Magyar nyelvű neurális beszédszintézis vizsgálata dialógus helyzetben

Zainkó Csaba[1], Csapó Tamás Gábor[1], Bartalis Mátyás[1], Németh Géza[1], Németh Norbert[2], Szász Gábor Krisztián[2], Szviridov István[2]

[1]Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék
[2]IdomSoft Zrt. Rendvédelmi Fejlesztési Ágazat
zainko@tmit.bme.hu

Kivonat Jelen tanulmányban olyan mély neurális hálózat alapú beszédszintetizátor rendszert (DNN-TTS) mutatunk be, amely hangsorozat bemenetet vár és a beszéd hullámformáját két lépésben állítja elő, melspektrogram köztes reprezentációt használva. Részletesen bemutatjuk és összehasonlítotjuk a Tacotron2+WaveGlow és FastPitch+HiFi-GAN (tőlünk független) rendszereket és komponenseiket. A magyar nyelvű adatokon végzett saját kísérletekben három beszélővel (két női és egy férfi) generálunk szintezitált beszédmintákat. Szubjektív, MUSHRA típusú meghallgatásos tesztjeink során a tesztalanyok a DNN-TTS beszédszintetizátorral előállított mondatokat lényegesen természetesebbnek minősítették, mint a HMM-TTS alaprendszert. A szintetizált beszédminták minősége (természetessége) ugyan nem éri el a természetes beszéd szintjét, de közel áll hozzá (Tacotron2: 58%, FastPitch: 73%, természetes: 89%). Összességében a tesztelők a FastPitch rendszert preferálták a Tacotron2-vel szemben természetesség szempontjából. A ChatBot dialógusba ágyazott tesztek eredménye szerint a női beszélők preferáltak, és a DNN-TTS rendszerekkel előállított beszéd érthetőbb, természetesebb, mint a HMM-TTS alaprendszer, és tesztelők a válaszokat is relevánsabbnak és részletesebbnek érezték az alaprendszerhez képest.
Kulcsszavak: beszédtechnológia, DNN, TTS

1. Bevezetés

A gépi szövegfelolvasók (más néven szöveg-beszéd átalakítók vagy beszédszintetizátorok, Text-To-Speech, TTS) legújabb generációi gépi tanuláson alapulnak. Kb. 2015-ig az úgynevezett rejtett Markov-modelleket (HMM, Zen és mtsai, 2009; Tóth és Németh, 2009; Tóth, 2013) alkalmazó beszédszintézis volt a fő irány, melyet a legújabb rendszerekben azóta fokozatosan felváltott a mély tanuláson (deep learning) alapuló gépi szövegfelolvasás (DNN-TTS, Zen és mtsai, 2013; van den Oord és mtsai, 2016; Zainkó és mtsai, 2017). A fejlődés során a korábbi technológiák nem avulnak el, mert marad komparatív előnyük, például a HMM szintetizátorok jellemzően kisebb erőforrást (CPU, memória, betanító adatbázis) igényelnek, mint a DNN-TTS megoldások.

A jelenleg használt beszédszintetizátorok bemenete a legtöbb esetben szöveg, egyes rendszereknél pedig hangsorozat. A beszédet a bemenetből általában

két lépésben állítják elő. Az első lépésben egy köztes reprezentációt állítanak elő, amely tipikusan a széles körben elterjedt mel-spektrogram. Ezt a köztes reprezentációt utána egy másik modell alakítja át hullámformává. Vannak úgynevezett end-to-end megoldások is, amelyek ezt egyetlen lépésben valósítják meg. Ezek minimálisan jobb minőséget tudnak, de kevésbé rugalmasak. Részletes áttekintő a beszédszintézis legújabb kutatási eredményeiről Tan és mtsai (2021) összefoglaló cikkében olvasható.

Jelen tanulmányban egy olyan rendszert mutatunk be, amely hangsorozat bemenetet vár és a beszéd hullámformáját két lépésben állítja elő. A bemeneti kvázi hangsorozatot a bemutatott modellektől független szabály alapú fontetikus átíróval állítjuk elő a megszólaltatni kívánt szövegből.

2. Módszerek

2.1. Adatok

TTS célokhoz stúdióminőségű beszéd a megfelelő, jellemzően 40 dB vagy jobb jel/zaj viszony szükséges a beszédjelen, ezért a saját fejlesztésű PPBA adatbázis (Olaszy, 2013; Tóth és mtsai, 2012) férfi és női beszélőinek hanganyagával végeztük a beszédszintézis kísérleteket. A stúdióban 44 100 Hz mintavételi frekvenciával és 16 bites lineáris kvantálással rögzítettük a beszédjelet. A tanítás során – hatékonyság növelés miatt – kisebb mintavételi frekvenciával dolgozunk, ezért jelen esetben 22 050 Hz-re újramintavételeztük a beszédet.

Jelen tanulmányban fonetikailag gazdag szöveget (Olaszy, 2013) használunk, amelyre jellemző, hogy a legváltozatosabb formában tartalmazza a hangkapcsolatokat. A hangsorozatot szabály alapú fonetikus átíróval állítjuk elő. A könnyebb kezelhetőség miatt nem fonetikus ábécét használunk, hanem közelítő betűképpel jelöljük a beszédhangokat. A szöveg feldolgozásakor a rendszer feloldja a rövidítéseket, átírja a számjegyekkel megadott adatokat, például összegeket, dátumokat. Így a modell bemenetére minden esetben olyan bemenet kerül, amely már csak hangokat vagy betűsorozatokat tartalmaz, számjegyeket vagy más jeleket (pl. százalékjelet) már nem.

