Artikulációs beszédszintézis megvalósítása dinamikus ultrahangfelvételek alapján

Trencsényi Réka[1], Czap László[2]

[1] *Debreceni Egyetem, Villamosmérnöki Tanszék*
[2] *Miskolci Egyetem, Automatizálási és Infokommunikációs Intézet*

Abstract

Starting from 2D dynamic ultrasound sources recording the movement of the vocal organs and the speech signal of the speaker in a simultaneous and synchronised manner, we produce machine speech by means of artificial intelligence. As visual objects, we use tongue and palate contours fitted automatically to the anatomic boundaries of the ultrasound images, and for training, we extract geometric information from these contours, as the change of their shape fundamentally describes the movement of the vocal organs during articulation. The geometric data consist of radial distances between the tongue and palate contours and coefficients of the discrete cosine transform of the curves, respectively. Relying on this dataset, parameters connected to the acoustic content of the speech signal are trained by the network. These parameters can be interpreted in the framework of the acoustic tube model of the vocal tract, and according to this, reflection coefficients and areas of the articulation channel are to be trained. In this study, sentences are synthesised using linear predictive coding and the acoustic tube model.

1. Bevezetés

A beszédkutatás egyik legfontosabb tématerülete a beszédszintézis, ami elemi alkotóját képezheti az ember-gép kapcsolatnak. Ez esetben a gép kommunikációs szerepe abban nyilvánul meg, hogy kódoló adóvá válik, azaz beszédet produkál. Napjainkban a beszédszintézis legelterjedtebb irányzata a szövegfelolvasók készítése, melyek írott szövegeket szólaltatnak meg. A beszédszintetizátorok megalkotásának célja a természetes emberi beszéd közben kialakuló akusztikai produktum élethű utánzása. Ebben a megközelítésben a beszéd hullámformája adja a kiindulópontot, amit kétfajta megoldásban alkalmaznak gépi beszéd előállítására. Az egyik csoportba az úgynevezett forráskódolású

Email addresses: trencsenyi.reka@science.unideb.hu (Trencsényi Réka), czap@uni-miskolc.hu (Czap László)

Beszédtudomány – Speech Science 2025/1. 90–116. DOI: 10.15775/Besztud.2025.1.90-116

technikák tartoznak, melyek segítségével a beszédjelből kivonják a lényegi információkat és ezeket bemeneti adatsorozatként kezelik a szintézis során (Kaur & Singh, 2022). A másik megoldás az emberi hangot közvetlenül használja fel a beszédépítéshez olyan módon, hogy a beszédjelből különböző hosszúságú hullámforma-részleteket vágnak ki és tárolnak el, majd az így kapott elemek megfelelő kiválasztásával és összefűzésével megkonstruálják a kívánt beszédhullámot (Panda & Nayak, 2017). Ezeken túlmenően, tágabb módszertani szempontok alapján megkülönböztetünk még szabályalapú, illetve statisztikai elven működő beszéd-előállítási eljárásokat. Az előbbi esetében megfigyelések és tapasztalatok szerint felállított szabályokkal koordinálják a szintézis egyes lépéseit (Carlson & Granström, 2008), az utóbbi esetében pedig valószínűségeken alapuló belső rendszerállapotok révén jutnak el a beszédprodukcióhoz. A statisztikai elvű módszerek egyik tipikus válfaja a gépi tanulóalgoritmusok szerkesztése és alkalmazása, ami a jelenlegi tudományos kutatások egyik legaktívabban prosperáló irányzataként tartható számon (Mahum et al., 2023). A magyarországi viszonyokat tekintve, a hazai kutatók az 1970-es években kapcsolódtak be a témakör tanulmányozásába. Az első magyar szövegfelolvasó rendszer, a HungaroVox 1980-ban készült el, majd folyamatos fejlesztések nyomán sorban követték egymást a ScriptoVox (Olaszy & Gordos, 1987), a Brailab (Kiss et al., 1987), a PC TALKER (Király, 1989), a PCROBOT, a MultiVox (Olaszy, 1989), a ProfiVox (Olaszy et al., 2000), illetve a FlexVoice (Balogh et al., 2000) rendszerek, melyek szépen kirajzolják a fejlődés ívét a nagyon robotos hangzású beszédtől a teljesen emberi hangzású, jól érthető beszédig.

A szövegfelolvasó rendszerek a beszédszintézis klasszikus ágát képviselik, ahol hosszú évtizedek során rendkívül sok tapasztalat és gazdag tudásanyag halmozódott fel, amit a szakirodalom számos közleménye is igazol (Arik et al., 2017; Mullah, 2015; Shiga et al., 2020). Emellett azonban olyan területek is kezdenek egyre élénkebben előtérbe kerülni, melyek kevésbé kidolgozottak, és rengeteg nyitott probléma vár még megoldásra. Ide sorolható például az artikulációs beszédszintézis (Csapó et al., 2017; Tóth et al., 2018; Juanpere & Csapó, 2019; Csapó, 2020; Csapó et al., 2020; Arthur & Csapó, 2021; Csapó

et al., 2022; Denby et al., 2023), ami az akusztikai produktum utánzását emberi hangminták helyett a hangképzés és artikuláció gépi leképezése révén próbálja megvalósítani. Ennek egyik technológiai vonulata a robotok beszédének előállításához szükséges artikulációs elektromechanikus beszédkeltőkre irányuló kísérletezés (Ashok et al., 2022; James et al., 2021; Pang et al., 2023). Szintén a jövő tendenciáinak kedvez a gégétől a száj-, illetve orrnyílásig terjedő artikulációs csatorna, más néven vokális traktus modellezésére épülő beszédszintézis, ami főként vizuális információkra támaszkodik. Számos tanulmány hitelesen alátámasztja, hogy az emberi beszéd fiziológiai folyamatairól nyert vizuális információk nagymértékben elősegítik a beszédképzés komplex mechanizmusának megértését, és ezen keresztül a beszédszintézis módszereinek hatékony fejlesztését (Birkholz et al., 2020; Leppävuori et al., 2021; Skordilis et al., 2017). A napjainkban rendelkezésünkre álló radiológiai és monitorozó eljárások – úgymint mágneses rezonanciás képalkotás (MRI), komputertomográfia (CT), ultrahang (UH), elektropalatográfia (EPG), elektromágneses artikulográfia (EMA) vagy elektroglottográfia (EGG) – nélkülözhetetlen szerepet játszanak az akusztikai-artikulációs konverzió problémájának kezelésében. A fentebb említett képalkotó és monitorozó technikák segítségével generált morfológiai és geometriai adatok felhasználásával maradéktalanul feltérképezhetők az adott beszédjelhez tartozó artikulációs mozgások. Nem triviális feladat azonban az artikuláció akusztikummal való összekapcsolása, azaz a vokális traktus morfológiai és geometriai adataira alapozott beszédprodukció megvalósítása. Ilyen jellegű kutatási eredmények már felfedezhetők a szakirodalomban (Cao et al., 2023; Gonzalez-Lopez et al., 2020; Jin et al., 2022), de ez a terület sok szempontból még a mai napig is nyitott. Az eddig publikált tanulmányok főképpen a vokális traktus geometriai modelljének megalkotására fókuszálnak, aminek alapját az esetek többségében MRI-, UH- vagy EMA-felvételek képezik (Denby & Stone, 2024; Otani et al., 2023; Toutios & Narayanan, 2013), vagy pedig a beszédjelből származó lényegi információkkal manipulálnak (Kaburagi, 2014, 2015), a tökéletes minőségű gépi beszéd tényleges megvalósításához vezető út azonban bőven tartogat még kihívásokat. Ezenkívül az artikuláció geometriai és akusztikai jellemzői között

fennálló fizikai kapcsolatok természetét és hátterét illetően is számos nyitott kérdés vár még megválaszolásra. A problémafelvetés aktualitását mutatja, hogy az artikulációs-akusztikai kapcsolatrendszer feltárása, illetve gyakorlati leképezése alapvető fontosságú lehet például a klinikai célú beszédterápiában, a nem anyanyelvi nyelvtanulási tréningek kialakításában vagy a néma beszéd megszólaltatásához szükséges szintetizátorok konstrukciójában és fejlesztésében, ami szolgálhatja többek között a gégeeltávolításon átesett emberek rehabilitációját is.

