Miskolci Egyetem

GÉPÉSZMÉRNÖKI ÉS INFORMATIKAI KAR

A BESZÉDMINŐSÉG AUTOMATIKUS ÉRTÉKELÉSE

PhD értekezés

Készítette:

Pintér Judit Mária

okleveles mérnökinformatikus

Hatvany József Informatikai Tudományok Doktori Iskola

Doktori iskola vezető:

Prof. Dr. Szigeti Jenő

A matematikai tudomány kandidátusa

Tudományos vezető:

Dr. Czap László

Egyetemi docens, PhD

Miskolc

2015

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.

Nyilatkozat

Alulírott Pintér Judit Mária kijelentem, hogy ezt a doktori értekezést magam készítettem, és abban csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerinti vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Miskolc, 2015. április 29.

Pintér Judit Mária

A disszertáció bírálatai és a védésről készült jegyzőkönyv megtekinthető a Miskolci Egyetem Gépészmérnöki és Informatikai Karának Dékáni Hivatalában, valamint a doktori iskola weboldalán az Értekezések menüpont alatt: http://www.hjphd.iit.uni-miskolc.hu

Témavezetői ajánlás

Pintér Judit Mária: **”A beszédminőség automatikus értékelése”** című PhD értekezéséhez

Pintér Judit Mária érdeklődése hallgató korában fordult a számítógépes beszédfeldolgozás felé. Szakdolgozatában a beszédfelismerés különböző paramétereit optimalizálta, amellyel Gábor Dénes Diplomatervezési díjat érdemelt ki. Kutatásait TDK munkák formájában folytatta. Diplomamunkájában a beszédfelismerés ipari alkalmazására dolgozott ki megoldást.

Bíztatásomra megkezdett kutatásait doktorandusz hallgatóként is folytatta, amelynek témáját a Miskolci Egyetem és a Debreceni Egyetem együttműködésében folyó a hallássérült gyerekek beszélni tanítását segítő rendszer problémái határozták meg. A beszédfeldolgozás egyes kérdéseire – például a szegmentálás – hibátlan artikuláció és jó minőségű felvételek esetén sincs tökéletes megoldás. Fokozottan merülnek fel a problémák hibás artikulációjú, torz és akadozott beszéd esetén, amivel a hallássérültek felvételeinél szembesültünk. Az ismert megoldások jelentős továbbfejlesztését igényelte a nem szokványos feladatok megoldása. Pintér Juditot nagy munkabírású, érdeklődő, motivált hallgatóként ismertem meg. A tanszék életét ötleteivel, javaslataival felpezsdítette.

Kutatási eredményeit hazai és nemzetközi konferenciákon és folyóiratokban publikálta, ezek anyagi hátterét sikeres pályázatokkal teremtette meg.

Az értekezés Pintér Judit Mária önálló eredményeit tartalmazza és a Hatvany József Informatikai Tudományok Doktori Iskola szabályzatában megkövetelt követelményeknek mindenben megfelel. A fentiek alapján a jelölt számára a PhD cím odaítélését messzemenően támogatom.

Miskolc, 2015. április 29.

Dr. Czap László

tudományos vezető

Köszönetnyilvánítás

Az értekezés 2012-ben kezdett kutatómunkám eredményeit foglalja össze, melyet a Hatvany József Informatikai Tudományok Doktori Iskola keretein belül végeztem, a Miskolci Egyetem Automatizálási és Infokommunikációs Intézeti Tanszékén.

Először szeretnék köszönetet mondani tudományos vezetőmnek, Dr. Czap Lászlónak, aki már egyetemi hallgató korom óta készséggel támogatja tudományos tevékenységemet. Számtalan szakmai és baráti tanáccsal látott el, melyek segítettek a kitűzött cél elérésében. Szakmai tudásának átadásával járult hozzá, hogy ez az értekezés elkészülhessen.

Hálásan köszönöm kollégáim segítőkészségét, türelmét és megértését, amellyel támogattak disszertációm elkészítése során.

Köszönöm projekt munkatársaimnak és a kutatómunkában résztvevő pedagógusoknak segítőkészségüket és szakmai hozzájárulásokat a vizsgálatok alapját képező hangminták rögzítéséhez és disszertációm elkészítéséhez.

