KÍSÉRLETEK A WAVENET MÓDSZER ALKALMAZÁSÁRA MAGYAR BESZÉDSZINTÉZISHEZ

Zainkó Csaba – Gyires-Tóth Bálint – Németh Géza – Olaszy Gábor

Budapesti Műszaki és Gazdaságtudományi Egyetem,
Távközlési és Médiainformatikai Tanszék

Bevezetés

A gépi beszédkeltés fejlődését a tudományos eredmények mellett alapvetően meghatározzák az aktuálisan elérhető gépi számítási és tárolási kapacitások is. A formáns alapú beszédszintézis a digitális szűrőkön és azok vezérlésén alapult, kihasználva az adott korban elérhető eszközök lehetőségeit (Klatt 1987; Olaszy 1985). A hullámforma összefűzéses eljárások a 90-es évek elején kezdtek elterjedni, amikor már lehetőség volt a szintézishez szükséges beszédhangminták időtartománybeli reprezentációjának tárolására és futás idejű feldolgozására (Olaszy et al. 2000). Később a háttértárak és memóriakapacitások növekedése lehetőséget nyitott a korpusz alapú beszédszintézisnek, amely akár több gigabájt mennyiségű előre rögzített hangfelvételből intelligens eljárással válogatja össze a szintetizáláshoz szükséges hullámforma elemeket. A korpusz alapú beszédszintetizátorokhoz (Fék et al. 2006) szükséges nagy mennyiségű felvételek lehetővé tették, hogy az addig döntően szabály alapú gépi megoldások után fejlődésnek induljanak a statisztikai elveken működő megoldások (Zen et al. 2007). A rejtett Markov-modell (*Hidden Markov Model, HMM*) alapú beszédszintetizátorok már nem hullámforma hangszeletekből építik fel a szintetizált beszédet, hanem gépi tanulás útján a beszédjel spektrális tartalmából meghatározott statisztikai paramétersorozatok segítségével beszédkódolót vezérelnek (Zen et al. 2009; Tóth et al. 2012). A fenti eljárásokról részletes információk olvashatók Németh–Olaszy (2010) összefoglaló munkájában. A WaveNet is a gépi tanulásra alapozott technológia, azonban közvetlenül a hullámformára alkalmazva. Ennek a módszernek a megvalósítását az új tudományos eredményeken felül segítette az ugrásszerűen növekvő gépi számítási kapacitás (Fan et al. 2014; Wu et al. 2015; Zen et al. 2013). A módszer azért ígéretes, mert helyettesíti a HMM-nél alkalmazott beszédkódolót, ami az átalakítás során torzítást okozhat.

Magyar nyelvű WaveNet beszédszintetizátor kutatásának megindításával azt cél tűztük ki, hogy beszédminőség tekintetében megközelítsük a termé-

Zainkó Csaba – Gyires-Tóth Bálint – Németh Géza – Olaszy Gábor 2018. Kísérletek a Wavenet módszer alkalmazására a magyar beszédszintézishez. *Beszédkutatás 2018*. 236–249.

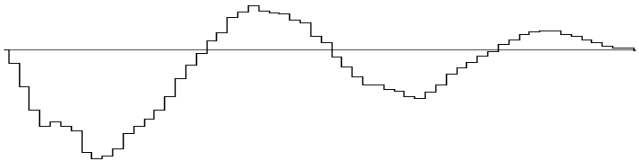
10.15775/Beszkut.2018.236-249.

szetes ejtésű beszédet. Jelen cikkünkben azt a hipotézisünket vizsgáljuk, hogy a korábbi WaveNet kísérleteinknél (Zainkó et al. 2017b) alkalmazott modellek finomításával javítani tudjuk-e a beszéd minőségét.

Módszertan

A WaveNet eljárás konvolúciós neurális hálózatok (*Convolutional Neural Network, CNN*) segítségével végzi a gépi tanulást (Le Cun–Bengio 1995; Oord et al. 2016b). A neurális hálózatok konvolúciós rétegei igen hatékonyan képesek a jellemzőket megtanulni (*feature learning*). Ez annyit jelent, hogy magukból a nyers adatokból tanulja meg a rendszer, hogy milyen absztrakció írja le legjobban azokat. Ilyen jellegű magyar nyelvű WaveNet kísérleteket 2016 novemberében kezdtük el (Zainkó et al. 2017a).

A tanulás a digitalizált beszéd hullámformájának egyes kvantált mintái alapján történik (1. ábra), tehát közvetlenül a fizikai jelből. A kimeneten pedig a szintetizálandó hullámformát reprezentáló minták sorozata jelenik meg. A WaveNet hálózat kialakítását Oord et al. (2016c) képekre alkalmazott PixelCNN neurális hálózat architektúrája és Jozefowicz et al. (2016) szövegre alkalmazott megoldása inspirálták. A WaveNet esetén azt feltételezték, hogy ha a PixelRNN hálózat képes 64x64 pixeles képeket modellezni, akkor az audiojelek finom struktúráját is lehetséges egy hasonló módszerrel kezelni.