A tanító adatok mondatszintűek, egy elem egy mondatból áll. A felhasznált adatok esetében 10 beszélőtől, beszélőnként 2000 mondatot használunk. A 2000 mondat minden beszélő esetén azonos. A mondatokat 3 halmazra osztjuk szét a mély tanulásos eljárások tanításánál szokott módon (tanító / validációs / teszt). A modellek tanításakor a teljes adathalmaz mondatainak kb. 2%-a volt a teszthalmaz, 7,5%-a a validációs halmaz, a maradék része a tanító halmaz. Az azonos mondatok minden beszélőnél ugyanabba a halmazba kerültek, tehát ha egy mondat a validációs halmaz része volt, akkor az minden beszélő esetében a validációs halmazban volt. Keverés esetében a halmazok függetlensége nem lett volna biztosítva.

2.2. A DNN modell bemutatása

Jelen kutatás során 2-2 modellt alkalmaztunk, mindegyik hálózat mély tanulás alapú, sok réteget tartalmazó struktúra. A beszédszintézis során nem end-to-

end megoldásokat használtunk, hanem három komponens segítségével állítottuk elő a beszédet. Ennek az az előnye, hogy jobban lehet kontrollálni a folyamatokat, lehetőség van a köztes reprezentációk módosítására, ellenőrzésére, illetve az egyes modellek egyedi tanítására, optimalizálására. Az end-to-end megoldások többnyire gyorsabbak és az adott halmazon akár minimálisan jobb minőséget is képesek előállítani, használatuk végleges rendszerek esetén indokolt, kutatási, kísérleti megoldások esetében a rugalmatlanságuk miatt kevésbe kifizetődő.

A beszédszintetizátorok összeállítása során 3 komponenst használtunk:

1. Szövegfeldolgozó, betű-hang konverzió (Profivox)
2. Hang-mel spektrogram konverter (Tacotron2, FastPitch)
3. Mel spektrogram-Hullámforma konverter (WaveGlow, HiFi-GAN)

A szövegfeldolgozó és betű-hang konverziós megoldás a korábbi diád, triád hullámforma-összefűzéses megoldás moduljait használja, részletesen Olaszy és mtsai (2000) tárgyalja.

A hang-mel spektrogram konverziós modellek esetében két rendszert vizsgáltunk, mindkettőt a saját beszédadatbázissal tanítottuk be. A Tacotron2 modell egy figyelmi mechanizmust (attention mechanism) is alkalmazó seq-to-seq modell, míg a FastPitch egy gyorsabb tanítást és futtatást is biztosító transzformer alapú megoldás.

A hullámforma előállításhoz a WaveNet alapú WaveGlow modellt használtuk, illetve a GAN megoldást tartalmazó HiFi-GAN architektúrát. A WaveGlow modellhez képest a HiFi-GAN lényegesen gyorsabb, illetve rendelkezésre áll a HiFi-GAN esetében több, változó minőségű és sebességű konfiguráció, amely lehetőséget biztosít a követelményekhez jobban illeszkedő verzió kiválasztására.

Tacotron2 A Tacotron2 (Shen és mtsai, 2018) bemenete valamilyen karaktersorozat, kimenete 80 csatornás mel-spektrogram. A modell esetében a következő mel-spektrogram paramétereket alkalmaztuk:

- Mel-csatornák száma: 80
- Minimális Mel frekvencia: 0 Hz
- Maximális Mel frekvencia: 8000 Hz
- Mintavételi frekvencia: 22050 Hz
- Ablakméret: 1024 minta
- Eltolás mérete (Hop size): 256 minta

A Tacotron2 bemenetén egy karakter beágyazás található, a modellben szabadon definiálható, hogy hány különböző bemeneti karakter használható. A szövegelőfeldolgozásnak biztosítania kell, hogy az előre definiált karakterkészleten kívül más karakter ne kerülhessen a bemenetre, mert abban az esetben a modell nem futtatható, hiba keletkezik. Mivel a modellünk bemenetére már az előfeldolgozott szöveg kerül, így biztosított, hogy a nem kívánt karakterek kezelve legyenek.

Annak érdekében, hogy a kiejtés felett nagyobb kontroll álljon rendelkezésre, a bemeneti betűsorozat esetében betűképpel jelölt hangsorozatot alkalmaztunk, pl.: *"[START] sz é p i d ő l e sz [END]"*. A bemeneten megkülönböztetjük a rövid és a hosszú hangokat. A hangkódok mellett szükségesek még start és stop

karakterek, illetve szünetek és írásjelek (pl. kérdőjel). A konfigurációtól függően 90-95 különböző bemeneti karaktert használtunk a tanításnál.

A Tacotron2 módosításával lehetőség van több beszélővel együttes tanításra is, ekkor egy vektor segítségével megadható, hogy éppen melyik beszélő tartozik az adott bemeneti és kimeneti adatokhoz. A modell így több-beszélős lesz, ami a tapasztalatok alapján jobb beszédminőséget eredményez. Az adott Tacotron2 modell esetében a PPBA (Olaszy, 2013) 10 beszélőjével tanítottuk be a hálózatot. A futtatásnál a vektor megadásával választható, hogy melyik beszélő hangján szeretnénk a szintetizálást elvégezni. A 10 beszélő között található az FF1, NOI1-es hang is, ezért ezeket közvetlenül elő lehet állítani, további beszélőadaptáció nélkül. A NOI2-es hang továbbtanítás során jött létre, a betanított 10 beszélős modellt tanítottuk tovább csak a NOI2-es hang segítségével. A Tacotron2 tanítása időigényes, 3 Bi-LSTM réteget is tartalmaz, amelyekre jellemző a lassú taníthatóság. A 10 beszélős modell továbbtanításával jelentős gyorsulást értünk el, a NOI2 hang egy nagyságrenddel rövidebb idő alatt betanítható volt így, kb. 4-5 óra egy NVIDIA Titan Xp-n.