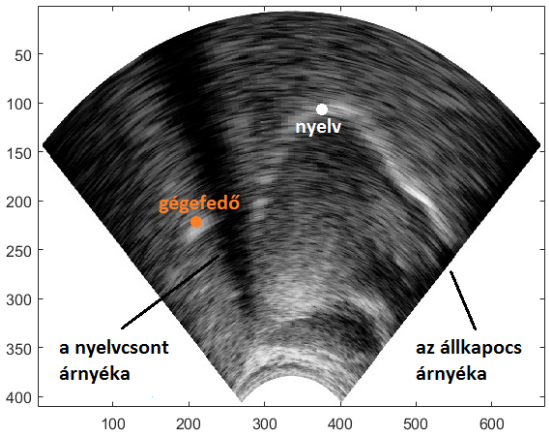
2. Az UH-felvételek jellemzése

Vizsgálataink alapvető eszközeit olyan audiovizuális források képezték, melyek ultrahangos (UH) eljárással készültek. A beszéd közben rögzített dinamikus felvételek képi formában megjelenítik a beszélő hangképző szerveit, miközben hallható a beszélő által kibocsátott akusztikus jel. A képi keretek sorozatán megfigyelhető néhány aktív hangképző szerv (nyelv, gégefedő) folytonos mozgása. A hang mint beszédjel időbeli eltolódás nélkül igazodik a felvételek képkockáihoz, így pontosan követhetők és egymáshoz rendelhetők a beszéd artikulációs és akusztikai mozzanatai. Végeredményben tehát létrejön az adott bemondáshoz tartozó, időben szinkronizált hang- és képcsomag. Az UH-felvételek kétdimenziós mozgóképek formájában álltak rendelkezésünkre. A felvételek síkját az emberi testet bal és jobb oldali részekre osztó függőleges szimmetriasík, az ún. középszagittális sík definiálja, amely lehetővé teszi a hangképző szervek kétdimenziós vetületi mozgásának, relatív elhelyezkedésének, illetve anatómiai szerkezetének részletes tanulmányozását. Az UH technikával többnyire csak a száj- és garatüreg egy része monitorozható kívülről, egy a beszélő álla alatt elhelyezett UH-fej alkalmazásával, melynek speciális elhelyezkedéséből adódóan az UH-felvételeken csak a nyelv és a gégefedő mozgása jeleníthető meg. A többi hangképző szerv ellenben nem látható, hiszen az ajkak és a gége kívül esik az eszköz letapogatási zónáján, a kemény és lágy szájpad pedig nem detektálható közvetlenül az UH-hullámok sajátos szájüregi visszaverődései miatt. Mindemel-

lett az UH-nyalábok nem képesek áthatolni a nyelvcsonton és az állcsonton, így a nyelv hátsó és elülső részeire árnyék vetül, aminek következtében a nyelv csak részlegesen mutatkozik meg. Itt jegyezzük meg, hogy a fogak nem láthatók a kereteken, mivel az UH-os eljárással kizárólag lágy szövetek tekinthetők át az emberi szervezetben.

Az 1. ábra egy statikus UH-keretet mutat be, ahol a vokális traktus látótérbe eső elemei különböző színekkel vannak felcímkézve. Látható, hogy a nyelvhát és a gégefedő csúcspontja világos sávokként rajzolódnak ki, a nyelvcsont és az állkapocs árnyékai pedig sötét zónák formájában észlelhetők.

Az általunk használt UH-csomag az MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoportjának Micro rendszerével készült. Az UH-felvételeken 40 különböző mondat hangzik el egy magyar női beszélő bemondásaiban.

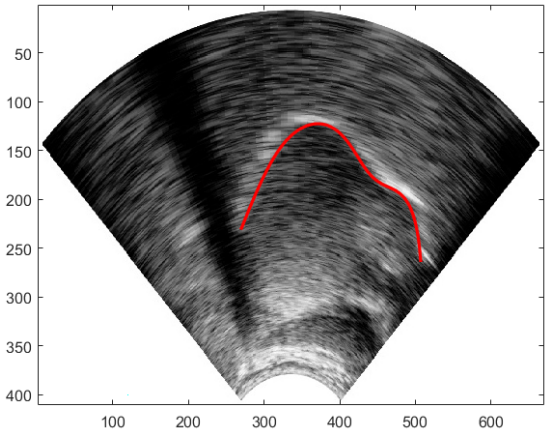


1. ábra. A vokális traktus látótérbe eső elemeinek elhelyezkedése egy statikus UH-kereten.

3. Az UH-felvételek anatómiai kontúrvonalainak megállapítása

A beszédhangok artikulációja során a nyelv alakja és pozíciója rendkívül fontos szerepet játszik, amit a nyelv felszíni formáinak tanulmányozásával lehet a legjobban leírni. Ebben nagy segítséget nyújthatnak az előző fejezetben bemutatott UH-felvételek, melyeken a nyelvszövet kétdimenziós vetületének határvonala nagyfokú pontossággal meghatározható a középszagittális síkban, ahol

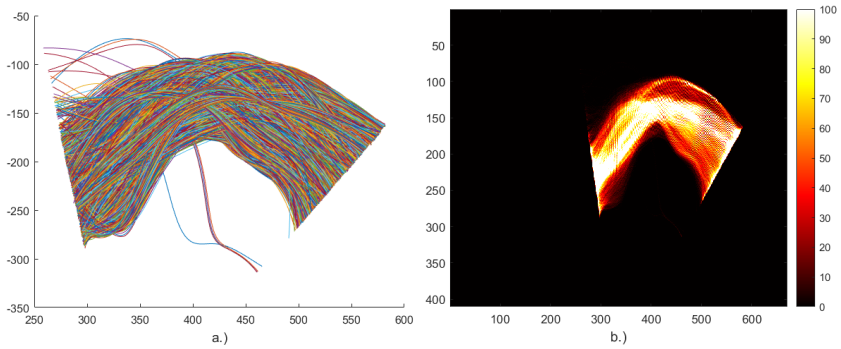
figyelemmel kísérhető a nyelv fel-le, illetve előre-hátra irányú mozgása. Az UH-keretek esetében a nyelvfelszín elmosódott, világos sávként azonosítható tartománya a nyelv és a felette lévő levegő határán kialakuló UH-hullámvisszaverődés eredményeként jön létre, így a nyelv felszíni vonalát a világos sáv alsó pereme jelöli ki (1. ábra). Mindezeket figyelembe véve, a nyelvkontúr követése a nyelvhát vonalát meghatározó képpontok sorozatának megkeresését jelenti a releváns képtartományon belül. A nyelvkontúrkövetés elsődleges célja a különböző beszédhangokhoz tartozó nyelvállások és nyelvalakok statikus vagy dinamikus leírása, illetve a koartikuláció során létrejövő hangátmeneteket jellemző nyelvmozgások vizsgálata. A kvalitatív analízis mellett a nyelvkontúr a beszéd kvantitatív jellegű tanulmányozásának is jó kiindulópontja lehet, hiszen a nyelvkontúrból származtatható számszerű értékek elősegíthetik az artikulációs modellek mélyebb megértését és fejlesztését. A vizsgálataink során egy olyan kontúrkövető algoritmust dolgoztunk ki és fejlesztettünk MATLAB-környezetben, amely a dinamikus programozás technikáját alkalmazza (Zhao & Czap, 2019). A módszer segítségével illesztett nyelvkontúrt a 2. ábra UH-keretén figyelhetjük meg az /o/ hang esetében.



2. ábra. A radiális geometriában megjelenő nyelvkontúr görbéje egy /o/ hanghoz tartozó UH-kereten.