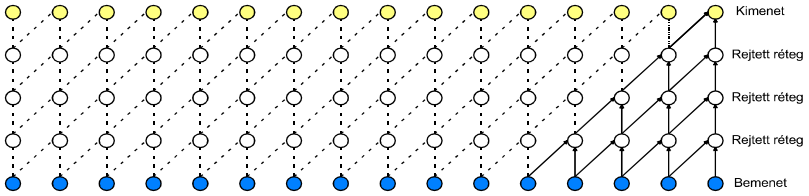


1. ábra
Egy beszédhangrészlet hullámformáját leíró mintavételi pontok az idő-amplitúdó síkon. Ezek helyét és amplitúdó számértékét használja a WaveNet a gépi tanuláshoz

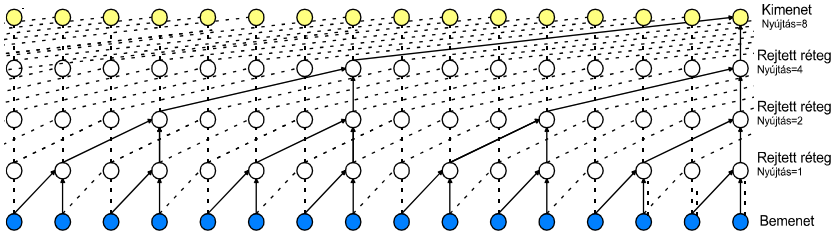
A WaveNet rendszer kialakításához a PixelCNN-nél (Oord et al. 2016b) használt modell felépítését vették alapul, ahol egy képpont generálását a korábbi képpontoktól függő feltételes valószínűségek segítségével adták meg. Az $x=\{x_1,\ldots x_T\}$ hullámformához tartozó feltételes valószínűségeket az (1) képlet adja meg. Minden x_t minta függ a korábbi időpillanatok mintáitól.

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}) \qquad (\mathbf{1})$$

Komplex adatstruktúrák esetén a konvolúciós hálózatban ahhoz, hogy nagyszámú korábbi mintát figyelembe tudjunk venni, alapvetően nagyszámú rejtett réteg vagy nagyméretű szűrők alkalmazása szükséges (2. ábra). Ezeknek viszont óriásira nőhet a számítási költsége mind a tanítás, mind a generálás során, ezért az ún. nyújtott konvolúciós (*dilated convolution*) architektúrát alkalmazták. Ennek lényege, hogy a rétegek nagyobb részénél, nem az előző időpillanat mintájához tartozó pontokat vonják be a konvolúcióba, hanem paraméterként megadható módon elemeket hagynak ki (3. ábra). A 2. és a 3. ábrán is 5 rétegű hálózatot látunk. Míg az első hálózat esetében a kimenet az azt megelőző 5 mintától függ, addig a nyújtott konvolúció esetén nagyságrendileg azonos számítás mellett, 16 mintától.



2. ábra
Példa egy 5 rétegű konvolúciós hálózati megoldásra
(Oord et al. 2016a alapján)



3. ábra
Példa egy 5 rétegű nyújtott konvolúciós hálózati megoldásra
(Oord et al. 2016a alapján)

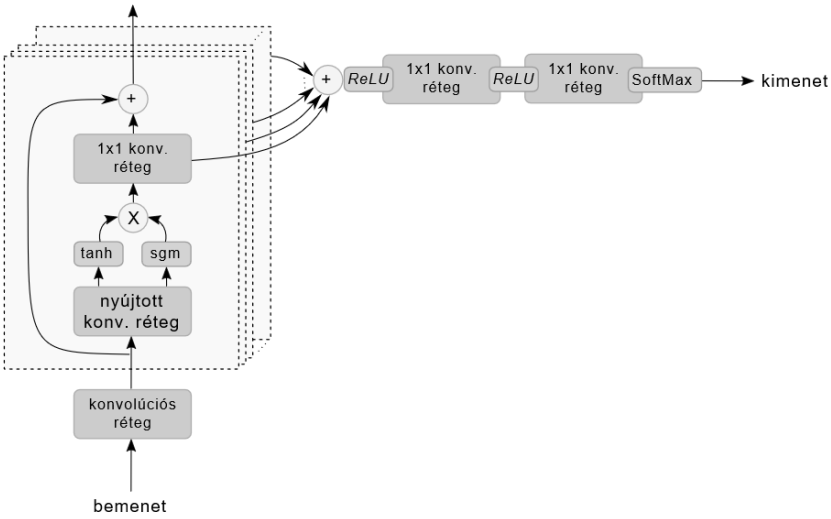
Az audiojelek előállítására tekinthetünk regressziós feladatként, de a digitális jelfeldolgozáshoz és átvitelhez széles körben használt logaritmikus kódolás segítségével osztályozási feladattá lehet átalakítani a problémát. A WaveNet esetében az ITU-T (1988) μ-law kódolását használták. A beszédfeldolgozásban tipikusan használt 16 bites lineáris PCM jelet – amely 65536 különböző kvantálási szinttel rendelkezik – átalakítják egy 256 logaritmikus kvantálási szinttel rendelkező μ-law kódolásba. Ezt a 256 szintű reprezentá-

ciót utána „one-hot” kódolással adják a hálózat bemenetére, ahol a 256 bemenet közül mindig csak egy tartalmaz nullától eltérő értéket.