WaveGlow A WaveGlow modell (Prenger és mtsai, 2019) mel-spektrogramból állítja elő a hullámformát. A konfigurációs paraméterek megegyeznek a Tacotron2-vel. A 80 csatornás mel reprezentációból egy WaveNet alapú megoldás segítségével áll elő a beszédjel. A modell elég általános, nem feltétlenül szükséges, hogy tanításnál a célhang segítségével adaptáljunk. Egy női beszélővel betanított hálózat a hasonló hangmagasságú női hangokat is magas minőségben képes előállítani. Viszont ha lényegesen eltérő alapfrekvenciájú beszéd kell (pl. férfi hang alacsony F0-lal), már jól hallható minőségromlás áll elő. A WaveGlow számításigényes és a tanítási folyamat is lassú: az NVIDIA által publikált egybeszélős modell tanításához is 580 000 iteráció volt szükséges egy 8 GPU-s környezetben.

Rendelkezésre áll egy NVIDIA által publikált modell, amely az LJ Speech adatbázis egyetlen női beszélőjével lett betanítva [1]. Méréseink szerint ez női beszélők esetében megfelelő, de férfi beszélők esetén nem használható, mély hangú férfi hangoknál durva torzítások jelentkeznek.

Azért, hogy a férfi beszéd is megfelelő minőségben előállítható legyen, a PPBA adatbázis segítségével tanítottunk egy WaveGlow modellt. Ez a hálózat az adatbázis 5 női és 5 férfi hangjával lett tanítva, így alkalmas a férfi beszélők beszédének előállítására is. A modell 635 000 iterációig volt tanítva.

A modell futtatása is időigényes, GPU-n ad csak megfelelő sebességet.

FastPitch A FastPitch modell (Łańcucki, 2021) több lényeges elemében is eltér a Tacotron2-től. Lényeges különbség, hogy nem tartalmaz LSTM rétegeket, helyette transzformer alapú megoldást használ. Ez lényegesen gyorsabb tanítást és futtatást eredményez. További lényeges eltérés, hogy a modell elkülönítve tartalmaz az alapfrekvenciáért és a hangidőtartamokért felelős komponenseket, így

[1] `https://drive.google.com/file/d/1cjKPHbtAMh_4HTHmuIGNkbOkPBD9qwhj`

azok külön-külön történő vezérlésére is lehetőség nyílik. Ez azt jelenti, hogy ellentétben a Tacotron2-vel, a FastPitch esetében lehetőség van a sebesség, illetve az alapfrekvencia egyszerű paraméterezésére, módosítására.

A modell bemenete hasonló a Tacotron2-höz: lehet betű vagy hang alapú is. A kompatibilitás megvalósítása érdekében: a Tacotron2-höz hasonló betűképpel jelölt hangsorozat bemenetet valósítottunk meg: 90-95 bemeneti szimbólum áll rendelkezésre, a rövid és hosszú beszédhangok, néhány kiegészítő szimbólummal.

A kimenet mel-spektrogram paraméterei megegyeznek a Tacotron2-nél ismertetett paraméterekkel, így a FastPitch kimenete használható a WaveGlow modellel is.

HiFi-GAN A HiFi-GAN (Kong és mtsai, 2020) egy GAN hálózat, amely a mel-spektrogram bemenetet alakítja át hullámformává. Az általunk használt megoldás 22 050 Hz-es mintavételi frekvenciát használ, és a HiFi-GAN szerzők által használt (v1) paraméterezést használjuk (Kong és mtsai, 2020).

A HiFi-GAN kevésbé univerzális mint a WaveGlow, azaz a generálandó beszélő hangjához közelebb kell lennie a tanító hangoknak. Amennyiben eltérő hangot (akár azonos nemű, de alapfrekvenciában jobban eltérő) állítunk elő, akkor a WaveGlow-hoz képest jelentősebb minőségi degradációt tapasztalhatunk.

A HiFi-GAN tanítása több szempontból is nehéz. Egyrészt az ilyen GAN hálózatok lassan taníthatók, másfelől a hiba (loss) értéke nem teljesen korrelál a minőséggel. Megfigyelhető volt, hogy hosszabb tanítás esetén, egyre kisebb validációs hiba érték mellett is a hangminőség nem javult, hanem erősödő torzítás miatt egyre romlott.

A felhasznált modellek továbbtanítás segítségével jöttek létre, a publikált súlyokat [2] tanítottuk tovább a PPBA 10 beszélőjével. Így az egybeszélős modellt elég univerzálissá tudtuk fejleszteni ahhoz, hogy mind férfi, mind női beszélők hangjait elő tudja állítani.

2.3. Tesztkörnyezet kialakítása

A TTS-ek szubjektív értékelése során a tesztelőket a dialógus és a felhasználás környezete is befolyásolhatja. A vizsgálatunk során arra törekedtünk, hogy ezen hatásokat amennyire tudjuk közelítsünk és így a kapott szubjektív értékelések a valós felhasználásnál ténylegesen fennálló érzeteket tükrözze. A tesztmondatok előállítása egy dialógusrendszer működési logikáját és felépítését követte, a minták előállításánál a komponensek integrációja úgy valósult meg, ahogy a valós felhasználás során is megvalósult volna.

Csatoló felület A TTS szolgáltatást HTTP hívásokkal lehet elérni, a felület elrejti a mögöttes technológiát, a korábbi HMM megoldással megegyező API biztosítja a kompatibilitást. A csatolón keresztül mondatonként lehet a szöveget szintetizálni. Az eredményt a webszerver egyetlen audióban adja vissza.

