakran rosszabb pontszámokat ad, mint dysarthriás mintákra(!), mivel csakis hangkonverzión átesett mintákon tanították be. Azt kell mondanunk tehát, hogy semmiképp beszélhetünk kiforrott technológiáról, amely alkalmas lenne a beszédminőség és -érthetőség objektív mérésére.

4. Összegzés

A cikkben összefoglaltuk első tapasztalatainkat a hangkonverziós algoritmusok alkalmazhatóságára nézve a dysarthriás beszéd minőségjavításában. Ezek az első tapasztalatok nem túl kedvezőek: úgy látjuk, hogy a jelenlegi, főként GAN neuronhálókra épülő technológia nagyon érzékeny a paraméterbeállításokra, valamint nem teszi lehetővé annak finomhangolását, hogy pontosan milyen mértékű és jellegű transzformáció történjen. Úgy gondoljuk, hogy mind az alkalmazott reprezentáción (vokóderen), mind a tanítási megszorítások megfogalmazásán változtatni, bővíteni kell a továbblépéshez. Fontos lenne továbbá megbízható objektív módszereket találni a beszédminőség javulásának kiértékelésére.

Köszönetnyilvánítás

Az SZTE Informatikai Intézetének munkáját az Innovációs és Technológiai Minisztérium és a NKFIH támogatta a Mesterséges Intelligencia Nemzeti Laboratórium keretében. A kutatást az EFOP-3.6.1-16-2016-00008 Intelligens élettudományi technológiák, módszertanok, alkalmazások fejlesztése és innovatív folyamatok, szolgáltatások kialakítása a szegedi tudásbázisra építve c. pályázatának "Nyelvhasználati és jelhasználati jellegzetességek stroke-on átesett személyeknél, különös tekintettel a diagnosztikus és rehabilitációs lehetőségekre" c. pilot programja, valamint a 2019-1.2.1-EGYETEMI-ÖKO-2019-00018 pályázat keretében megvalósuló Proof of Concept Alap Pályázat „Dysarthriával élők támogatása beszédminőségük eszközös javítása által" című projektje támogatta. A TKP2021-NVA-09 számú projekt az Innovációs és Technológiai Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a TKP2021-NVA pályázati program finanszírozásában valósult meg.

Hivatkozások

Aronson, A.: Motor speech signs of neurologic disease. In: Darby, J.K. (szerk.) Speech Evaluation in Medicine, pp. 159–180. Grune and Stratton (1981)

Biadsy, F., Weiss, R.J., Moreno, P.J., Kanevsky, D., Jia, Y.: Parrotron: An endfsta-to-end speech-to-speech conversion model and its applications to hearing-impaired speech and speech separation. arXiv preprint arXiv:1904.04169 (2019)

Camillo, L., Ortiz, K.: Vocal analysis (auditory-perceptual and acoustic) in dysarthrias. Pro-Fono Revista de Atualizacao Cientifica 19(4), 381–386 (2007)

Chen, L.W., Lee, H.Y., Tsao, Y.: Generative adversarial networks for unpaired voice transformation on impaired speech. arXiv preprint arXiv:1810.12656 (2018)

Ferro, R., Onin, N., Roebel, A.: CycleGAN voice conversion of spectral envelopes using adversarial weights. In: EUSIPCO. pp. 406–410 (2020)

Horváth, S., Hirshberg, J.: Diszartria/diszartrofónia. In: Hirschberg J., Hacki T., M.K. (szerk.) Foniátria és társtudományok II., pp. 80–86. Eötvös Kiadó (2013)

Huang, W.C., Hayashi, T., Wu, Y.C., Kameoka, H., Toda, T.: Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining. arXiv preprint arXiv:1912.06813 (2019)

Huang, W.C., Hayashi, T., Wu, Y.C., Kameoka, H., Toda, T.: Pretraining techniques for sequence-to-sequence voice conversion. IEEE/ACM Transactions on Audio, Speech and Language Processing pp. 745–755 (2021a)

Huang, W.C., Kobayashi, K., Peng, Y.H., Liu, C.F., és mtsai: A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion. arXiv preprint arXiv:2106.01415 (2021b)

Jia, Y., Weiss, R.J., Biadsy, F., Macherey, W., Johnson, M., Chen, Z., Wu, Y.: Direct speech-to-speech translation with a sequence-to-sequence model. arXiv preprint arXiv:1904.06037 (2019)

Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N.: StarGAN-VC2: Rethinking conditional methods for stargan-based voice conversion. arXiv preprint arXiv:1907.12279 (2019)

Kaneko, T., Kameoka, H., Tanaka, K., Hojo, N.: MaskCycleGAN-VC: Learning non-parallel voice conversion with filling in frames. In: Proc. ICASSP. pp. 5919–5923 (2021)

Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T.S., Watkin, K., Frame, S.: Dysarthric speech database for universal access research. In: Ninth Annual Conference of the International Speech Communication Association (2008)

Kim, J.W., Jung, H.Y., Lee, M.: Vocoder-free end-to-end voice conversion with transformer network. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2020)

Kim, T., Cha, M., Kim, H., Lee, J.K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: International Conference on Machine Learning. pp. 1857–1865. PMLR (2017)

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W.Z., Sotelo, J., de Brébisson, A., Bengio, Y., Courville, A.: Melgan: Generative adversarial networks for conditional waveform synthesis. arXiv preprint arXiv:1910.06711 (2019)

Li, N., Liu, S., Liu, Y., Zhao, S., Liu, M.: Neural speech synthesis with transformer network. In: Proc. AAAI. pp. 6706–6713 (2019)

Lo, C.C., Fu, S.W., Huang, W.C., Wang, X., Yamagishi, J., Tsao, Y., Wang, H.M.: Mosnet: Deep learning based objective assessment for voice conversion. arXiv preprint arXiv:1904.08352 (2019)

Luong, M., Tran, V.A.: Flowvocoder: A small footprint neural vocoder based normalizing flow for speech synthesis. arXiv preprint arXiv:2109.13675 (2021)

Markó, A., Gráczi, T., K., B.S.: A diszfónia terápiájának hatékonysága a beteg beszédtechnikai képzettségének függvényében. Alkalmazott nyelvtudomány 12(1–2), 83–103 (2007)

Mohammadi, S., Kain, A.: An overview of voice conversion systems. Speech Communication 88, 65–82 (2017)

Morise, M., Yokomori, F., Ozawa, K.: WORLD: a vocoder-based high-quality speech synthesis system for real-time applications. IEICE Trans. on Information and Systems 99(7), 1877–1884 (2016)

Moulines, E., Sagisaka, Y.e.: Voice conversion: state of the art and perspectives. Speech Communication (speciális különszám) 16(2) (1995)

Novitasari, S., Tjandra, A., Sakti, S., Nakamura, S.: Sequence-to-sequence learning via attention transfer for incremental speech recognition. arXiv preprint arXiv:2011.02127 (2020)

Prenger, R., Valle, R., Catanzaro, B.: Waveglow: A flow-based generative network for speech synthesis. In: Proc. ICASSP. pp. 3617–3621 (2019)

Purohit, M., Patel, M., Malaviya, H., Patil, A., Parmar, M., Shah, N., Doshi, S., Patil, H.A.: Intelligibility improvement of dysarthric speech using mmse discogan. In: Proc. SPCOM. pp. 1–5. IEEE (2020)

Sündermann, D.: Voice conversion: State-of-the-art and future work. In: Forschritte der Akustik. pp. 735–736 (2005)

Tanaka, K., Kameoka, H., Kaneko, T., Hojo, N.: AttS2S-VC: Sequence-to-sequence voice conversion with attention and context preservation mechanisms. In: Proc. ICASSP. pp. 6805–6809 (2019)

Turk, O., Arslan, L.: Subband based voice conversion. In: Proc. ICSLP (2002)

Tóth, L., Kovács, G., Ivaskó, L., Tóth, A., Jakab, K., Vécsei, L.: Stroke-on átesett dysarthriás betegek beszédének gépi elemzése - kezdeti eredmények. In: Orvosi Informatika. A XXXI. Neumann Kollokvium kiadványa. pp. 43–49 (2018)

Wang, Y., Skerry-Ryan, R., Stanton, D., Wu, Y., Weiss, R.J., és mtsai: Tacotron: Towards end-to-end speech synthesis. arXiv preprint arXiv:1703.10135 (2017)

Yang, S.H., Chung, M.: Improving dysarthric speech intelligibility using cycle-consistent adversarial training. arXiv preprint arXiv:2001.04260 (2020)

Zhang, J.X., Ling, Z.H., Liu, L.J., Jiang, Y., Dai, L.R.: Sequence-to-sequence acoustic modeling for voice conversion. IEEE/ACM Trans. ASLP 27(3), 631–644 (2019)

Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)