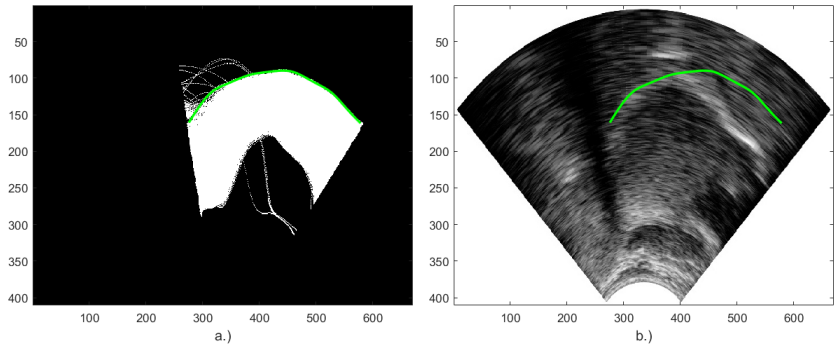
Az említett UH-felvételeken a nyelvkontúr mellett sokszor hasznos lehet a szájpadkontúr kijelölése is. Láttuk, hogy az UH-képeken egyáltalán nem mutatkozik meg sem a kemény, sem a lágy szájpad. Ezen oknál fogva különösen gondos megfontolásokat igényel a szájpadkontúr meghatározását célzó technika módszeres kidolgozása. Az UH-szájpadkontúr becslésének az volt az alapvető koncepciója, hogy a felvételeken megtaláljuk az artikuláció során a nyelvfelszín által érintett, legmagasabb helyzetben lévő szájüregi pontok halmazát, aminek nyomán valószínűsíthetjük a nyelv és a szájpad határvonalát. Ehhez természetesen olyan mássalhangzók vizsgálatára kell szorítkoznunk, melyek képzése során a nyelv biztosan érintkezik a szájpad palatális (kemény) és veláris (lágy) zónájával. Ez a feltétel a rendelkezésünkre álló, különböző bemondásokat tartalmazó UH-csomag esetében automatikusan teljesül, hiszen a rögzített mondatokban szereplő mássalhangzók képzésekor (pl. /k l ʃ t/) a nyelv más-más helyeken kerül kontaktusba a szájpad ívével. Így a szájpad kontúrjának kirajzolását lényegében egy szélsőérték-keresési probléma megoldásaként valósítottuk meg. Ehhez elsőként egy közös koordináta-rendszerben összegyűjtöttük a 40 UH-bemondás összes keretéhez tartozó 11111 darab nyelvkontúr görbéit, és ezzel párhuzamosan megalkottuk az összes nyelvkontúr lenyomata által kialakuló sűrűségtérképet is, melynek minden pixelje egy olyan valószínűségi mérték szerint kap értéket, hogy az adott képpont milyen gyakorisággal fordul elő lehetséges nyelvkontúrpontként. Az összes nyelvkontúr halmazát, illetve a görbeseregnek megfelelő sűrűségtérképet a 3. ábra demonstrálja.

A következő lépésben az összes nyelvkontúr által adott görbesereget egy olyan bináris sűrűségtérképen ábrázoltuk, amely csak 0 és 1 értékeket vehet fel aszerint, hogy a képmátrix adott pontját érinti-e bármely nyelvkontúrnak bármely pontja. Ennek értelmében a nyelvkontúrok által lefedett pixeleket 1 értékek jellemzik, a fennmaradó képpontokat pedig 0 értékekkel látjuk el, aminek folytán a kép fekete hátteréből fehér tartományként emelkedik ki az összes nyelvkontúr halmaza. A bináris sűrűségtérkép fekete-fehér kontrasztja kiváló terepet biztosít a kontúrkereső algoritmusunk futtatására, hiszen az eljárás segítségével detektálható a fekete és fehér domének maximális világosságú felső



3. ábra. A 40 UH-bemondás összes keretéhez tartozó 11111 darab nyelvkontúr halmaza (balra), illetve az összes nyelvkontúr által alkotott görbeseregnek megfelelő valószínűségi sűrűségtérkép (jobbra).

határvonala, ami éppen a szájpadkontúrt közelíti. A kontúrkereső algoritmussal kirajzolt szájpadkontúrt a 4. ábra zöld görbéje rögzíti a bináris sűrűségtérképen, illetve egy /k/ hanghoz tartozó UH-kereten.



4. ábra. A kontúrkereső algoritmussal detektált szájpadkontúr a bináris sűrűségtérképen (balra), illetve egy /k/ hanghoz tartozó UH-kereten (jobbra).

4. Artikulációs beszédszintézis

A gépi beszéd előállításának kiindulópontját az UH-felvételek képezték. A beszédszintézist olyan vizuális információkra támaszkodva valósítottuk meg, amik az említett kétdimenziós képi forrásokból közvetlenül vagy közvetve kinyerhetők. A szükséges geometriai adatok egy részét a felvételekre illesztett nyelv- és

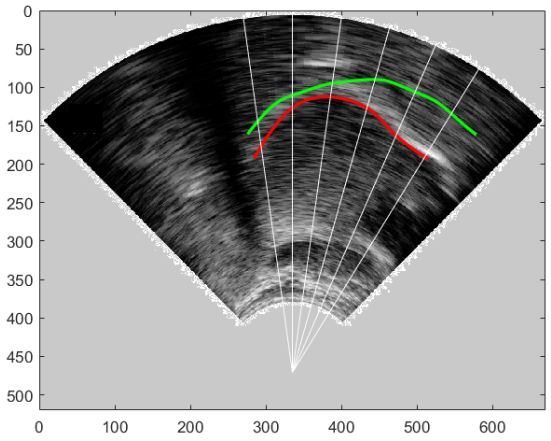
szájpadkontúrok segítségével származtattuk úgy, hogy MATLAB-környezetben kidolgoztunk egy algoritmust, melynek alkalmazásával a vokális traktusban dinamikus módon megmérhetők a szájpad és a nyelvfelszín közötti szagittális radiális távolságok. Az adathalmaz másik részét a nyelvkontúrokból kivont DCT-együtthatók formájában határoztuk meg. A kapott geometriai adatokat a gépi tanulás eszközei révén próbáltuk meg összekapcsolni a beszédet jellemző különböző artikulációs paraméterekkel, melyeket az akusztikus csőmodell vagy a lineáris predikció elvének keretében értelmeztük. Ennek során folyamatos beszédet kívántunk produkálni.

4.1. Dinamikus távolságmérés az UH-felvételeken

A vizuális adatok UH-felvételekből történő kinyeréséhez olyan anatómiai kontúrvonalakra volt szükségünk, melyek a lehető legteljesebb mértékben képesek lefedni a vokális traktus látótérbe eső tartományait. Ez az elvárás az UH-felvételek esetében már eddig is maximálisan teljesült, hiszen az előző fejezet szerint megkonstruált nyelv- és szájpadkontúrokon kívül nem detektálhatók további görbék a vokális traktusban. Ennek technikailag az az akadálya, hogy a részlegesen megjelenő szájüregen kívül az artikulációs csatorna többi része nem hozzáférhető az UH-letapogatás számára. A távolságméréshez két, különböző elveken alapuló mérési mechanizmust dolgoztunk ki az adatok dinamikus kinyerésére.

Az első megközelítésben a mérést radiális geometriában valósítottuk meg. Ehhez igazodva, elsőként egy rögzített középpontból indulva, sugárirányú metszeteket képeztünk az UH-felvételek releváns körcikkei által definiált tartományokban, amint azt az 5. ábra fehér vonalai is érzékeltetik a k hanghoz rendelt képkocka segítségével. A radiális metszetek a 0°-nál elhelyezkedő vertikális egyenes adott középpont körüli elforgatásával hozhatók létre a választott szögtartományon belül. Az 5. ábrán megrajzolt radiális metszetek a $[-8°, 32°]$ intervallum által megszabott szögtartományt fedik le úgy, hogy az egyes metszetek 8°-onként követik egymást. A releváns szögtartományt úgy állítottuk be, hogy az összes tanulmányozott bemondás képkockáira megfelelő legyen, azaz a

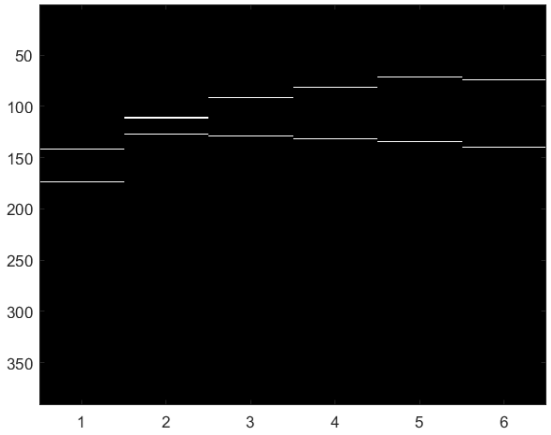
nyelvkontúr a nyelv mozgása során ne lépjen ki a felvett szögintervallumból. Az algoritmus második lépésében az 'intersect' függvény segítségével megkerestük a radiális metszetek és a nyelv-, illetve a szájpadkontúr által képzett metszéspontok koordinátáit, ami az 5. ábrával egyetértésben azt jelenti, hogy a fehér-piros, illetve a fehér-zöld görbepárok közös pontjait kell megtalálni. A görbepár metszéspontjainak birtokában kiszámolhatók a nyelv- és szájpadkontúr között mérhető radiális távolságok.



