XX. Fiatal Műszakiak Tudományos Ülésszaka, 2015. Kolozsvár, 119–122. *DOI: 10.33895/mtk-2015.03.22*
http://hdl.handle.net/10598/28606

GYENGE MINŐSÉGŰ BESZÉD SZEGMENTÁLÁSA

SEGMENTATION OF LOW QUALITY SPEECH

Czap László[1], Pintér Judit Mária[2]

[1,2]Miskolci Egyetem, Gépészmérnöki és Informatikai kar, Villamosmérnöki Intézet, Automatizálási és Infokommunikációs Intézeti Tanszék, Cím: 3515, Magyarország, Miskolc- Egyetemváros, Telefon / Fax: +36-46-565-140, czap@uni-miskolc.hu, pinterjm@uni-miskolc.hu

Abstract

We are working on computer based assessment of speech production that needs the segmentation. Comparison of the reference and actual utterance requires dynamic time warping. There are computer algorithms that work well for good quality of speech. Speech of hearing impaired people is usually unrhythmical. Their performance for deformed, sometimes drawling and jerky pronunciation is very weak. We are going to modify the algorithm for correct segmentation of low quality speech.

***Keywords**: speech recognition, feature extraction, teaching hearing impaired children.*

Összefoglalás

A beszéd automatikus minősítésének számítógépes megoldásán dolgozunk, aminek kulcskérdése a helyes szegmentálás. A referencia és a vizsgált beszéd összehasonlításához dinamikus idővetemítésre van szükség, amire a számítógépes beszédfeldolgozásban kidolgozott eljárások állnak rendelkezésre. Ezek a módszerek jó minőségű beszédre készültek. A hallássérültek beszédére különösen jellemző az egyes hangok megszokottól eltérő idejű artikulációja. A torz hangokra, a rendkívül elnyújtott, akadozó beszédre gyenge eredményt szolgáltatnak. Célunk a szegmentálásra szolgáló módszerek továbbfejlesztése annak érdekében, hogy a szinte érthetetlen beszédre is használható szegmentálási eredményeket kapjunk.

***Kulcsszavak:** beszédfelismerés, lényegkiemelés, hallássérült gyerekek oktatása.*

1. Bevezetés

Az „Alap- és alkalmazott kutatások hallássérültek internetes beszédfejlesztésére és az előrehaladás objektív mérésére” címet viselő projekt a Debreceni és a Miskolci Egyetem közös kutatása, amelynek a célja egy komplex rendszer létrehozása, mely a beszédfolyamat audiovizuális megjelenítését szolgáltatja, egyrészről a beszéd hangképeinek másrészről az artikulációnak a vizuális megjelenítésével (beszélő fej), egy oktatási keretrendszerbe foglalva [1]. Ezek mellett számos olyan funkciókat is tartalmaz majd a rendszer (prozódia megjelenítés, automatikus minősítés, tudásalapú rendszer implementálása), amelyek a későbbiekben lehetővé teszik az egyéni gyakorlást nem csak számítógépen, hanem mobil eszközön is. A felsorolt funkciók közül kiemelten fontos szerepet játszik az egyéni gyakorlásban az automatikus minősítés és visszajelzés a gyakorlást végző személy számára. Az aktuális kiejtést általában a

spektrum alapján hasonlítják össze a referencia bemondással. Ez az összehasonlítás nem egyezik azzal az ítélettel, amit a hallgató az érthetőségről megállapít. A szubjektív értékelést szurdopedagógusok bevonásával a torz beszédre elvégezve megalkotható a minőségi skála. A cél ennek a skálának és a visszajelzésnek a létrehozása, ami alapján az előrehaladás is nyomon követhetővé válik. A skála megalkotásához alapmintákat gyűjtöttünk be a célkorosztályban a beszédminőség felmérésére, ezt a második fejezet taglalja részletesebben. A helyes és az aktuális kiejtés összehasonlításához elengedhetetlen volt a különböző lényegkiemelési eljárások vizsgálata, a szegmentálási pontosságuk meghatározása és egy olyan dinamikus idővetemítési eljárás megalkotása, amely gyenge minőségű beszéd esetén is hasonló hatékonysággal működik, mint helyes kiejtésnél. Cikkünkben a megalkotott eljárás sajátosságait foglaljuk össze és az elvégzett előzetes összehasonlítással szemléltetjük annak hatékonyságát.

2. Szóadatbázis az automatikus minősítés megalkotásához

A már említett minősítési skála megalkotásához szükséges adatbázis mintáit eltérő beszédprodukciós fejlettségi fokú személyektől gyűjtöttük be és minősíttettük laikus hallgatókkal (akik ritkán találkoznak hallássérültekkel), valamint a szurdopedagógusokkal, a szépen beszélő ép hallóktól az alig érthetően beszélő hallássérültekig.