A WaveNet esetében ún. kapuzott aktivációs (*gated activation*) egységeket alkalmaznak két aktivációs függvény szorzataként:

$$\mathbf{z} = \tanh\left(W_{f,k} * \mathbf{x}\right) \odot \sigma\left(W_{g,k} * \mathbf{x}\right) \tag{2}$$

A * jelöli a konvolúciót, az x a réteg bemenet, a $W_{f,k}$ a k-dik réteghez tartozó szűrő súlymátrixa, a $W_{g,k}$ pedig a kapu súlymátrixa. A σ a szigmoid függvényt jelöli, a kör pedig az elemenkénti szorzást.



4. ábra
Rétegek kapcsolata egymáshoz és a kimenethez a WaveNet esetében
(Oord et al. 2016c alapján)

A 4. ábrán látható, hogy a kimenetekhez minden rétegből kivezetjük az adatokat, és azok összegzése és 1x1-es konvolúciója után egy szoftmax függvény adja meg a kimeneti kvantált amplitúdó osztályt.

A WaveNet hálózat ebben a formában csak feltétel nélküli hullámforma generálásra alkalmas, tehát nem irányítható. Ahhoz, hogy generáláskor szabályozni tudjuk paraméterek segítségével a generálási folyamatot – amire beszédszintézisnél szükség van (pl. bemeneti szöveg, prozódiai jellemzők) – a hálózat rétegeibe vezetjük be a szükséges információkat. A magyar WaveNet kísérletek elvégzéséhez GPU (Grafikus processzor) alapú mély tanuló keretrendszert használtunk. Az adatbázisok előfeldolgozása C++ nyelven történt, a

tanítást és a generálást pedig Python alapú TensorFlow (v1.0.0) keretrendszerrel végeztük. A keretrendszer 5.1-es cudnn-t és 7.5-ös CUDA drivert használt. A tanítást GeForce GTX TITAN X-en végeztük.

A különböző tanításokhoz és kísérletekhez az alábbi beszédadatbázisokat használtuk:

FONETIKA korpusz (Olaszy 2013): 10 beszélővel felolvasott 2000 mondatos párhuzamos beszédadatbázis, amelyben a hangkapcsolatok fonetikailag kiegyenlítettek. Az adatbázis összesen 10x2000=20.000 mondatot tartalmaz.

MK_MÁV egybeszélős felolvasott beszédkorpusz, amely a MÁV állomások hangos utastájékoztató rendszeréhez optimalizált mondatokból áll (Zainkó et al. 2015). A hangfelvételek stúdióban készültek professzionális rádióbemondó közreműködésével. A korpusz 3225 annotált mondatot tartalmaz.

A hibamérés fontos eleme a WaveNet modell tanításának. A neurális hálózatok tanítása során a túltanítás elkerülésére a leállási feltételt gyakran a hibafüggvény értékének alakulásához kötjük. A hibafüggvény azt adja meg, hogy mekkora a különbség a tanító adat és a modell által aktuálisan előre jelzett (generált) minta között. Például, ha adott tanítási cikluson keresztül nem csökken a hiba (vagy elkezd növekedni) egy tanító adatoktól elkülönített, ún. validációs halmazon, akkor leállítjuk a tanítást. A validációs halmazt úgy alakítjuk ki, hogy az eredeti tanító adatok egy részét elkülönítjük, azok nem vesznek részt a tanításban. A WaveNet hálózat tanítása két szempontból is speciális. Egyrészt a generált hangmintákat megvizsgálva jellemző, hogy nem minden esetben a legkisebb hibát produkáló hálózatok adják a legjobb szubjektív értékelést a tesztelőknél. A másik tényező az, hogy a validációs halmaz elemeivel való teljes összehasonlításhoz a mondatokat le kell generálni. Ez nagyon időigényes folyamat, ezért a hatékony tanítás érdekében maradtunk a hiba egyszerű mérésénél, tehát a hálózat hibáját vizsgáltuk a validációs halmazon anélkül, hogy generálást végeztünk volna. A hiba mérésekor tehát csak mindig a következő előrejelzett minta értékét hasonlítjuk össze a valódi minta értékével. Ez gyorsabb, de kevésbé pontos.

A WaveNet eljárás során egymástól független tanítási és generálási (szintézis) fázisokat különböztetünk meg. A tanítási fázis alatt a tanító adathalmaz segítségével a modellt úgy próbáljuk meg módosítani, hogy a validációs halmazon mért hiba folyamatosan csökkenjen. A tanítás során a kísérletekkel határozzuk meg, hogy milyen méretű hálózat alkalmas jobb beszédminőség elérésére. Szintén a kísérletekkel vizsgáljuk, hogy milyen vezérlőparaméterek használata esetén tudjuk az emberihez minél közelebbi hangminőséget elérni. A generálási (szintézis) fázisban a modellt kismértékű zajjal indítjuk el, majd a vezérlő paramétereket figyelembe véve mintánként generálja a beszéd hullámformáját. A generálás esetében mindig az előzőleg generált mintákat használja a hálózat. Mivel a következő minta generálásához szükséges az

előzőleg generált minta, ezért nem tudjuk a GPU-k párhuzamos számítását kihasználni és egyszerre több mintát előállítani.