5. ábra. A szagittális radiális távolságok méréséhez felvett sugárirányú metszetek egy /k/ hanghoz tartozó UH-kereten.

A második megközelítésben a módszer kifejlesztése a radiális geometriának négyszöges elrendezésbe történő transzformációján alapult. Az eljárás fő lépése ugyanis az volt, hogy az UH-felvételek 5. ábrán rögzített középpontjából kiindulva és a kijelölt szögtartományt és szögbeosztást megőrizve, a létrejövő radiális irányok mentén mintavételeztük a nyelv- és szájpadkontúrokat, majd a kapott minták sorozatát mátrixos struktúrába rendeztük, ami a minták láncolatának négyszöges síkba való kifeszítését jelenti. A művelet eredményeként létrejövő, négyszöges geometriába alakított mintasorozatok láthatók a 6. ábrán, ahol a blokk felső, illetve alsó pixelláncai az 5. ábra szájpad-, illetve nyelvkontúrjaiból származnak. Ahogyan azt a 6. ábra vízszintes tengelye is mutatja, 6 oszlopot tartalmaznak a mátrixos struktúrák, ami azzal van összefüggésben, hogy

8°-onként beosztva a $[-8°, 32°]$ intervallumot, éppen 6 különböző szögérték áll elő. A mintasorozatok kényelmes alapot biztosítottak a távolságméréshez, amit úgy hajtottunk végre, hogy az összetartozó nyelv- és szájpadkontúrt leképező mátrix minden oszlopában kiszámítottuk a két vízszintes fehér vonal függőleges koordinátáinak különbségét.



6. ábra. Egy UH-keret szájpad- és nyelvkontúrjainak négyszöges geometriába transzformált mintasorozatai.

Kontroll céljából összehasonlítottuk a két távolságmérő módszer által produkált eredményeket, és arra jutottunk, hogy igen jó egyezés tapasztalható a két eljárással kapott kimenetelek között, hiszen a két adatsor elemeiben legfeljebb néhány pixelnyi eltérés mutatkozik. Ez a tendencia a felhasznált UH-keretek túlnyomó részében helytálló.

4.2. Diszkrét koszinusztranszformáció

A nyelvkontúrok mennyiségi jellemzésének egy lehetséges módja a diszkrét koszinusztranszformáció (Discrete Cosine Transform – DCT) alkalmazása, amely a matematikai transzformációknak egy speciális válfaja (Rao & Yip, 2014). A DCT segítségével egy N elemű valós $x_1, x_2, \ldots, x_N$ adathalmazt egy ugyanolyan elemszámú, szintén valós $X_1, X_2, \ldots, X_N$ értékhalmazra konvertálhatunk az

$$X_k = \sqrt{\frac{2}{N}} \sum_{n=1}^{N} x_n \frac{1}{\sqrt{1+\delta_{k1}}} \cos\Big(\frac{\pi}{2N}(2n-1)(k-1)\Big) \qquad (1)$$

formula szerint, ahol $k = 1, 2, \ldots, N$. (1) alapján az inverz operáció

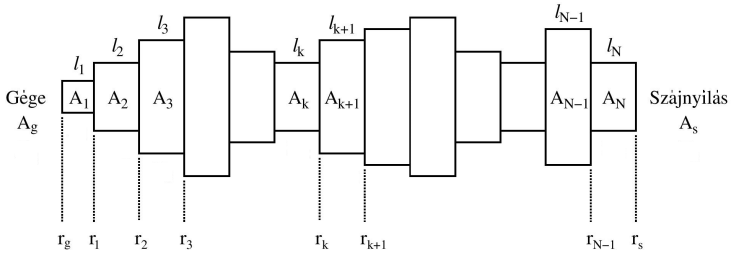
$$x_n = \sqrt{\frac{2}{N}} \sum_{k=1}^{N} X_k \frac{1}{\sqrt{1+\delta_{k1}}} \cos\Big(\frac{\pi}{2N}(2n-1)(k-1)\Big) \qquad (2)$$

alakban írható fel, mellyel az eredeti adattömb kapható vissza. Az (1) és (2) összefüggésekben δ_{k1} a Kronecker-deltát jelöli a $\delta_{k1} = 0$ (ha $k \neq 1$) és $\delta_{k1} = 1$ (ha $k = 1$) definíciónak megfelelően. Esetünkben az $x_1, x_2, \ldots, x_N$ paraméterek a nyelvkontúrok pontjainak vízszintes vagy függőleges koordinátáit testesíthetik meg, az $X_1, X_2, \ldots, X_N$ értékek pedig az adott görbéhez rendelt DCT-együtthatókat adják meg. A DCT inverz műveletével együtt kiválóan alkalmas görbék simítására, aminek igen nagy jelentősége van az UH-felvételekre illesztett nyelvkontúrok egyenetlenségeinek kiküszöbölésében is. Ennélfogva a DCT-együtthatók fontos információkat foglalnak magukba a nyelvkontúr alaki tulajdonságainak vonatkozásában.

4.3. Az akusztikus csőmodell

Az emberi beszédkeltés leképezésének egyik leggyakrabban alkalmazott és leghatékonyabb eszköze az akusztikus csőmodell (Fant, 1960). Ennek keretében a gégétől a száj- és orrnyílásig terjedő vokális traktust az egyik végén zárt, a másik végén nyílt rezonátorrendszerként kezeljük, mely akusztikai szűrőként működve egy több frekvenciakomponenst is tartalmazó hanghullámból csak adott frekvenciasávokba eső komponenseket enged át és sugároz ki a szabad térbe. A folytonos vokális traktust egy térben szabályosan kvantált csővel közelítjük, azaz felosztjuk N számú egymás után csatlakozó, egyenként állandó keresztmetszetű, rövid csőszakaszra, ahogyan azt a 7. ábra is szemlélteti. Az egyes csőszakaszok hosszúságát és keresztmetszeti paramétereit rendre az $l_1, l_2, l_3, \ldots, l_k, l_{k+1}, \ldots, l_{N-1}, l_N$, illetve az $A_1, A_2, A_3, \ldots, A_k, A_{k+1}, \ldots, A_{N-1}, A_N$ szimbólumok jelölik. A gégéhez és a szájnyíláshoz két extra keresztmetszetet

rendelünk A_g és A_s címkékkel. A 7. ábrán vázolt kép szerint gondolkodva azonnal felismerhető, hogy a szomszédos csőszakaszok csatlakozásánál történő ugrásszerű keresztmetszetváltásoknál a beszédhanghullámok visszaverődést szenvednek, amit fizikailag az $r_1, r_2, r_3, \ldots, r_k, r_{k+1}, \ldots, r_{N-1}$ reflexiós tényezőkkel írhatunk le. Ezt a halmazt kiegészítik a gégénél és a szájnyílásnál értelmezett r_g és r_s reflexiós tényezők. Az akusztikus csőmodell szerint $r_g = 1$ és $A_g = 1$. A modell tehát N csőszakasz alkalmazásakor $N + 1$ reflexiós tényezőt és $N + 2$ keresztmetszetet definiál.



7. ábra. A vokális traktus modellezése egyenként állandó keresztmetszetű, egyforma hosszúságú csőszakaszokkal.