[2] `http://drive.google.com/drive/folders/1-eEYTB5Av9jNqlOWGBlRoi-WH2J7bp5Y`

Szövegelőfeldolgozó modul A szöveg előfeldolgozása elengedhetetlen lépése a beszéd szintézisnek. Itt történnek azok a feldolgozási lépések, amelyek tetszőleges szövegek felolvasását lehetővé teszik, pl: számok, dátumok, összegek feloldása, speciális karakterek értelmezése, rövidítések feloldása, stb. A modul szöveg be- és kimenettel rendelkezik, alapvetően szabály vagy szótár alapú megoldás. A számok feloldása jelenleg is szabály alapon a legmegbízhatóbb, mivel a legtöbb esetben nem környezetfüggő a feloldása, egyszerű logikával is megoldható.

Betű-hang konverzió Mind a Tacotron2, mind a FastPitch képes kezelni betűszintű bemenetet, de ebben az esetben az üzemeltető számára kevesebb kontroll áll rendelkezésre, kevésbé tudja befolyásolni a kiejtést. A szavak, mondatok kiejtése a tanító adathalmazon alapul, speciális karaktersorozatok (pl. történelmi személynevek) kiejtése egyáltalán nem, vagy csak nehezen megoldható. Ennek kezelésére a DNN modulok bemenetére a betűsorozat helyett az ahhoz tartozó hangkód-sorozatot adjuk meg. A hangkód-sorozatot közelítő betűképpel kódoljuk, így nem szükséges a modelleken változtatni. A hangkód-sorozat előállítására a ProfiVox beszédszintetizátor betű-hang konverziós modulját használjuk (Olaszy és mtsai, 2000). A modul segítségével nem csak a hangkód-sorozatot határozzuk meg, hanem a hangidőtartamokat és a fonológiai hanghosszúsági kategóriát is (rövid vagy hosszú hang). A hangidőtartamok pontos megvalósítása nem történik meg, mivel a DNN modellek saját időzítési adatokat generálnak, de a bemeneten jelöljük a hangkódnál a fonológiai hanghosszúsági kategóriát, azaz hogy rövid vagy hosszú hangról van-e szó.

A hullámforma előállítása A DNN modellek előállítása történhet tisztán CPU vagy CPU+GPU-s környezetben. A modellek futtatása jelentősen gyorsabb GPU-s környezetben. A Tacotron2 és WaveGlow modellek csak CPU (azaz GPU rendelkezésre állása nélkül) esetén lassúak, a jelentős konverziós idő miatt online, valós idejű dialógus rendszerben nem használhatók. GPU-s gyorsítás esetén a valós idő elérhető, amely már megfelelő erre a célra. A FastPitch és HiFi-GAN modellek nagyságrenddel gyorsabb futást tesznek lehetővé, szerver környezetben több CPU használatával a valós idejű működés elérhető, GPU-s gyorsítással pedig több csatorna párhuzamos kiszolgálása is megvalósítható egyetlen gépen.

A hullámformát a bemutatott modelleknél két lépésben állítja elő a rendszer. Az első lépésben a mel spektrogram reprezentációt állítja elő a Tacotron2 vagy a FastPitch modell. Ezek kompatibilisek egymással, ugyanolyan paraméterezéssel készültek, a két modell kimenete felcserélhető. Ezt a mel spektrogram reprezentációt utána WaveGlow vagy HiFi-GAN modellek segítségével alakítja át a rendszer hullámformává. Az így elkészült hullámformán további feldolgozás már nem történik, az elkészült beszédjel a csatoló HTTP felületen keresztül jut el a dialógus-vezérlőhöz, majd a későbbiek során a megfelelő időben lejátszódik.

2.4. Előzetesen elvárt hangminőség

A különböző rendszerek angol nyelven már más szerzők által implementálásra kerültek (Shen és mtsai, 2018; Łańcucki, 2021; Prenger és mtsai, 2019; Su és mtsai,

2020), és különböző összehasonlító vizsgálatokat is végeztek rajtuk. Egybeszélős modell esetén a FastPitch kismértékben, míg több beszélős esetben lényegesen jobb minőséget adott, mint a Tacotron2. A különbség abból adódott, hogy több beszélős környezetben a FastPitch azonos teljesítményt nyújtott, míg a Tacotron2 gyengébbet. Sebesség tekintetében a FastPitch kb. 60-szor gyorsabb (Łańcucki, 2021). A mel-spektrogram - hullámforma konverziós modellek esetében a beszédszintetizátor elemeit kihagyva vizsgálták a minőséget. A hullámformákat átalakították mel-spektrogrammra, majd ezt a vizsgált modellel visszakonvertálták. Az eredmények szerint a WaveGlow 0.55 ponttal gyengébben végzett, mint a HiFi-GAN v1-es verziója, miközben a HiFi-GAN kb. 7-szer gyorsabban fut. A kb. 60-szor gyorsabb HiFi-GAN v3-as verzió is jobban teljesített a WaveGlow modellnél az összehasonlításban.

Az irodalomban megtalálható összehasonlítás alapján azt várjuk, hogy a HMM rendszernél a DNN megoldások jobb minőséget adnak. A két fajta DNN megoldás közül a FastPitch+HiFi-GAN páros várhatóan jobb eredményt ad a teszteken. A beszédszintetizátorok minősége általában függ a beszélő alany hangjától. Ennek vannak objektív és szubjektív okai is. Szubjektív, hogy mennyire tartják kellemesnek, barátságosnak az adott személy hangját, objektív pedig, hogy a hangnyomás-változás fizikai folyamatai mennyire illeszkednek a gépi algoritmusokhoz. Ezek alapján nem tudunk előzetes becslést adni, hogy a két női és az egy férfi hang közül melyik teljesít jobban.