A szintézis gyorsítható, Le Paine (2016) Fast-WaveNet módszerével. A forward lépéseknél olyan számításokat végzünk el minden egyes lépésnél, amit már korábban egyszer kiszámoltunk. La Paine rámutatott, hogy a részeredmények eltárolásával a generálás $O(2^L)$-ről O(L)-re gyorsítható, ahol L a rejtett rétegek száma. A saját méréseink először nem támasztották alá ezt a gyorsulást. A generálás folyamatát elemezve megállapítottuk, hogy a Fast-WaveNet esetében a numerikus számítás mennyisége annyira lecsökken, hogy a futási idő legnagyobb részét a GPU-ra történő adatátvitel és a számítások után az eredmények memóriába való visszaolvasása adta.

Kísérletek

Az első kísérletünkben azt vizsgáltuk, hogy a különböző vezérlő paramétereket hogyan tudjuk felhasználni a modell vezérlésére, illetve milyen hatással vannak közvetlenül a hibára, közvetve pedig a generált beszéd minőségére. A tanító adatbázis összetétele és a hálózat mérete változtatásának hatását a második kísérletben vizsgáltuk. Ebben három különböző modellt készítettünk és ezekkel szintetizáltunk tesztmondatokat. Ahhoz, hogy a fejlődést nyomon tudjuk követni, korábbi (Zainkó et al. 2017b) modellek segítségével is elkészítettük ugyanazokat a mondatokat, majd meghallgatásos teszttel értékeltettük őket.

A mély rétegek szabályozása

A WaveNet hálózat nem csak a bemeneti rétegen keresztül vezérelhető, hanem minden rétegben módosíthatjuk a szűrő és a kapu súlyok hatását (Oord et al. 2016a). Az (1)-es képletet módosítva a feltételes eloszlásunk a következő formába írható át:

$$p(\mathbf{x} \mid \mathbf{h}) = \prod_{t=1}^{T} p(x_t \mid x_1, \ldots, x_{t-1}, \mathbf{h}) \tag{3}$$

A plusz bemeneteket ***h***-val jelöltük. A ***h*** lehet egy globális paraméter, amely hosszú időn keresztül állandó, például a beszélő azonosítója. Amennyiben egy *y=f(h)* függvénnyel a bemeneti mintákhoz illesztett paraméterlistát generálunk (például hangkódok vagy egyéb nyelvi jellemzők), akkor a konvolúciós egységekben lévő aktivációt - (2)-es képlet - a következőképpen módosíthatjuk (ahol $V_{f,k}$*y és $V_{g,k}$*y egy-egy 1x1-es konvolúció):

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}) \tag{4}$$

A kísérleteinkben a -2, +2 beszédhangos környezettel kibővített beszédhangot (5 hangnyi területet) kódoltuk, illetve az alapfrekvencia logaritmusát.

Ezek mellett további hangkörnyezetet leíró jellemzőket használtunk fel. Például azt, hogy a középső hang zöngés, vagy magánhangzó-e, illetve azt, hogy mennyi a hang időtartama. Ezek mellett prozódiai információkkal kapcsolatos paramétereket is felhasználunk, mint például a vizsgált hangot tartalmazó szó pozíciója a mondatban vagy azt, hogy a szó a mondat melyik prozódiai egységében található (kezdő, közbenső, utolsó, vagy csak egy ilyen egységből áll a mondat).

Több-beszélős adatbázis esetében nem használtuk a $\boldsymbol{h}$ globális paraméterezés lehetőségét, a beszélő azonosítóját is az y bemenetre konvertáltuk át, illetve többnyelvű modell esetén a beszéd nyelvi azonosítóját is hasonlóan kezeltük.

A paraméterek szerepe a generálásban

A vezérlő információk nem egyforma mértékben hatnak a kimeneti hullámforma minőségére. Az első kísérletben azt vizsgáltuk, hogy ha egy paramétert nullázunk (hatását kiiktatjuk), akkor mennyivel növekszik meg a hiba a validációs halmazon. Amennyiben a hiba alig változik, az azt jelenti, hogy kevésbé fontos a hálózat számára ez a paraméter, ha pedig az adott paraméter elhagyása jelentősen növeli a hibát, akkor az fontosabb a vezérlés szempontjából.