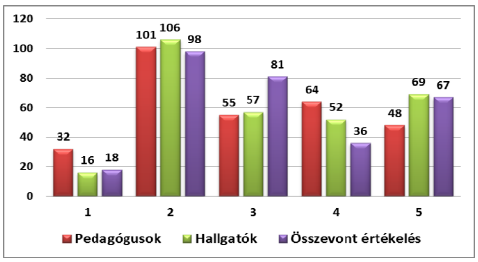
Az adatbázisban pontosan 2421 szó szerepel (egyes szavak többszörösen is előfordulnak, de a bemondók eltérőek, ezért azok érthetősége is), amit 13 pedagógus és 23 hallgató értékelt oly módon, hogy a szurdopedagógusok a saját oktatási intézményük diákjait nem értékelték. Az értékelés alapját a pedagógusok által meghatározott 5 fokozatú skála képezte. A skála értelmezése:

- Érthetetlen (1): az artikuláció teljesen torz; felismerhetetlenek a magán-és mássalhangzók; a szótagszám visszaadása sem megfelelő vagy nem kivehető; a levegővétel, a levegővel való gazdálkodás helytelen; rossz a tempó, a ritmus; dallamtalan, dinamikátlan vagy túl feszített a hangadás;
- Nehezen érthető (2): súlyos torzítások, hangelhagyások, hangcserék; csak a magánhangzók egy része kivehető; a légzés elégtelensége miatt létrejövő torzítások, pl. túl levegős vagy fojtott; eltérő, zavaró hangszín, ritmus, tempó jellemzi;
- Közepesen érthető (3): a magánhangzók ejtése helyes, a szótagszám megfelelő; súlyos beszédhibák előfordulhatnak pl. diszlália, orrhangzósság, fejhangzósság, stb. Prozódiai elégtelenségek;
- Jól érthető (4): csekély mértékű beszédhibák; enyhe prozódiai elégtelenségek;
- Hallók beszédével azonos szinten érthető (5): legfeljebb 1-2 hanghiba fordulhat elő.

A szegmentálási vizsgálatok elvégzéséhez ebből a 2421 szóból választottunk ki 300 szót. A kiválasztott szókészlet elég változatos nemcsak a szavak hosszúsága alapján, hanem a hangkapcsolatok előfordulásának szempontjából is.

Az 1. ábrán a szavak értékelés szerinti eloszlása látható: nemcsak az összevont, hanem külön a hallgatóké és a pedagógusoké. Azért is érdemes külön szemléltetni az értékelések eloszlását, mert a hallgatók az esetek 63 %-ban jobbra értékelték a bemondást a szurdopedagógusokkal szemben. Továbbá az is megfigyelhető, hogy az egyesre értékelt mintákból keveset tartalmaz az adatbázis. Ennek oka, hogy a 2421 szó értékelésében sem fordult elő több ilyen érthetetlen bemondás, ezt kompenzálva a

kettesre értékeltekből többet válogattunk be a tesztelésre szánt adatbázisba.



1. ábra *A szavak eloszlása értékelések alapján*

3. Dinamikus idővetemítés

Általánosságban elmondható, hogy ha ugyanazon mondat vagy szó két kiejtését össze akarjuk hasonlítani (pl.: a referencia bemondást az aktuális bemondással), akkor azokat hangszinten azonos időbeli hosszra kell skálázni, azaz egymáshoz vetemíteni. Helyesen beszélő emberek esetén is a beszéd tartalmaz nemlineáris megnyúlásokat és rövidüléseket, amik nem feltétlenül számítanak hibás ejtésnek. Ezek az előfordulások hallássérült személyek esetén pedig hatványozottan jelennek meg, ezért ebben az esetben nem érdemes lineáris időskálázást alkalmazni.

Egy korábban kifejlesztett vetemítési módszer, amely több nyelvre is alkalmazható, szigorúan csak a beszédhangok mentén illeszti össze a két mintát és csak azokon belül valósít meg lineáris skálázást. A nemlineáris vetemítő eljárás gépi automatikus szegmentáláson alapszik, amelynél általános akusztikai hangosztályokat használtak [2].

Egy másik vetemítési eljárás az optimális időillesztést, mint minimális hosszúságú, illetve súlyú út keresését tekinti egy adott gráfban.

– bármely x_i vektort csak egyszer ismételhetünk meg (tehát legfeljebb duplázhatunk, de már nem triplázhatunk);
– ha x_i-t elhagytuk, akkor a szomszédait (x_{i-1}-et és x_{i+1}-et) nem hagyhatjuk el, tehát két szomszédos szegmens már nem hagyható el;
– a szegmensek sorrendje nem cserélhető fel.

4. Az idővetemítés szabályainak módosítása a gyenge minőségű beszéd szegmentálására

Az ismert és az előzőekben példaként leírt szabályok alapján egy referencia szóhoz az idővetemítés nem bizonyult sikeresnek.