Szüleimnek és testvéremnek köszönöm a belém vetett hitüket és az értem hozott áldozataikat, amellyel lehetővé tették egyetemi tanulmányaimat.

Végül, de nem utolsó sorban köszönöm barátaimnak, hogy mindenben segítettek és biztattak és külön köszönöm Demény Anitának, hogy végig mellettem állt és támogatott.

Tartalomjegyzék

Rövidítések listája

ADTW	Adapted Dynamic Time Warping
AMDF	Average Magnitude Difference Function
AMDF	Average Magnitude Difference Function
ANN	Artificial Neural Network
DTW	Dynamic Time Warping
HMM	Hidden Markov-model
MFCC	Mel-Frequency Cepstral Coefficients
MOS	Mean Opinion Score
NN	Neural networks
PEAKS	Program for Evaluation and Analysisof all Kinds of Speech Disorders
PLP	Perceptual Linear Prediction

Ábrajegyzék

Táblázatjegyzék

1 Bevezetés

Az értekezésben a TÁMOP-4.2.2.C-11/1/KONV-2012-0002 azonosító számú, "Alap- és alkalmazott kutatások hallássérültek Internetes beszédfejlesztésére és az előrehaladás objektív mérésére" c. projekt kapcsán végzett kutatómunkám eredményeit mutatom be. A fő fejezetek a beszédkeltés fiziológiáját, a beszédfelismerés alapjait, a siketek és nagyothallók beszédtanulási nehézségeit és lehetőségeit, valamint a kutatási projektben megalkotott beszédminőség automatikus kiértékelésének gyakorlati alkalmazását és megvalósításának kulcslépéseit mutatják be.

A beszéd nem más, mint akusztikus hullámok keltése, azaz beszédhangok, fonémák (hangok olyan elemi, elvont egysége, amely szavakat különböztet meg egymástól, önálló jelentéssel nem rendelkezik) kibocsátása. A beszéd nem csupán fonémák sorozata, hanem fontos a hangsúlyozás, a hanglejtés és számos más szupraszegmentális jellemző is. Ezek alapján egyértelmű, hogy a beszéd az emberek legfőbb kommunikációs eszköze, amiért az akusztikus beszédfelismerést igen sok területen és különböző céloknak megfelelően alkalmazzák [S10].

1.1 A beszéd komplexitása, fiziológiai és fizikai alapjai

Egyedül az ember rendelkezik a beszélni tudás képességével. Az előzőekben megadott definíció alapján a beszéd hanghullámok révén közvetíti az információt. A beszédhez egyidejűleg három dolognak a megléte szükséges:

- nyelvi rendszer, ami meghatározza az akusztikai elemeket;
- a beszélő ember ép biológiai szervei a beszédképzéshez és a beszédjel felfogásához;
- végül a közeg, amely a hanghullámot továbbítja.

Ha a fentebb felsoroltak közül bármelyik is hiányzik vagy sérül, a beszéd információközvetítési hatásfoka csökken. A beszéd – mint akusztikai jel - komplex szerkezetét a fenti három elem határozza meg a beszélés minden pillanatában és ebből adódóan az emberi beszéd nem determinisztikus jellegű [42]. A nem determinisztikus jelleg a beszédünkben használt különböző kifejezések különféle kiejtései közötti variáltságot jelenti, azaz ha ugyanazon kifejezés különböző kiejtéseit rögzítjük, nem ugyanazt a hullámformát, digitalizálás után pedig nem ugyanazt a byte sorozatot fogjuk kapni. Ez a variáltság beszélők között és egy adott beszélő különböző kiejtései között is adott. A beszélők közötti variáltság a második pontból következik. Minden ember hangképző szervének a felépítése más, de vannak korcsoportokra, nemekre, dialektusokra jellemző vonások.