A modellek tanítása

A második kísérletben azt vizsgáltuk, hogy miként lehet egy sokbeszélős modellt felhasználni egyetlen beszélő jó minőségű szintetizálásához. A korábbi WaveNet kísérletek rámutattak, hogy több beszélő adatainak párhuzamos használata segíti, hogy a rendszer ne az adott beszélő egyéni tulajdonságait, hanem a több beszélő közös tulajdonságait tanulja meg. Ez azt jelenti, hogy a rendszer képes az adott nyelv jellemzőinek egy részét megtanulni. Lehetőség van arra, hogy az elkészült modellt úgy vezéreljük, hogy egy adott beszélő hangját generálja, de a beszélők egy adott csoportjának együttes beszédét is generálni tudja, tehát egy átlaghangon állíthat elő beszédet. Ez lehetőséget ad arra, hogy az egyes beszélők hibáit a rendszer figyelmen kívül hagyja, jobb minőségű, a nyelvre jobban jellemző beszédet szintetizáljon.

A modell tanítása két lépésből állt. Első lépésként sok beszélő segítségével egy átlaghangot tanítottunk. A FONETIKA adatbázis 10 beszélője segítségével tanítottuk a rendszert. A mondatok és a beszélők is véletlen sorrendben követték egymást. A tanítás után a modell képes volt mind a 10 beszélő hangján beszédet előállítani vagy azok tetszőleges kombinációjával. Például lehetőség van nő vagy férfi átlaghang megszólaltatására, vagy magyar átlaghangú beszéd generálására, amelynél a beszélő neme nem megállapítható. A személyparaméterek folyamatos változtatásával személyek közötti átmenet is megvalósítható.

A HMM beszédszintetizálási modelleknél jellemző tanítási módszer, hogy sok beszélő átlaghang modelljét adaptálják egy beszélőre. WaveNet esetében is lehetőség van erre, így a kísérleteink során a fent elkészült több-beszélős modellt tanítottuk tovább, már egyetlen hanggal úgy, hogy a továbbiakban már nem adtuk meg a modellnek, hogy melyik beszélő hangját tanulja, hanem az átlaghangot módosítja már.

A tanítási folyamat második lépésében az MK_MÁV adatbázis egyetlen beszélőjével folytattuk a tanítást, a beszélő paramétereket átlaghangnak állítottuk be. A modell ekkor már csak ennek az adott beszélőnek a hangjellegzetességeit tanulta.

A modellek tanítása során 3 különböző rendszer készítettünk (egyszerű, közepes, nagy), amelyeknél a tanító adatok és a tanítás módja, paraméterek száma megegyezik, csak a hiper-paraméterekben van különbség. Az egyszerű modell esetében a konvolúciós szűrő mélysége 32 volt, a másik kettőnél 64. A rétegek száma az egyszerű és a közepes modell esetében 40 réteg volt, a nagy esetében pedig 80 rejtett réteg szerepelt a hálózatban.

Mondatok generálása

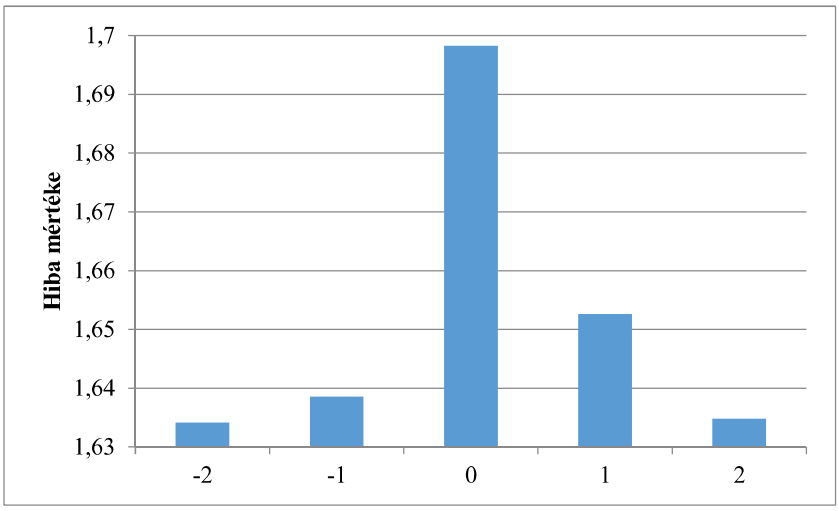
WaveNet tanítása időigényes, de a nagy mennyiségű feldolgozott adatot figyelembe véve hatékonynak tekinthető. A modell paramétereitől függően egy iterációs lépés kb. 1,4-2 mp-ig tart, amely során 100.000 mintát, 6,25 mp hanganyagot használunk fel. A tanítás és a generálás is 16kHz-es mintavételi frekvenciával történt. A mérések szerint egy minta generálása, kb. 0,25 mp-ig tart. Így nagyságrendileg 1 mp hanganyag előállítása kb. 2 óra.

A szintézis Le Paine (2016) Fast-WaveNet módszerével gyorsítva és a GPU-t kihagyva, csak CPU-n (központi processzoron) futtatva 1 mp hanganyag előállítása csak kb. 4 percet vett igénybe. A program CPU-ra történő optimalizálásával és a számítások párhuzamosításával az 1 másodperces hanganyag előállítását sikerült végül 13 mp-re leszorítani.