3. A szintetizált beszédminták vizsgálata

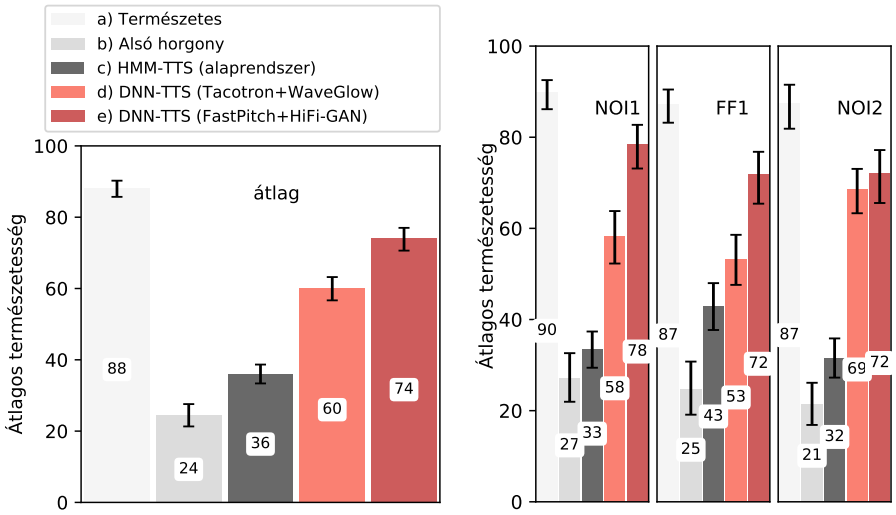
A három beszélőtől (két nő: NOI1 és NOI2 és egy férfi: FF1) mondatokat szintetizáltunk, majd ezek egy részét kiválasztottuk internetes meghallgatásos tesztekhez. A beszédszintézis kutatásának egyik célja, hogy a jövőben az emberek könnyedén használhassák az alkalmazásával készített szolgáltatásokat, és a ráépülő alkalmazásokat. Így a tesztelés folyamán a felhasználói élmény felmérése kiemelten fontos, hiszen lehet bármilyen hasznos egy alkalmazás, ha a felhasználóknak kényelmetlen, nehézséget okoz a használata, akkor nem fogják használni.

3.1. MUSHRA típusú szubjektív meghallgatásos teszt

Először úgynevezett MUSHRA-jellegű (MUltiple Stimuli with Hidden Reference and Anchor; ITU-R Recommendation BS.1534) szubjektív meghallgatásos tesztet végeztettünk el, hogy felmérjük, melyik modell hogyan teljesít, ezáltal képet kapva az eredményességükről. A MUSHRA teszt széles körben elterjedt beszédszintetizátorok minőségének összehasonlítására.

A teszt során a teszthalmazban szereplő szintetizált beszédmintákat kellett meghallgatniuk a kitöltőknek. Egy-egy oldalon mindegyik hangminta ugyanahhoz a mondathoz tartozott. Referenciaként szerepelt az eredeti (nem szintetizált) hangfájl is. A minőségi skála alsó pontjának érzékelését segítendő a tesztek mindig tartalmaztak egy rossz minőségű ún. „alsó horgony” mondatot, melyet az eredeti beszédminta torzításával állítottunk elő (MP3 kódolás, 8kbps). A kitöltők

feladata az volt, hogy minden egyes mintát osztályozzanak egy skálán aszerint, hogy mennyire hangzik természetesnek az adott beszéd (0: nagyon rossz, 100: teljesen természetes). A kitöltők nem tudták, melyik minta melyik modellhez tartozott és a minták sorrendje tesztesetenként meg is volt keverve.



1. ábra. A MUSHRA teszt eredményei: a különböző TTS modellek átlagos szubjektív értékei az egész tesztet nézve (fent) és beszélőnként (lent), a 95%-os konfidenciaintervallumokat is feltüntetve.

A MUSHRA-jellegű meghallgatásos teszt eredményét mutatja be az 1. ábra, a három beszélőre külön-külön (alul), valamint összesítve (felül). Összesen három modellt hasonlítottuk össze: az alaprendszert (HMM-TTS: Tóth, 2013; Csapó, 2013; Csapó és Németh, 2017), továbbá a DNN-TTS két változatát: Tacotron2+WaveGlow és FastPitch+HiFi-GAN, melyeket a cikk korábbi részében ismertettünk. A tesztben 10 ép hallású, magyar anyanyelvű kísérleti személy vett részt (1 nő és 9 férfi; 24–63 évesek; átlag: 38 év), 4 rendelkezett beszédtechnológiai ismeretekkel. Átlagosan 18 percig tartott a hangminták meghallgatása. A „természetes” minták érték el a legjobb eredményt, míg az „alsó horgony” a legrosszabbat – ezeknek a referenciáknak az volt a célja, hogy a tesztelők tudják mihez viszonyítani a szintetizált mondatokat. A tesztből kiderült, hogy a HMM-TTS alaprendszer viszonylag gyenge eredményt tudott elérni, átlagosan 36%-ra értékelték a tesztelők. A kutatás aktuális szakaszában javasolt DNN-TTS mind a Tacotron2+WaveGlow, mind a FastPitch+HiFi-GAN változatban jobb eredményt tudott elérni, mint az alaprendszer (Tacotron2+WaveGlow: 60%, FastPitch+HiFi-GAN: 74%). A meghallgatásos tesztben előállt rangsort Mann-Whitney-Wilcoxon teszttel is összevetettük, 95%-os konfidenciaszintet használ-

va. Ennek alapján a DNN-TTS segítségével szintetizált mondatok (mind Tacotron2, mind FastPitch) szignifikánsan jobb minőségűnek bizonyultak a horgonyhoz és az alaprendszerhez (HMM-TTS) képest, ugyanakkor szignifikánsan kevésbé voltak természetesek az eredeti referencia mondatokhoz képest.

A tendenciák hasonlóak mindhárom beszédhang esetén: HMM-TTS (alaprendszer) < DNN-TTS (Tacotron2+WaveGlow) < DNN-TTS (FastPitch+HiFi-GAN). NOI1 és FF1 esetén a FastPitch lényegesen természetesebb, mint a Tacotron2, míg a NOI2 beszélő esetén a két DNN-TTS közti különbség kevésbé jelentős (és a statisztikai teszt szerint nem különbözik szignifikánsan).