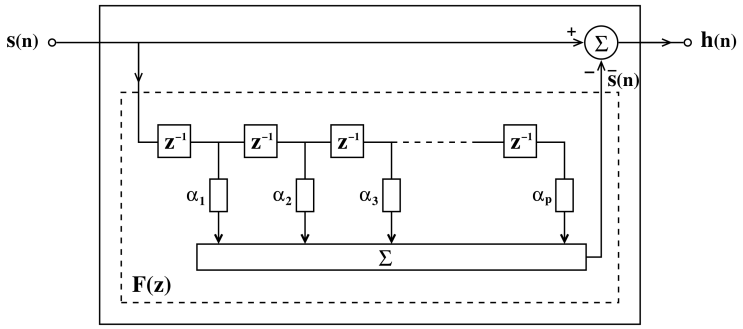
4.4. A lineáris predikció elve

A lineáris predikció (LPC – Linear Predictive Coding) a digitális jelek feldolgozásának és becslésének igen széles körben elterjedt és sikeres eszköze (Fant, 1960). Az eljárás azon alapul, hogy a jel adott pillanatbeli értékét az azt megelőző időpillanatokhoz tartozó jelértékek segítségével megpróbáljuk előrejelezni, idegen szóval predikálni. A predikció abban az esetben lineáris, ha a becsült jel a becslés során alkalmazott értékek lineáris függvénye. Amennyiben az s jel mint idősor n-edik elemét az azt megelőző p darab minta alapján állapítjuk meg, akkor a jelzett lineáris viszony az

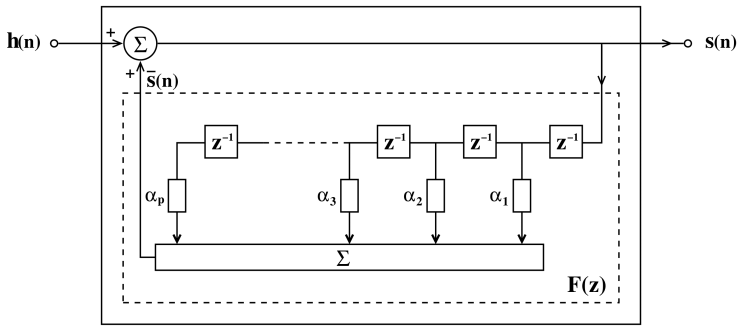
$$\overline{s}(n) = \sum_{i=1}^{p} \alpha_i \, s(n-i) \tag{3}$$

lineáris kombináció formájában jön létre, ahol az α_i koefficienseket predikciós együtthatóknak vagy LPC-együtthatóknak nevezzük. A p paraméter a predikció fokszámát adja meg.

A predikció 8. ábrán vázolt analízismodelljében a rendszer kimenetén az $s(n)$ eredeti és az $\overline{s}(n)$ becsült jel különbségeként előállítjuk a $h(n)$ hibajelet, a predikció 9. ábrán vizualizált szintézismodelljével pedig rekonstruálhatjuk az eredeti jelet a hibajel és a becsült jel összegeként. A 8. és 9. ábrák szaggatott vonallal határolt moduljaiban a z^{-1} faktorok egy ütemnyi késleltetést valósítanak meg, és a késleltetők láncolata mentén szorzásokat végzünk. Az i-edik lépésben kialakuló produktumot az aktuális leágazásnál található α_i koefficiens szorozza, végül a létrejövő tagokat a $\sum$ tömb összesíti.



8. ábra. A lineáris predikció analízismodellje.



9. ábra. A lineáris predikció szintézismodellje.

4.5. A neurális hálózatok felépítése

A beszédszintézisre irányuló törekvéseink során a gépi tanulás eszköztárához folyamodtunk, mivel a szintézishez szükséges artikulációs paramétereket geometriai adatokra alapozva, neurális hálózatok beiktatásával konstruáltuk meg. Folyamatos beszéd szintézisét irányoztuk elő, amihez az UH-felvételeken rögzített mondatokat használtuk fel. A hálózat bemeneti és kimeneti adathalmazát ugyanabból a forrásból származtattuk.

A rendszer bemenetén a nyelv- és szájpadkontúrok között mért szagittális radiális távolságokat vagy a nyelvkontúrokból kivont DCT-együtthatókat értelmeztünk. A távolságmérés során a korábban felvett 6 radiális metszettel dolgoztunk (5. ábra), a DCT-együtthatók számát pedig 9-re állítottuk be. A bemeneti paramétereket a korábbi fejezetekben ismertetett távolságmérő módszerek és transzformációs összefüggések szerint generáltuk. A rendszer kimenetén reflexiós tényezők vagy keresztmetszetek betanítását irányoztuk elő úgy, hogy 18 reflexiós tényezőt és 19 keresztmetszetet definiáltunk. A kimeneti paraméterek előállításához az adott bemondás akusztikus jeléből az 'lpc' függvény felhasználásával kiszámítottuk az LPC-együtthatókat, melyekből az 'lpcar2rf', majd az 'lpcrf2aa' függvények alkalmazásával meghatároztuk a reflexiós tényezők, illetve a vokális traktus keresztmetszeteinek halmazát. A neurális hálózat $(m \times n)$-es mátrixok formájában kezelt bemeneti és kimeneti blokkjainak dimenzióit a különböző rendszerbeállítások esetében az 1. táblázat foglalja össze. Az m sorindex a vizsgálatban részt vevő összes bemondáshoz tartozó képkeretek számát jelzi, míg az n oszlopindex a felvett radiális távolságok, DCT-együtthatók, reflexiós tényezők vagy keresztmetszetek számát fejezi ki.

Az 1. táblázatban feltüntetett paramétereket speciális technikával vontuk ki a beszédjelből. Folyamatos beszédről lévén szó, az egymáshoz kapcsolódó beszédhangok láncolata mentén haladva dinamikusan változik a vokális traktus alakja, és ezt a dinamikát a paraméterhalmaznak is tükröznie kell. Ezért a bemeneti adatok számításakor az adott bemondás összes keretét figyelembe kellett vennünk úgy, hogy minden egyes kerethez hozzárendeltük az aktuális nyelv- és szájpadkontúr alkalmazásával kapott radiális távolságokat és DCT-

1. táblázat. A neurális hálózat bemeneti és kimeneti adathalmazainak dimenziói az összes rendszerkonfiguráció esetében.

a bemenet mérete		**a kimenet mérete**	
radiális táv	**DCT-egyh.**	**reflexiós tény.**	**keresztmetszet**
9×6	–	9×18	9×19
4354×13	4354×9	4354×18	4354×19
6278×6	6278×9	6278×18	6278×19

együtthatókat. A kimeneti adatok kivonásakor pedig az adott bemondás teljes mintaszámát elosztva az összes keret számával, meghatároztuk a keretenkénti minták számát, így – a bemeneti oldalon érvényes módszerhez hasonlóan – ez esetben is minden egyes kerethez társítottuk az aktuális hangminták csoportjából eredeztetett reflexiós tényezőket és keresztmetszeteket.

A tanulóalgoritmus konstrukciójakor az skálázott konjugált gradiens (SCG – Scaled Conjugate Gradient) módszert alkalmaztuk a kimeneti paraméterek betanítására (Moller, 1993). A rendszerben egy rejtett réteget helyeztünk el, melybe 100 neuront ültettünk be. A hálózat kimeneti rétegének aktivációs függvényét a kimeneti paraméter típusához igazodva választottuk meg. Mivel a reflexiós tényezők értéke lehet pozitív és negatív is, esetükben megtartottuk a standard lineáris átvitelt, a keresztmetszetek tanításakor azonban a hiperbolikus tangens szigmoid függvény mellett döntöttünk, kizárva ezzel a negatív keresztmetszetek kialakulásának lehetőségét.