Eredmények

A paraméterek szerepe a generálásban

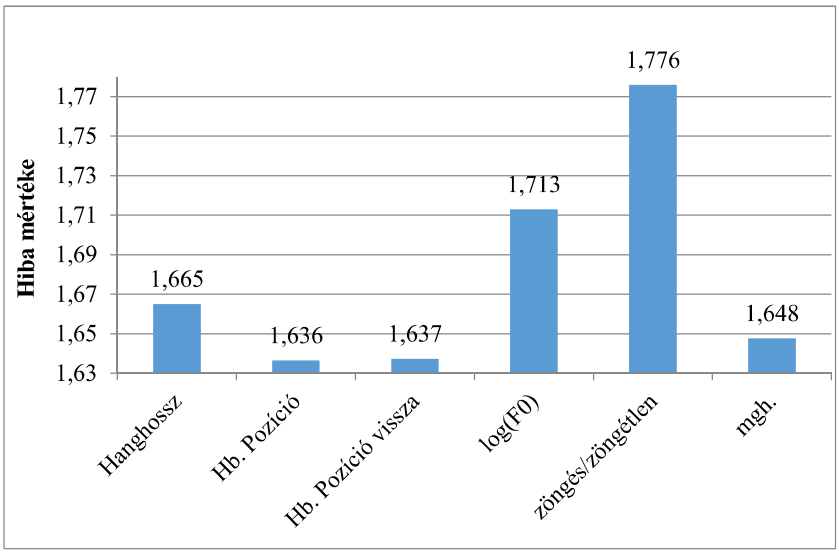
Az 5. ábrán azt láthatjuk, hogy ha az adott hang vagy annak környezetéből egy hangra vonatkozó információt törlünk, mennyire nő meg a hiba mértéke. Az ábrán a hálózat átlagos hibáját ábrázoltuk.



5. ábra
A hangok elhagyásának hatása a validációs hibára

A vártnak megfelelően a szintetizálandó hang elhagyása okozta a legnagyobb hibát. A második legfontosabbnak az adott hang után következő hang bizonyult, majd az adott hang előtt álló következett. A legkevésbé fontos a kísérlet alapján az adott hang előtt kettővel álló beszédhang.

A további 31 különböző vezérlő paraméter hatását is vizsgáltuk. A 6. ábrán látható a 6 legfontosabb paraméter hatása a hibára. A zöngésségi paraméter elhagyása okozta a legnagyobb hibát a rendszerben. A valós érték az ábrából kilógna, értéke 1,766. Második legfontosabbnak az alapfrekvencia értéke bizonyult, majd az adott hang hossza. A negyedik legfontosabb paraméter azt adja meg, hogy magánhangzó-e az adott hang. További fontos paraméter az, hogy a hangon belül a generálás hol tart, illetve mennyi van még hátra az adott beszédhang teljes megvalósításából. Az ábrán nem megjelenített paraméterek is hatással voltak kis mértékben a hibára.



6. ábra
Az első 6 legnagyobb validációs hibára ható paraméter

Mivel a WaveNet egy hullámforma generáló eljárás és magában nem alkalmas arra, hogy szövegből beszédet előállítson, ezért természetes mondatokból indulva állítottuk elő a tesztben szereplő mondatokat. Nem szövegből generáltuk a vezérlő paramétereket, hanem a természetes ejtésű mondatokból származtattuk. Ezzel az eljárással elkerülhető az egyéb modulok (fonetikai átíró, prozódiai modul) befolyása a hangminőségre, például a prozódia helytelen modellezése. A szintetizálás során a tanító adatbázisban nem szereplő mondatok vezérlő paramétereinek felhasználásával állítottuk elő a hullámformákat. Kiválasztottunk 8 különböző mondatot, majd mind a 3 új rendszerrel és a korábbi tanulmányban (Zainkó et al. 2017b) ismertetett 2 legjobb rendszerrel előállítottuk a szintetizált változatokat.

Meghallgatásos teszt

A második kísérlet eredményeinek értékeléséhez meghallgatásos MUSHRA (MUlti-Stimulus test with Hidden Reference and Anchor) tesztet készítettünk ITU-R (2001). A teszt lényege, hogy az alanyok a meghallgatásos összehasonlítás során egy referencia mondathoz hasonlítják a különböző módszerrel előállított mondatokat és megadják egy 100-as skálán, hogy mennyire találják a referenciához hasonlónak a meghallgatott mondatot. A meghallgatandó mondatok közé a referencia mondatot is elrejtettünk.

A tesztbe 2 korábbi rendszert is bevettünk, hogy a modellek fejlődését is tudjuk vizsgálni. A (Zainkó et al. 2017b) cikkben szereplő rendszerek többnyelvű modellt használnak, az egyik rendszer 2 nyelvű (magyar–angol) a másik háromnyelvű (magyar–angol–német). A tesztben csak magyar nyelvű mondatokat generáltunk. A tesztben összesen 8 mondat 5 változatát kellett értékelnie az alanyoknak. A mondatok a MK_MÁV adatbázis tematikájához illeszkedő mondatok voltak. A tesztelt rendszerek főbb tulajdonságait az 1. táblázatban adjuk meg.