A MUSHRA teszt konklúziójaként a %-os értékekből is látszik, hogy a korábbi HMM-TTS alaprendszerhez képest a DNN-TTS eredménye lényegesen jobb, és összességében a tesztelők a legmodernebb FastPitch rendszert preferálják.

3.2. Dialógusba ágyazott szubjektív meghallgatásos teszt

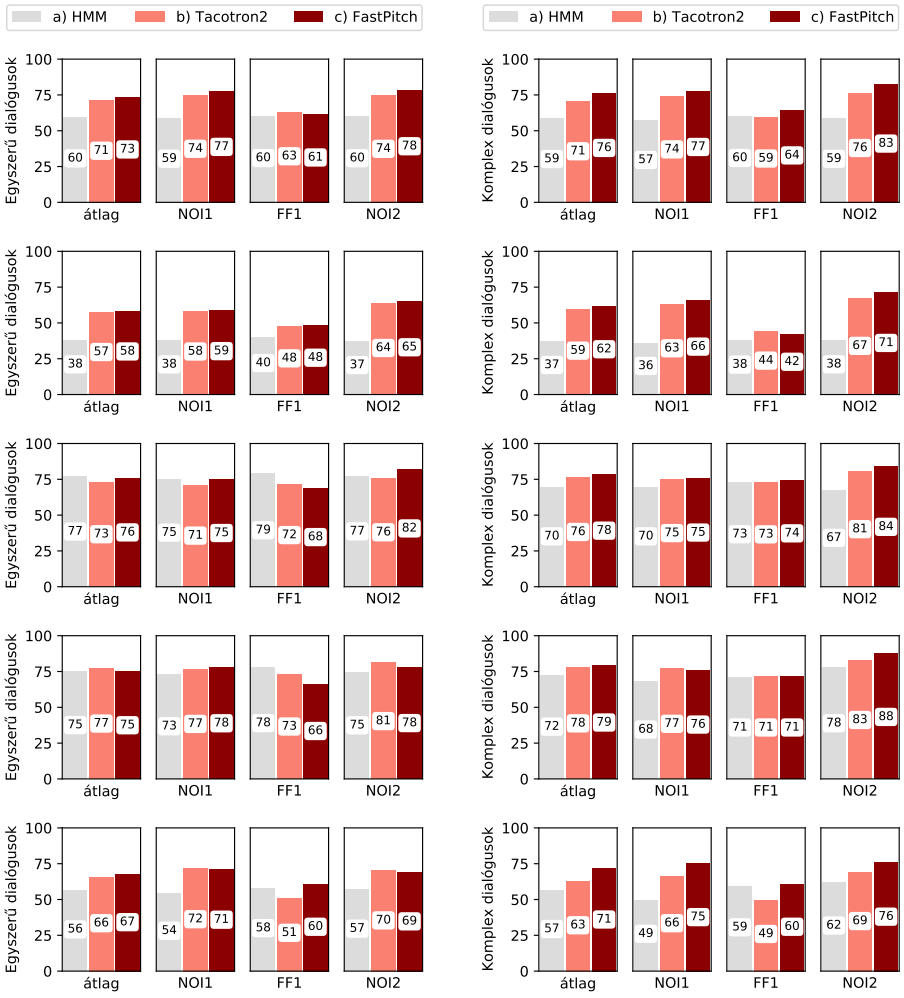
A második teszt, amelyet elvégeztünk, dialógusba ágyazva tartalmazta a szintetizált beszédmintákat. A dialógusok szöveges anyagai egy ChatBot rendszerből származtak. Most egy-egy oldalon csak egy, hosszabb hangminta szerepelt, videó formájában, mely egy elképzelt ChatBot dialógust imitált.

Összesen 126 videót generáltunk a teszthez, azaz beszélőnként (NOI1 / NOI2 / FF1) és rendszerenként (HMM-TTS / DNN-TTS-Tacotron2 / DNN-TTS-FastPitch) 14 dialógust. Ebből a 14-ből egy dialógus komplex volt, azaz több ügyfél kérdést és szintetizált mondatot tartalmazott; hosszuk kb. 1 perc volt. A többi 13 dialógus rövidebb volt, csak egy ügyfél kérdést és chatbot választ tartalmazott. Egy adott tesztelő összesen 9 dialógust értékelt egy adott beszélőtől: 3 komplexet (a 3 rendszerrel), és 6 egyszerűt (melyet a rendszer véletlenszerűen sorsolt a fenti 13-ból).

A kitöltők feladata itt az volt, hogy minden egyes mintát osztályozzanak öt kérdésre válaszolva: 1) 'Mennyire érthető a beszéd?', 2) 'Mennyire természetes a beszéd?', 3) 'Releváns-e a válasz?', 4) 'Elég részletes-e a válasz?', 5) 'Illeszkedik-e a hang a dialógushoz?' (0: negatív, 100: pozitív). A kitöltők nem tudták, melyik minta melyik modellhez tartozott és a minták sorrendje meg is volt keverve. Egy-egy kitöltő csak egy beszélő hangmintáit hallgatta (az véletlenszerűen dőlt el, hogy adott tesztelő melyik beszélőét).

A dialógusba ágyazott meghallgatásos tesztet másik tesztelői kör végezte: ebben 19 ép hallású, magyar anyanyelvű kísérleti személy vett részt (1 nő és 18 férfi; 20–63 évesek; átlag: 36 év), 6 rendelkezett beszédtechnológiai ismeretekkel. Átlagosan 9 percig tartott a dialógus videók meghallgatása és értékelése.