4.1 Az alkalmazott lényegkiemelés

Összefüggésben a következőkben tárgyalandó a referenciagenerálással olyan lényegkiemelési módot kellett választanunk, amely alkalmazkodik a mesterségesen generált referenciához. Neurális hálózatokat tanítottunk be, amelyek osztályozzák a hangokat és az osztályon belül kimeneti aktivitást generálnak az egyes hangokhoz. A neurális hálózatok által képezett osztályok

– szünet;
– magánhangzó (a, á, e, é, i, o, u, ü);
– fél magánhangzó (m, n, ny, r, l, j);
– réshang (f, sz, s, h, v, z, zs);
– zárhang. (p, t, ty, k, b, d, gy, g);

(Az osztályokhoz tartozó külön neurális hálózat kimeneteihez tartozó hangokat soroltuk fel.)

4.2 A referenciagenerálás

A dinamikus idővetemítésnél a keresett szót egy referencia bemondással vetjük össze és keressük egyes időkeretek megismétlésével illetve kihagyásával a referencia bemondáshoz leginkább hasonló ütemezést. A hallássérült gyerekek bemondásai közül a nehezen érthetők nem alkalmasak arra, hogy a referencia bemondáshoz valamilyen hasonlósági mérték szerint eléggé hasonlítsanak. Próbálkoztunk férfi, női, gyerek be-

mondáshoz és szintetizált hanghoz is vetemíteni a keresett szavakat, ezek azonban nem voltak sikeresek. Egy 300 bemondótól származó 4 és fél órás hangadatbázis alapján PLP lényegkiemelést alkalmazva euklideszi távolságot képeztünk a magyar beszédhangok átlagai között. A normalizált távolság megfordításával hasonlóságmértéket képeztünk az egyes hangok között. A referenciát úgy alakítjuk ki (ebben a feladatban nem a szó felismerése a cél, hanem a bemondás minősítése, ezért rendelkezésre áll a vizsgált szó fonetikus leírása). Az egyes hangok referencia időtartamát az adott beszédhang átlagos hosszára állítottuk be [3]. A neurális hálózatok kimenete optimális esetben 0 illetve 1. Mivel a hangok is gyakran rendkívül torzak és az osztályozás sem hibátlan a vetemítéshez a neurális hálózatok kimeneteit a hasonlósági mértékkel súlyozva a megfelelő osztály hangjaihoz is hozzárendeljük. Ily módon ha a hang torz vagy a neurális hálózat nem megfelelő kimenete mutatja a legnagyobb aktivitást, akkor is kapunk a megfelelő kimeneten 0-tól eltérő jelet. A referencia előállításánál az adott hanghoz tartozó időszegmensre 1-t állítunk be a megfelelő osztály kimenetén és az adott hanghoz tartozó kimeneten.

A dinamikus vetemítés 3. fejezetben ismertetett szabályaival a vetemítés jobb eredményt mutatott, mint azokban az esetekben, amikor referenciaként bemondott szavakat használtunk.

A hallássérült gyerekek bemondásában gyakran találkoztunk hangok között több tized másodperces szünetekkel és hosszan ejtett hangokkal.
Ezért:

– minden hang után beiktattunk a referencia előállítása során egy szünetet;
– a szünet akárhányszor ismétlődhet.

A korábban ismertetett szabályok szerint egy időintervallumot maximum kétszeresére lehet nyújtani. A hallássérült gyerekek bemondásában azonban ennél hosszabban ejtett hangokkal is gyakran találkoztunk. Ezért:

– egy időkeret kétszeri ismétlését is megengedjük, ezzel egy időintervallum háromszorosára nyújtható.

5. Következtetések

A cikkben leírt módosított vetemítési eljárás előzetes eredményei a korábbi megoldásoknál lényegesen jobb szegmentálási pontosságot mutatnak. A más eljárásokkal végzett vetemítési eredményekkel való kvantitatív összehasonlításon jelenleg is dolgozunk.

Köszönetnyilvánítás

A kutató munka a Miskolci Egyetem stratégiai kutatási területén működő Mechatronikai és Logisztikai Kiválósági Központ keretében, a TÁMOP-4.2.2. C-11/1/KONV-2012-0002 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Szakirodalmi hivatkozások

[1] Czap L., Pintér J.: *A beszédasszisztens koncepció*, Multidiszciplináris Tudományok; A Miskolci Egyetem Közleménye 3; 2013; 241-250 oldal.

[2] Kiss G., Vicsi K.: *Akusztikai hangosztályok felismerésén alapuló, nemlineáris idővetemítés megvalósítása a mondathanglejtés és a szóhangsúlyozás oktatásához*; Beszédkutatás 21, 2013, 247-261. oldal.

[3] Németh G., Olaszy G.: *A magyar beszéd,* Akadémiai kiadó, Budapest, 2010