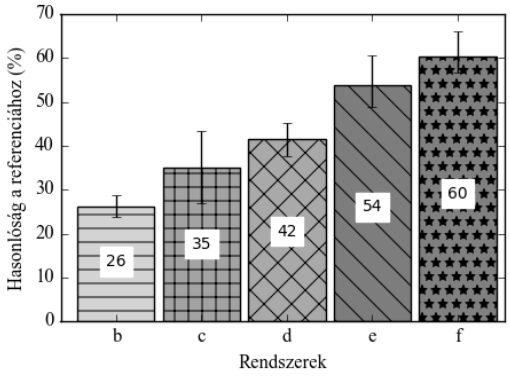
1. táblázat: A kísérletben felhasznált modellek főbb adatai

Azonosító	Rétegek száma	Szűrők mélysége	Nyelvek száma
b	40	64	2
c	40	64	3
d	40	32	1
e	40	64	1
f	80	64	1

A tesztet 27 felnőtt töltötte ki, 11 férfi és 16 nő. A tesztet átlagosan 12,6 perc alatt végezték el. A beszélők átlag életkora 50 (29–77) év volt.

Eredmények

A MUSHRA tesztben előállt rangsort Mann–Whitney–Wilcoxon teszttel is összevetettük, 95%-os konfidenciaszintet használva. Ez alapján a vizsgált rendszerek mondatai szignifikánsan különböző minőségűnek bizonyultak a horgony és a WaveNet által szintetizált mondatok mindegyikéhez képest, tehát mind az 5 rendszer szignifikánsan különbözik egymástól. A legrosszabb eredményt a Zainkó et al. (2017b) cikkben leírt korábbi rendszer érte el (b modell), tehát az új modellek sikeresebbek lettek a korábbinál. Az eredmények arányosak voltak a modell méretével, minél összetettebb, nagyobb volt a modell, annál jobb eredményt ért el. Az eredmények a 6. ábrán láthatóak (a rejtett referenciát nem jelenítettük meg). A tesztalanyok a referencia mondattal pontosan megegyező mintát 92%-osra értékelték. Mivel nem volt külön negatív referencia, ezért a teszt alapján abszolút minőségi értékelést nem tudunk adni, csak a rendszereket tudjuk egymáshoz hasonlítani.



7. ábra

Az 5 vizsgált WaveNet modell hangminősége

Összefoglalás

A magyar nyelvű WaveNet kísérleteinkkel sikerült elérni azt a beszédminőséget, amely alapján a modell felhasználható beszédszintetizálás céljára. A beszédminőséget szignifikánsan sikerült javítanunk a korábbi rendszereinkhez képest, a modell finomítása megfelelő eredményt hozott, javult a hangminőség (f modell). A modell legnagyobb hátránya jelenleg az, hogy nem tudjuk a más beszédszintézis technológiákhoz hasonló számítási sebességet megközelíteni. Ez jelentősen korlátozza az alkalmazási lehetőségeket. Ebben jelentős javulás várható a közeljövőben, mert a Google DeepMind kutatócsoportja már bejelentette, hogy angol és japán nyelvű megoldásaiban ezerszeres gyorsulást ért el (Oard 2017d).

A WaveNet vagy a várhatóan nagy számban megjelenő hasonló mély neurális hálózat alapú rendszerek egyértelműen jobb hangminőségű beszédet tudnak előállítani, mint a korábbi technológiák és a sebesség problémák megoldása után széles körű elterjedésük várható. Mivel ezek a technológiák már közelítik az emberi beszéd minőségét, és egyes esetekben már nem megkülönböztethető az emberi beszédtől a gépi változat, alkalmazásuknál fokozottan jogi és etikai problémák kerülhetnek előtérbe. Elképzelhető, hogy új kutatási irány lesz az, hogy a szintetizált beszéd teljesen emberi legyen, de ne hasonlítson egyetlen tanító adatbázisban szereplő beszélő hangjára sem (nyelvi átlag hang).

Irodalom

Fan, Yuchen – Qian, Yao – Xie, Fenglong – Soong, Frank K. 2014. TTS synthesis with bidirectional LSTM based recurrent neural networks. *Interspeech*. 1964–1968.

Fék, Márk – Pesti, Péter – Németh, Géza – Zainkó, Csaba – Olaszy, Gábor 2006. Corpus-based unit selection TTS for Hungarian. In International Conference on Text, Speech and Dialogue 367–373. Springer

ITU-R. Recommendation BS.1534: Method for the subjective assessment of intermediate audio quality (2001)

ITU-T. Recommendation G. 711. Pulse Code Modulation (PCM) of voice frequencies 1988.

Jozefowicz, Rafal – Vinyals, Oriol – Schuster, Mike – Shazeer, Noam – Wu, Yonghui 2016. Exploring the limits of language modeling 2016 (Feb 7) arXiv preprint arXiv:1602.02410.

Klatt, Dennis 1987. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America* 82/3. 737–793.

Le Cun, Yann – Bengio, Yoshua 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361/10.

Nagy, Péter – Zainkó, Csaba – Németh, Géza 2015. Synthesis of speaking styles with corpus-and HMM-based approaches. Cognitive Infocommunications (CogInfoCom), 6th IEEE International Conference on. IEEE.