A dialógus teszt eredményeit az 5 kérdésre és két dialógus típusra a 2. ábra tartalmazza, melyeket kérdésenként foglalunk össze a következőkben (különkülön részábra készült minden kombinációra). Mivel egy adott tesztelő csak egy beszélőtől (NOI1 / NOI2 / FF1) hallgatott hangmintákat a teszt során, ezért a beszélők közötti értékek nem feltétlenül vethetőek össze egymással, hiszen minden tesztelő saját szubjektív skálát alkalmaz (azaz van, aki a legjobb minőséget 100%-nak állítja be, míg más csak 90%-ra értékeli a legjobb rendszert). Ugyanakkor az átlagos eredményeken érdemes a tendenciákat vizsgálni.



2. ábra. A dialógusba ágyazott meghallgatásos teszt eredményei kérdésenként és dialógus típusonként (egyszerű / komplex – bal oldali felirat szerint): átlagos eredmények (balra) és beszélőnkénti eredmények (jobbra).

Kérdés: 'Mennyire érthető a beszéd?'

Az előzetes tapasztalataink szerint mind az alaprendszerrel (HMM-TTS), mind a DNN-TTS rendszerekkel jól érthető beszédet lehet létrehozni, melyet a meghallgatásos teszt is igazolt. Az egyszerű és komplex dialógusok esetén is 55% feletti értéket kaptunk mindegyik rendszerre és beszélőre, bár a női beszélők esetén a tesztelők nagyobb különbséget éreztek a HMM-TTS alaprendszer és DNN-TTS-ek között (azaz az alaprendszert kevésbé tartották jól érthetőnek NOI1 és NOI2 esetén, míg FF1 esetén hasonló a három rendszer érthetősége).

Kérdés: 'Mennyire természetes a beszéd?'

A beszédszintetizátorok értékelésénél általában a természetesség az, amit a kutatási eredmények során vizsgálnak; illetve a korábbi MUSHRA-jellegű teszt is részben ezt vizsgálta. A HMM-TTS alaprendszer itt egyértelmű hátránnyal rendelkezik: mivel ez egyszerűbb technológia, ezért az átlagos értékelés csak 36-40% körüli. A DNN-TTS rendszerek, az általunk is elvárt és nemzetközi tapasztalatok szerint temészetesebb beszédet tudnak létrehozni; melyet a teszt igazolt: a jelen tesztben 60% körüli értéket értek el. Lényeges különbség nem látszik a Tacotron2 és a FastPitch rendszerek között sem az egyszerű, sem a komplex dialógusok esetén – bár a komplex dialógus esetén a FastPitch értékelése némileg magasabb a két női beszélő esetén. Ami egyértelműen észrevehető, hogy a férfi beszélő (FF1) esetén a DNN-TTS rendszerek természetessége kevésbé kiemelkedő a HMM-TTS alaprendszerhez képest: ugyan a minőségbeli különbség látható az ábrákon a DNN-TTS javára, de az eltérés csak 4-8% körüli.

Kérdés: 'Releváns-e a válasz?'

A tesztelők mindhárom rendszer esetén relevánsnak tartották a válaszokat. Érdekes módon a HMM-TTS alaprendszer az egyszerű dialógusok esetén kissé magasabb értékeléseket kapott, mint a DNN-TTS rendszerek (a különbség leginkább FF1 esetén látványos), de az eltérések nem jelentősek. A komplex dialógusok esetén a HMM-TTS alaprendszer alacsonyabb értékelést kapott a releváns-e kérdésre.

Kérdés: 'Elég részletes-e a válasz?'

A válasz részletessége szempontjából az egyszerű dialógusoknál nincs lényeges különbség. A komplex dialógusoknál ugyanakkor megjelenik a DNN-TTS előnye: a tesztelők NOI1 és NOI2 esetén a részletességre is magasabb értéket javasoltak, ami talán azért lehet, mert ezeknek a rendszereknek a természetessége is magasabb (lásd a 2. kérdés eredményeit a természetességgel kapcsolatban).

Kérdés: 'Illeszkedik-e a hang a dialógushoz?'

A tesztelők szerint a FastPitch rendszer illeszkedik leginkább a dialógushoz, melyet a Tacotron2 követ, majd végül a HMM-TTS alaprendszer következik – mind az egyszerű, mind a komplex dialógusok esetén; bár a különbség inkább csak a komplex esetben jelentős a FastPitch javára. A tesztelők a két női beszélő hangját tartották leginkább a dialógushoz illeszkedőnek, akit a FF1 beszélő követ. A dialógus videókon női chatbot fej szerepel, így érthető, hogy a tesztelők a férfi hangot kevésbé tartották ehhez illeszkedőnek. A két női beszélő közötti preferencia eltérés abból is fakadhat, hogy a teszt megvalósítása miatt különböző

tesztelők hallgatták a különböző beszélőktől származó hangmintákat, és az egyes tesztelők eltérő skálát alkalmaztak.

4. Összefoglalás

Jelen tanulmányban olyan DNN-TTS rendszert mutatunk be, amely hangsorozat bemenetet vár és a beszéd hullámformáját két lépésben állítja elő. Részletesen bemutattuk és összehasonlítottuk a Tacotron2+WaveGlow és FastPitch+HiFi-GAN rendszereket és komponenseiket. A magyar nyelvű adatokon végzett kísérletekben három beszélővel (két női és egy férfi) generáltunk szintezitált beszédmintákat, és és vizsgáltuk azok beszédminőségét.

Szubjektív meghallgatásos tesztjeink során a tesztalanyok a DNN-TTS beszédszintetizátorral előállított mondatokat lényegesen természetesebbnek minősítették, mint a HMM-TTS alaprendszert. A szintetizált beszédminták minősége (természetessége) ugyan nem éri el a természetes beszéd szintjét, de közel áll hozzá (Tacotron2: 58%, FastPitch: 73%, természetes: 89%). Összességében a tesztelők a FastPitch rendszert preferálták a Tacotron2-vel szemben természetesség szempontjából, és ezek az eredmények szignifikánsan különböznek.