5. A beszédszintézis módszerei

5.1. Az akusztikus csőmodellen alapuló szintézis

Az akusztikus csőmodell keretében a vokális traktusban terjedő hanghullámot speciális fizikai mennyiségekkel jellemezhetjük, melyek közül programozástechnikai szempontból az $u(x,t)$ térfogatáram a legalkalmasabb a folyamatok modellezésére. A kétváltozós térfogatáram a vokális traktus tengelye mentén

mért x pozíciótól és a t időtől függ, és megmutatja az artikulációs csatorna adott keresztmetszetén egységnyi idő alatt átáramló levegő térfogatának mértékét. A térfogatáramot vektorfizikai mennyiségként kezeljük, mivel a nagyságán túl az iránya is mérvadó. A pozitív, illetve negatív x irányba terjedő térfogatáramot az $u^+(x,t)$, illetve az $u^-(x,t)$ szimbólumokkal láthatjuk el. A modell hipotéziseinek betartásával levezethető egy olyan komplett egyenletrendszer, mely egyértelműen összekapcsolja a szomszédos csőszakaszok határánál létrejövő térfogatáramokat, méghozzá oly módon, hogy a vokális traktus három sajátos tartományát külön-külön kezelve, eltérő szerkezetű egyenletcsoport alakul ki a gégénél található gerjesztési oldalon, a közbenső csőszakaszok régiójában, valamint a szájnyílásnál lévő sugárzási oldalon. A keletkező egyenletrendszer az

$$\begin{aligned}
u_1^+(t) &= \frac{1+r_g}{2}\, u_g(t) + r_g\, u_1^-(t) && \text{(a)} \\
u_{k+1}^+(t) &= (1+r_k)\, u_k^-(t-\tau_k) + r_k\, u_{k+1}^-(t) \\
u_k^-(t+\tau_k) &= -r_k\, u_k^+(t-\tau_k) + (1-r_k)\, u_{k+1}^-(t) && \text{(b)} \\
u_N^-(t+\tau_N) &= -r_s\, u_N^+(t-\tau_N) \\
u_s(t) &= -(1+r_s)\, u_N^+(t-\tau_N) && \text{(c)}
\end{aligned} \tag{4}$$

alakot ölti, ahol az (a), (b), (c) blokkok rendre a gerjesztési oldal, a közbenső csőszakaszok, illetve a sugárzási oldal térfogatáram-viszonyait jellemzik. A (4) egyenletrendszer tömörített jelölésformái szerint a k-adik csőszakasz elején pozitív vagy negatív irányba haladó térfogatáram az

$$u^{\pm}(x_k,t) = u_k^{\pm}(t), \tag{5}$$

míg a k-adik csőszakasz végénél pozitív vagy negatív irányba haladó térfogatáram az

$$u^{\pm}(x_k+l_k,t) = u_k^{\pm}(t\pm\tau_k) \tag{6}$$

szimbólumoknak megfelelően egyszerűsödik, $k = 1,2,\ldots,N$ indexelés mellett. Az (5) és (6) egyenlőségek bal oldali argumentumában szereplő helykoordinátát tehát a jobb oldali alsó index váltja fel, a (6)-ban megjelenő τ_k paraméter pedig azt az időeltolódást határozza meg, ami ahhoz szükséges, hogy

a $v = 343,14$ m/s sebességgel terjedő hanghullám áthaladjon az l_k hosszúságú csőszakaszon. A (4.a) és (4.c) egyenletekben szereplő $u_g(t)$ és $u_s(t)$ függvények speciálisan a gége által indukált gerjesztőjelet, illetve a szájnyíláson keresztül a szabad térbe sugárzott beszédjelet testesítik meg, az r_g, r_k, r_s faktorok pedig a 7. ábra szerint bevezetett reflexiós tényezőkkel azonosíthatók. A gépi beszéd produkciója a (4) egyenletrendszer programszintű ciklikus megvalósításával lehetséges, melynek során döntő szerepe van az $u_g(t)$ gerjesztőjelnek, hiszen a jel alakja alapvetően befolyásolja a szintézis eredményét. A gerjesztőjel jellegét természetesen az előállítandó hang típusa határozza meg attól függően, hogy zöngés vagy zöngétlen módon képződik. Zöngés gerjesztés esetén a gerjesztőjel periodikusan ismétlődő impulzusok formájában jön létre, zöngétlen gerjesztés esetén pedig véletlenszerűen változik. Az egymást követő beszédhangok típusának véletlenszerűségéhez igazodó zöngés-zöngétlen fluktuáció legegyszerűbben úgy indukálható, hogy gerjesztőjelként a lineáris predikció keretében értelmezett hibajelet alkalmazzuk. A csőmodell megvalósításához szükséges reflexiós tényezőket részben közvetlenül a neurális hálózat kimeneti csatornájából nyertük eleve betanított paraméterek formájában, másrészt pedig a betanított keresztmetszeteket átalakítottuk reflexiós tényezőkké az 'lpcaa2rf' függvény segítségével.

5.2. A lineáris predikció elvén alapuló szintézis

A predikciós elvű beszédszintézis végrehajtása a 8. és 9. ábrákon vázolt analízis- és szintézismodellek programszintű megalkotása révén lehetséges. Ehhez mindkét modell esetében ismernünk kell a teljes rendszer átviteli függvényét, ami a 8. és 9. ábrákon szaggatott vonallal határolt modul

$$F(z) = \sum_{i=1}^{p} \alpha_i \, z^{-i} \tag{7}$$

átviteli függvényére vezethető vissza, ahol z a komplex körfrekvenciát jelöli. Ennek megfelelően az analízismodell szerint megszerkesztett rendszer átviteli

függvénye

$$A(z) = 1 - F(z) = 1 - \sum_{i=1}^{p} \alpha_i \, z^{-i} = \sum_{i=0}^{p} \beta_i \, z^{-i} \tag{8}$$

alakban, a szintézismodell keretében felépített rendszer átviteli függvénye pedig

$$B(z) = \frac{1}{1 - F(z)} = \frac{1}{1 - \sum_{i=1}^{p} \alpha_i \, z^{-i}} = \frac{1}{\sum_{i=0}^{p} \beta_i \, z^{-i}} \tag{9}$$

formában kapható meg. A (8)-(9) egyenletekben megtörtént a (7) kifejezés helyettesítése, az utolsó egyenlőségek felírásakor pedig az α_i együtthatókat a β_i faktorokká transzformáltuk. A (8)-(9) átviteli függvényeket alapul véve, az analízis- és szintézismodell kimeneti jelei a 'filter' függvény segítségével generálhatók 'filter(a, b, x)' struktúrával, ahol az x-szel jelölt bemeneti adatokat a szűrésnek alávetett szintetizálandó hangminták vagy a hibajel alkotják, az a és b paraméterek pedig a vokális traktus átviteli függvényének számlálójában és nevezőjében megjelenő koefficienseket adják meg. Ennek megfelelően, a (8)-(9) kifejezésekhez igazodva, az analízismodell hibajelét az $a = \beta_i$ és $b = 1$, a szintézismodellel produkált beszédjelet pedig az $a = 1$ és $b = \beta_i$ együtthatók beállításával biztosíthatjuk. A hibajel előállításához szükséges β_i LPC-együtthatókat az 'lpc' függvény alkalmazásával vontuk ki az eredeti beszédjelből. A hangminták szintetizálásában részt vevő β_i LPC-együtthatókat pedig a betanított reflexiós tényezőkből, illetve keresztmetszetekből származtattuk az 'lpcrf2ar', illetve az 'lpcaa2rf' függvények felhasználásával.

6. Eredmények

Az eredeti és a szintetizált beszédjelek elérhetők és meghallgathatók az alábbi linken: https://drive.google.com/drive/folders/1LIz6y6wKZfMRIyE4EP1YW62SqXiZ-W8K?usp=drive_link. A 'MONDATOK' főmappában található almappák elnevezése minden esetben 'X_Y_Z' szerkezetű, ahol 'X' a szintézis során alkalmazott módszert rejti, 'Y' és 'Z' pedig a neurális hálózat bemeneti és kimeneti paramétereit egyértelműsíti, tehát az 'X_Y_Z' címke 'Y' adatokkal táplált és 'Z' adatokat betanuló neurális hálózat kimeneti eredményeit felhasználó, 'X'

módszerrel realizált szintézist takar. Ezzel összhangban az 'X' helyére 'cso' vagy 'lpc' kerül, ami az akusztikus csőmodellt vagy a lineáris predikció elvét jelzi. Az 'Y' pozícióban 'tav' vagy 'dct' feliratok lehetségesek, melyek a radiális távolságokat vagy a DCT-együtthatókat jelölik. A 'Z' elem 'ref' vagy 'ker' alakú, utalva a reflexiós tényezőkre vagy a keresztmetszetekre. Ezek mellett a főmappa tartalmazza az eredeti bemondások audiofelvételeit is. Az eredményeink elemzése során arra a következtetésre jutottunk, hogy a beszédszintézis minden esetben sikerrel zárult, mivel kvalitatív és kvantitatív szinten is relatíve jó egyezést tapasztaltunk az eredeti és a szintetizált bemondások között.