Németh, Géza – Olaszy Gábor (szerk.) 2010. A magyar beszéd. Akadémiai Kiadó, Budapest.

Olaszy Gábor 1985. A magyar beszéd leggyakoribb hangsorépítő elemeinek szerkezete és szintézise. A számítógépes beszéd-előállítás néhány kérdése. *Nyelvtudományi Értekezések* 121. Akadémiai Kiadó, Budapest.

Olaszy, Gábor 2013. Precíziós, párhuzamos magyar beszédadatbázis fejlesztése és szolgáltatásai. *Beszédkutatás 2013*. 261–270.

Olaszy, Gábor – Németh, Géza – Olaszi, Péter – Kiss, Géza – Gordos, Géza 2000. "PROFIVOX - A Hungarian Professional TTS System for Telecommunications Applications", *International Journal of Speech Technology* 3. 3-4. 201–216.

Oord, Aaron van den – Dieleman, Sander – Zen, Heiga – Simonyan, Karen – Vinyals, Oriol – Graves, Alex – Nal, Kalchbrenner – Senior, Andrew – Kavukcuoglu, Koray 2016a. WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Oord, Aaron van den – Nal, Kalchbrenner – Kavukcuoglu, Koray 2016b. Pixel Recurrent Neural Networks. arXiv preprint arXiv:1601.06759.

Oord, Aaron van den – Nal, Kalchbrenner – Vinyals, Oriol – Espeholt, Lasse – Graves, Alex – Kavukcuoglu, Koray 2016c. Conditional image generation with pixelcnn decoders. arXiv preprint arXiv:1606.05328.

Oord, Aaron van den – Li, Yazhe – Babuschkin, Igor – Simonyan, Karen – Vinyals, Oriol – Kavukcuoglu, Koray – van den Driessche, George – Lockhart, Edward – Cobo, Luis C. – Stimberg, Florian – Casagrande, Norman – Grewe, Dominik – Noury, Seb – Dieleman, Sander – Elsen, Erich – Kalchbrenner, Nal – Zen, Heiga – Graves, Alex – King, Helen – Walters, Tom – Belov, Dan – Hassabis, Demis 2017d. Parallel WaveNet: Fast High-Fidelity Speech Synthesis arXiv preprint arXiv: 1711.10433

Le Paine, Tom 2016 (nov.10). Fast Wavenet: An efficient Wavenet generation implementation https://github.com/tomlepaine/fast-wavenet

Tóth, Bálint, Pál – Németh, Géza 2009. Rejtett Markov-modell alapú szövegfelolvasó adaptációja félig spontán magyar beszéddel. VI. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 246–256.

Wu, Zhizheng – Valentini-Botinhao, Cassia – Watts, Oliver – King, Simon 2015. Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 4460–4464.

Zainkó, Csaba – Bartalis, Mátyás – Németh, Géza – Olaszy, Gábor 2015. A Polyglot Domain Optimised Text-To-Speech System for Railway Station Announcements. Sixteenth Annual Conference of the International Speech Communication Association.

Zainkó, Csaba – Tóth, Bálint Pál – Németh, Géza 2017a. Polyglot Magyar nyelvű WaveNet kísérletek, In XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017), Szeged. 205–216.

Zainkó, Csaba – Tóth, Bálint Pál – Németh, Géza – Olaszy, Gábor 2017b. Polyglot speech generation with WaveNet, In Technical report of Smartlab, http://smartlab.tmit.bme.hu/downloads/pdf/zainko/Zainko_Polyglot_2017.pdf Letöltés: 2017. November 15.

Zen, Heiga – Senior, Andrew – Schuster, Mike 2013. Statistical parametric speech synthesis using deep neural networks. In IEEE International Conference on Acoustics, Speech and Signal Processing. 7962–7966.

Zen, Heiga – Toda, Tomoki – Nakamura, Masaru – Tokuda, Keiichi 2007. Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. IEICE transactions on information and systems 90/1. 325–333.

Zen, Heiga – Tokuda, Keiichi – Black, Alan 2009. Statistical parametric speech synthesis. *Speech Communication*, 51/11. 1039–1064.

Köszönetnyilvánítás

A szerzők köszönetüket fejezik az NVIDIA Corporation-nak, hogy a kutatást egy NVIDIA TITAN X GPU-val támogatták. A kutatást támogatta továbbá a DANSPLAT (EUREKA 9 944) és a VUK (*AAL*-2014-1-183.) projekt.

Experiments to apply the WaveNet method for Hungarian speech synthesis

The WaveNet architecture is suitable to generate high quality speech, it was demonstrated for English by Google DeepMind. In this paper we describe our experiments of using WaveNet for Hungarian speech generation. We investigated the effects of different control parameters and compared the quality of generated speech with different hyper-parameter settings and with different Hungarian speech databases. We examined the most influential control parameters and we conducted a listening test to investigate the evaluation of Hungarian WaveNet models. Significant increase in quality of synthesized speech with larger models and with a modified approach was achieved.