A dialógus tesztek eredménye szerint a női beszélők preferáltak, és a DNN-TTS rendszerekkel előállított beszéd érthetőbb, természetesebb, mint a HMM-TTS alaprendszer, és tesztelők a válaszokat is relevánsabbnak és részletesebbnek érezték az alaprendszerhez képest.

Valós dialógus rendszerbe integráláskor célszerű megvizsgálni a különböző technológiákhoz szükséges hardverkövetelményeket is: a HMM-TTS CPU-n is gyorsan fut sok csatornán; a Tacotron2+WaveGlow rendszerhez GPU szükséges a valós idejű szintézishez, míg a FastPitch+HiFi-GAN modellek nagyságrenddel gyorsabb futást tesznek lehetővé (a Tacotron2+WaveGlow-hoz képest): szerver környezetben több CPU használatával a valós idejű működés elérhető, GPU-s gyorsítással pedig több csatorna párhuzamos kiszolgálása is megvalósítható.

Köszönetnyilvánítás

A kutatás az „Infokommunikáció, információtechnológiai kutatások-fejlesztések Nemzeti Laboratórium” c. projekten belül „A mesterséges intelligencia alkalmazásának kutatása” kutatási irányához kapcsolódik. Köszönjük a meghallgatásos tesztek résztvevőinek (IdomSoft Zrt. és BME TMIT kollégák) a teszt kitöltését. Köszönjük Olaszy Gábornak a cikk korábbi változatával kapcsolatos megjegyzéseket és segítséget. Csapó Tamás Gábor kutatásait az MTA Bolyai János kutatói ösztöndíja, valamint az Új Nemzeti Kiválóság Program Bolyai+ (ÚNKP-22-5-BME-316) pályázata támogatta. Zainkó Csaba kutatásait az NVIDIA Corporation GPU-val támogatta.

Hivatkozások

Csapó, T.G.: A gépi beszéd-előállítás természetességének növelése rejtett Markov-modell alapú szövegfelolvasó rendszerben [Increasing the naturalness of synthesized speech in hidden Markov-model based text-to-speech synthesis]. Phd thesis, Budapest University of Technology and Economics (2013), http://dokutar.omikk.bme.hu/collections/phd/Villamosmernoki_es_Informatikai_Kar/2014/Csapo_Tamas_Gabor/ertekezes.pdf

Csapó, T.G., Németh, G.: Folytonos paraméterű vokóder rejtett Markov-modell alapú beszédszintézisben - magyar nyelvű kísérletek 12 beszélővel. In: MSZNY 2017. pp. 308–315 (2017)

Kong, J., Kim, J., Bae, J.: HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (szerk.) Advances in Neural Information Processing Systems. vol. 33, pp. 17022–17033 (2020), https://proceedings.neurips.cc/paper/2020/file/c5d736809766d46260d816d8dbc9eb44-Paper.pdf

Łańcucki, A.: FastPitch: Parallel text-to-speech with pitch prediction. In: Proc. ICASSP. pp. 6588–6592. Toronto, ON, Canada (2021)

Olaszy, G., Németh, G., Olaszi, P., Kiss, G.: Profivox—A Hungarian text-to-speech system for telecommunications applications. International Journal of Speech Technology 3(3-4), 201–215 (2000), http://link.springer.com/article/10.1023/A:1026558915015

Olaszy, G.: Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai [Development and services of a Hungarian precisely labeled and segmented, parallel speech database] (in Hungarian). Beszédkutatás 2013 [Speech Research 2013] pp. 261–270 (2013)

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W., Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio. CoRR abs/1609.0 (2016), http://arxiv.org/abs/1609.03499

Prenger, R., Valle, R., Catanzaro, B.: Waveglow: A Flow-based Generative Network for Speech Synthesis. In: Proc. ICASSP. pp. 3617–3621. Brighton, UK (2019)

Shen, J., Pang, R., Weiss, R.J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R., Saurous, R.A., Agiomvrgiannakis, Y., Wu, Y.: Natural TTS Synthesis by Conditioning Wavenet on MEL Spectrogram Predictions. In: Proc. ICASSP. pp. 4779–4783. Calgary, Canada (2018)

Su, J., Jin, Z., Finkelstein, A.: HiFi-GAN: High-Fidelity Denoising and Dereverberation Based on Speech Deep Features in Adversarial Networks. In: Proc. Interspeech. pp. 4506–4510. International Speech Communication Association (jun 2020), https://arxiv.org/abs/2006.05694v2

Tan, X., Qin, T., Soong, F., Liu, T.Y.: A Survey on Neural Speech Synthesis (jun 2021), https://arxiv.org/abs/2106.15561v3

Tóth, B.: Rejtett Markov-modell alapú gépi beszédkeltés. Phd thesis, BME TMIT, Hungary (2013)

Tóth, B.P., Németh, G.: Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel. In: MSZNY 2009. pp. 246–256. Szeged, Hungary (2009)

Tóth, B.P., Németh, G., Olaszy, G.: Beszédkorpusz tervezése magyar nyelvű, rejtett Markov-modell alapú szövegfelolvasóhoz. Beszédkutatás 2012 [Speech Research 2012] 20, 278–295 (2012)

Zainkó, C., Tóth, B.P., Németh, G.: Magyar nyelvű WaveNet kísérletek. In: MSZNY 2017. Szeged (2017)

Zen, H., Senior, A., Schuster, M.: Statistical parametric speech synthesis using deep neural networks. In: Proc. ICASSP. pp. 7962–7966. Vancouver, Canada (2013), http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6639215

Zen, H., Tokuda, K., Black, A.W.: Statistical parametric speech synthesis. Speech Communication 51(11), 1039–1064 (nov 2009), http://linkinghub.elsevier.com/retrieve/pii/S0167639309000648