A kvalitatív értékelés során azt állapítottuk meg, hogy a szintetizált bemondások mindegyike elég jól érthető és a beszéd felismerhető az eredeti hanginformáció nélkül is, bár megjegyezzük, hogy a szintetizált jelekben zajkomponensek is tapasztalhatók, ami valamelyest rontja az akusztikai élményt, de természetesen a kutatómunkánk későbbi fázisaiban szeretnénk majd a torzításokat a lehető legjobb mértékben kiküszöbölni és javítani a beszéd minőségén.

Az eredményeink értékeléséhez néhány szintetizált bemondást szubjektív audioteszt formájában véleményezésre bocsátottunk egy erre a feladatra felkért, sok résztvevős célcsoport körében, akik független minősítőként semmilyen módon nem kapcsolódtak a kutatáshoz. A felmérés célja az volt, hogy a szubjektív audioteszt kimenetelének ismeretében egyértelműen állást lehessen foglalni, hogy az 1. táblázatban rendszerezett konfigurációk közül melyik bizonyul a leghitelesebbnek a beszédszintézisben, azaz melyik neurális hálózat beállításával lehet a legjobb minőségű beszédet előállítani. Ezzel a vizsgálattal nem a modellszintű hatékonyságot szerettük volna tesztelni, ezért az akusztikus csőmodell és a lineáris predikció révén kapott hangminták összehasonlítására irányuló kísérlet nem történt. Számunkra inkább az volt a legfőbb eldöntendő kérdés, hogy az 1. táblázat mely paraméterkombinációjának alkalmazásával érhető el a legtisztábban érthető gépi beszéd. A szubjektív minősítés megvalósításához kiválasztottunk egy lineáris predikció szerint szintetizált mondatot, és azt négy különböző változatban tártuk a célcsoport elé. A kiválasztott mondat A, B, C, D címkékkel ellátott négy verzióját a 2. táblázat sűríti, ahol azonosítható a

szintézis alapjául szolgáló neurális hálózat bemeneti és kimeneti paramétertípusainak összekapcsolása. A minősítésben részt vevő személyeket nem szerettük volna semmilyen módon befolyásolni, így nem hoztuk a tudomásukra, hogy az A, B, C, D felvételekhez milyen paraméterek vannak hozzárendelve, mindössze annyit közöltünk velük, hogy a négy mondatot négy különböző metódus szerint generáltuk, és pusztán az auditív percepció alapján kellett eldönteniük, hogy melyik esetben érthető legjobban a beszédjel.

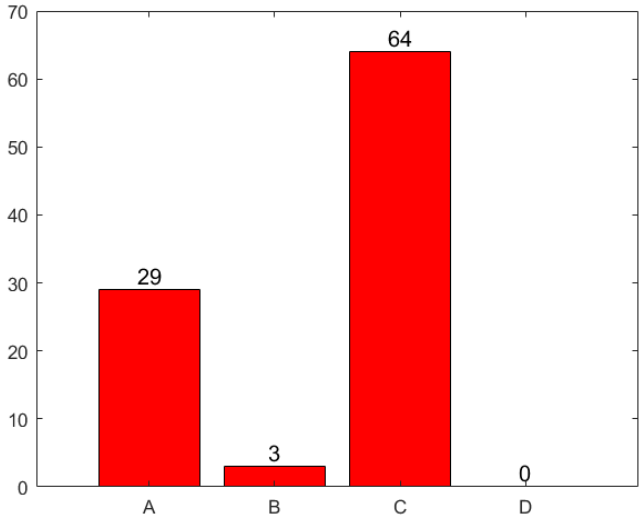
2. táblázat. A szubjektív tesztelésre bocsátott szintetizált audiofelvétel négy különböző változata a neurális hálózat paramétereinek függvényében.

	bemeneti paraméter	**kimeneti paraméter**
A	radiális távolság	reflexiós tényező
B	radiális távolság	keresztmetszet
C	DCT-koefficiens	reflexiós tényező
D	DCT-koefficiens	keresztmetszet

A szubjektív értékeléshez összesen 96 alany csatlakozott. Véleményük megoszlását a 10. ábra szemlélteti, ahol megfigyelhető, hogy 3 résztvevő kivételével senki nem voksolt a B és D variánsokra, a szavazatok lényegében az A és C verziók szintjén differenciálódnak többségében C szerinti állásfoglalással a 29:64 aránynak megfelelően, vagyis az A-hoz viszonyítva kb. kétszer annyian jelölték be a C-t. Ez az eredmény tehát arra enged következtetni, hogy a keresztmetszetekkel szemben a reflexiós tényezők betanításával jobb minőségű gépi beszéd állítható elő, és a tanítóalakzatok szintjén a radiális távolságokkal szemben előnyt élveznek a DCT-koefficiensek.

7. Összegzés

Jelen tanulmányban az artikulációs beszédszintézis különböző aspektusait vizsgáltuk MATLAB-környezetben. Elemzéseink során olyan ultrahangos (UH)



10. ábra. Az auditív percepción alapuló szubjektív tesztben részt vevők szavazatainak megoszlása négy különböző módon szintetizált mondat (A, B, C, D) próbájakor.

technikával készült audiovizuális forrásokra támaszkodtunk, melyek a kétdimenziós szagittális síkban vizualizálják a vokális traktus hangképző szerveinek relatív helyzetét és mozgását, miközben rögzítik a beszélő által kibocsátott audiojelet. Így a kép- és hangtartalom a szinkronizációnak köszönhetően egyértelmű módon összekapcsolódik.

Az UH-felvételek jó alapot biztosítottak ahhoz, hogy geometriai és akusztikus paramétereket nyerhessünk ki a kép- és hangforrásokból. A geometriai adatokhoz való hozzáférést nagyban megkönnyítette az automatikus kontúrkövető algoritmusunk alkalmazása, melyekkel elvégeztük az UH-keretek nyelvkontúrjainak dinamikus letapogatását. Az így kapott görbecsoportot kiegészítettük az UH-felvételekre rajzolt szájpadkontúrral, amit egy általunk kidolgozott eljárással konstruáltunk meg. A nyelvkontúrok ismeretében származtattuk a görbék diszkrét koszinusztranszformációjában részt vevő együtthatókat (DCT-együtthatók). Emellett a nyelv- és szájpadkontúrokból kiindulva kifejlesztettünk két különböző módszert a nyelv és szájpad anatómiai felszínei között mérhető szagittális radiális távolságok dinamikus meghatározására. Az akusztikus

paramétereket a beszédjelek időfüggvényeiből eredeztettük a lineáris predikció elvéhez kapcsolódó LPC-együtthatók formájában, melyeket az akusztikus csőmodellben értelmezett reflexiós tényezőkké, illetve a vokális traktus keresztmetszeti adataivá konvertáltunk.

Az artikulációs beszédszintézis során folyamatos beszédet állítottunk elő. A szintézis előkészítéseként olyan neurális hálózatokat szerkesztettünk, amik bemeneti adatként fogadják a vokális traktus szagittális radiális távolságaival és a nyelvkontúrokból kivont DCT-együtthatókkal felépített mátrixokat, a kimeneten pedig a beszédjelből származtatott reflexiós tényezők, valamint keresztmetszetek által alkotott struktúrákat produkálnak. A betanított paraméterek felhasználásával beprogramoztuk az akusztikus csőmodellt, illetve a lineáris predikció analízis- és szintézismodelljeit, melyek segítségével véghez vittük a gépi beszédprodukciót.

Köszönetnyilvánítás

Őszintén hálásak vagyunk és kegyeletteljes köszönetet mondunk Csapó Tamás Gábornak, hogy – amíg közöttünk volt – végtelenül segítőkész és önzetlen módon a rendelkezésünkre bocsátotta az MTA-ELTE Lendület Lingvális Artikuláció Kutatócsoportjának Micro rendszerével készült UH-felvételeket.

Hivatkozások

Arik, S., Chrzanowski, M., Coates, A., Diamos, G., Gibiansky, A., Kang, Y., Li, X., Miller, J., Ng, A., Raiman, J., Sengupta, S., & Shoeybi, M. (2017). Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning, PMLR* (p. 195–204). volume 70.

Arthur, F., & Csapó, T. (2021). Towards a practical lip-to-speech conversion system using deep neural networks and mobile application frontend. In *The International Conference on Artificial Intelligence and Computer Vision* (p. 441–450).

Ashok, K., Ashraf, M., Thimmia Raja, J., Hussain, M., Singh, D., & Haldorai, A. (2022). Collaborative analysis of audio-visual speech synthesis with sensor measurements for regulating human–robot interaction. *International Journal of System Assurance Engineering and*, *Management*, 1–8.

Balogh, G., Dobler, E., Grobler, T., Smodies, B., & Szepesvári, C. (2000). Flexvoice: A parametric approach to high-quality speech synthesis. In *IEE Seminar on State of the Art in Speech Synthesis* (p. 189–194).

Birkholz, P., Kürbis, S., Stone, S., Häsner, P., Blandin, R., & Fleischer, M. (2020). Printable 3d vocal tract shapes from mri data and their acoustic and aerodynamic properties. *Scientific Data*, *7*, 255.

Cao, B., Ravi, S., Sebkhi, N., Bhavsar, A., Inan, O., Xu, W., & Wang, J. (2023). Magtrack: A wearable tongue motion tracking system for silent speech interfaces. *Journal of Speech, Language, and Hearing Research*, *66*, 3206–3221.

Carlson, R., & Granström, B. (2008). *Rule-based speech synthesis*. Springer Handbook of Speech Processing.

Csapó, T. (2020). Speaker dependent articulatory-to-acoustic mapping using real-time mri of the vocal tract. In *INTERSPEECH 2020* (p. 2722–2726).

Csapó, T., Gosztolya, G., Tóth, L., Shandiz, A., & Markó, A. (2022). Optimizing the ultrasound tongue image representation for residual network-based articulatory-to-acoustic mapping. *Sensors*, *22*, 8601.

Csapó, T., Grósz, T., Gosztolya, G., Tóth, L., & Markó, A. (2017). Dnn-based ultrasound-to-speech conversion for a silent speech interface. In *INTERSPEECH* (p. 3672–3676).

Csapó, T., Zainkó, C., Tóth, L., Gosztolya, G., & Markó, A. (2020). Ultrasound-based articulatory-to-acoustic mapping with waveglow speech synthesis. In *INTERSPEECH 2020* (p. 2727–2731).

Denby, B., Csapó, T., & Wand, M. (2023). Future speech interfaces with sensors and machine intelligence. *Sensors*, *23*.

Denby, B., & Stone, M. (2024). Speech synthesis from real time ultrasound images of the tongue. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing* (p. –685). volume 1.

Fant, G. (1960). Acoustic theory of speech production.

Gonzalez-Lopez, J., Gomez-Alanis, A., Donas, J., Pérez-Córdoba, J., & Gomez, A. (2020). Silent speech interfaces for speech restoration: A review. *IEEE access*, *8*, 177995–178021.

James, J., Balamurali, B., Watson, C., & MacDonald, B. (2021). Empathetic speech synthesis and testing for healthcare robots. *International Journal of Social Robotics*, *13*, 2119–2137.

Jin, Y., Gao, Y., Xu, X., Choi, S., Li, J., Liu, F., Li, Z., & Jin, Z. (2022). Earcommand: "hearing" your silent speech commands in ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *6*, 1–28.

Juanpere, E., & Csapó, T. (2019). Ultrasound-based silent speech interface using convolutional and recurrent neural networks. *Acta Acustica united with Acustica*, *105*, 587–590.

Kaburagi, T. (2014). Determining the length and cross-sectional area of the vocal tract jointly from formants using acoustic sensitivity function. *Acoustical Science and Technology*, *35*, 290–299.

Kaburagi, T. (2015). A method for estimating vocal-tract shape from a target speech spectrum. *Acoustical Science and Technology*, *36*, 428–437.

Kaur, N., & Singh, P. (2022). Speech waveform reconstruction from speech parameters for an effective text to speech synthesis system using minimum phase

harmonic sinusoidal model for punjabi. *Multimedia Tools and Applications*, *81*, 26101–26120.

Király, J. (1989). A pc talker beszédszintetizátor és digitális hangrögzítő-visszajátszó rendszer. O.

Kiss, G., Arató, A., Lukács, J., Sulyán, J., & Vaspöri, T. (1987). Brailab, a full hungarian text-to-speech microcomputer for the blind. In *Proceedings of World Conference on Phonetics* (p. 1–4).

Leppävuori, M., Lammentausta, E., Peuna, A., Bode, M., Jokelainen, J., Ojala, J., & Nieminen, M. (2021). Characterizing vocal tract dimensions in the vocal modes using magnetic resonance imaging. *Journal of Voice*, *35*, 804– 27.

Mahum, R., Irtaza, A., & Javed, A. (2023). Text to speech synthesis using deep learning. In *Intelligent Multimedia Signal Processing for Smart Ecosystems* (p. 289–305). Cham: Springer International Publishing.

Moller, M. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural networks*, *6*, 525–533.

Mullah, H. (2015). A comparative study of different text-to-speech synthesis techniques. *International Journal of Scientific Engineering and Research*, *6*, 287–292.

Olaszy, G. (1989). Multivox – a flexible text-to-speech system for hungarian, finnish, german, esperanto, italian, and other languages for ibm-pc. In *First European Conference on Speech Communication and Technology (EUROSPEECH '89* (p. 2525–2528). Paris: France.

Olaszy, G., G., O., P., K., G., Z., C., & Gordos, G. (2000). Profivox — a hungarian text-to-speech system for telecommunications applications. *International Journal of Speech Technology*, *3*, 201–215.

Olaszy, G., & Gordos, G. (1987). Automatic text-to-speech system applied in a reading machine. In *Proceedings of European Conference on Speech Technology (EUROSPEECH '87* (p. 25–29). Edinburgh, UK.

Otani, Y., Sawada, S., Ohmura, H., & Katsurada, K. (2023). Speech synthesis from articulatory movements recorded by real-time mri. In *Proceedings of the Annual Conference of the International Speech Communication Association, Interspeech* (p. 127–131).

Panda, S., & Nayak, A. (2017). A waveform concatenation technique for text-to-speech synthesis. *International Journal of Speech Technology*, *20*, 959–976.

Pang, B., Teng, J., Xu, Q., Song, Y., Yuan, X., & Li, Y. (2023). Chinese personalised text-to-speech synthesis for robot human–machine interaction. *IET Cyber-Systems and Robotics*, *5*, 12098.

Rao, K., & Yip, P. (2014). *Discrete cosine transform: algorithms, advantages, applications*. Academic Press.

Shiga, Y., Ni, J., Tachibana, K., & Okamoto, T. (2020). Text-to-speech synthesis. In Y. Kidawara, E. Sumita, & H. Kawai (Eds.), *Speech-to-Speech Translation*. Singapore: Springer Briefs in Computer Science. Springer.

Skordilis, Z., Toutios, A., Töger, J., & Narayanan, S. (2017). Estimation of vocal tract area function from volumetric magnetic resonance imaging. In *2017 IEEE International Conference on Acoustics, Speech and Speech and Signal Processing (ICASSP* (p. 924–928).

Toutios, A., & Narayanan, S. (2013). Articulatory synthesis of french connected speech from ema data. In *Interspeech* (p. 2738–2742).

Tóth, L., Gosztolya, G., Grósz, T., Markó, A., & Csapó, T. (2018). Multi-task learning of speech recognition and speech synthesis parameters for ultrasound-based silent speech interfaces. In *INTERSPEECH* (p. 3172–3176).

Zhao, L., & Czap, L. (2019). A nyelvkontúr automatikus követése ultrahangos felvételeken. *Beszédkutatás*, *27*, 331–343.