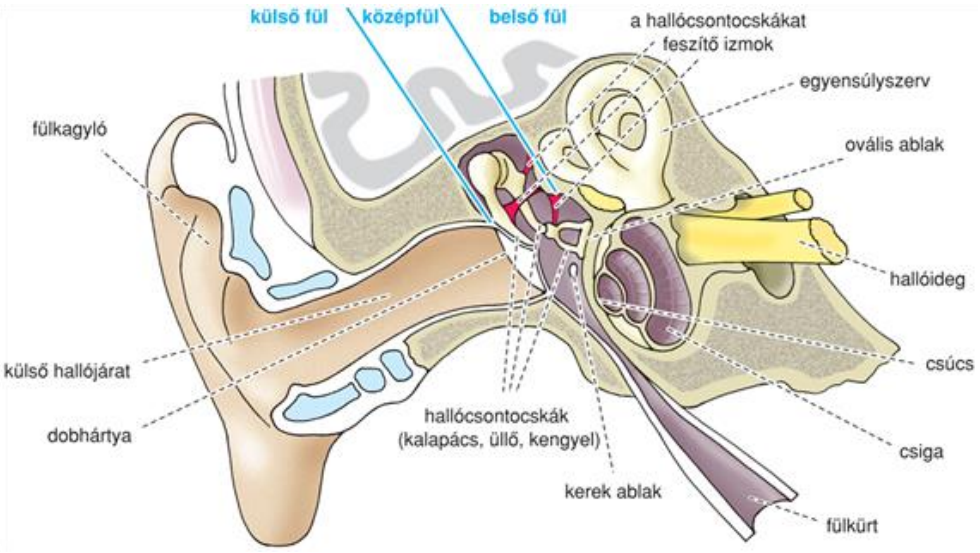
A beszédfelismerés kutatása során fontos a hangképzés vizsgálata is, mert következtethetni lehet arra, hogy mik a hangrezgések információtartalmát befolyásoló körülmények, illetve így válik igazán érthetővé, hogy egy adott gondolat kifejezése hány különféle hullámformában valósulhat meg.

1.1.1 A hallás folyamata

A levegő rezgéseit a fül bonyolult rendszere idegi jelzésekké transzformálja. A beszédfelismerés során ezt a transzformálást szeretnénk automatizálni. Ehhez nem csak azt kell érteni, hogy milyen jellemzői vannak a sugárzott beszédhang-hullámoknak, hanem azt is, hogy a vevő oldalán miképp működik a hang érzékelése és észlelése, hiszen az emberi hallórendszer komplex akusztikai, mechanikai, hidrodinamikai elektromos jelátalakító, idegvezetési és agyi szerkezet. Nemcsak számos ingerre reagál, hanem a beszédhangot és az alaphangot (hangmagasságot, hangfekvést), sőt, a hangforrás irányát is precízen beazonosítja. A hallási funkciók nagy részét a fül végzi el, de nagymértékben függ attól az adatfeldolgozástól is, amely a központi idegrendszerben történik [42].

Nem lehet azt mondani, hogy ilyen jellegű ismeretek nélkül nem lehetne működő beszédfelismerőt létrehozni, hiszen a csatornából vett akusztikai rezgések tartalmazzák azt az információt, ami elégséges a dekódoláshoz. Például a legelső, dinamikus idővetemítést használó megoldások nem vették figyelembe az emberi hallás sajátosságait. De a későbbiekben elterjedt rejtett Markov-modell alapú felismerők esetén fontos szerepet tölt be a sebesség, ugyanis az ilyen statisztikai alapú megoldások számításigényesek, és ezért a másodpercenkénti minták száma nagyban meghatározza a sebességet. De nem csak a fentebb említettek miatt, hanem általában, egyszerűbb és ésszerűbb a hangrezgéseknek csak azon komponenseivel foglalkozni, amelyek egy beszédet hallgató ember számára információt hordoznak. Ezeket a komponenseket pedig úgy lehet kinyerni, ha ismerjük, hogy az emberi fül mire érzékeny. Ezért a következőkben a fül felépítését részletezem kiemelve azon szerveket, melyek fontos szerepet játszanak a hallott hang átalakításában, szűrésében és ingerületté alakításában. Így világosabbá válnak a későbbiekben megfogalmazott elő-feldolgozási lépések.

Alapvetően három részre tagolható a fül, a működési funkciók függvényében: külső fül, középfül, belső fül, ahogy ez a 1. ábrán is látható.



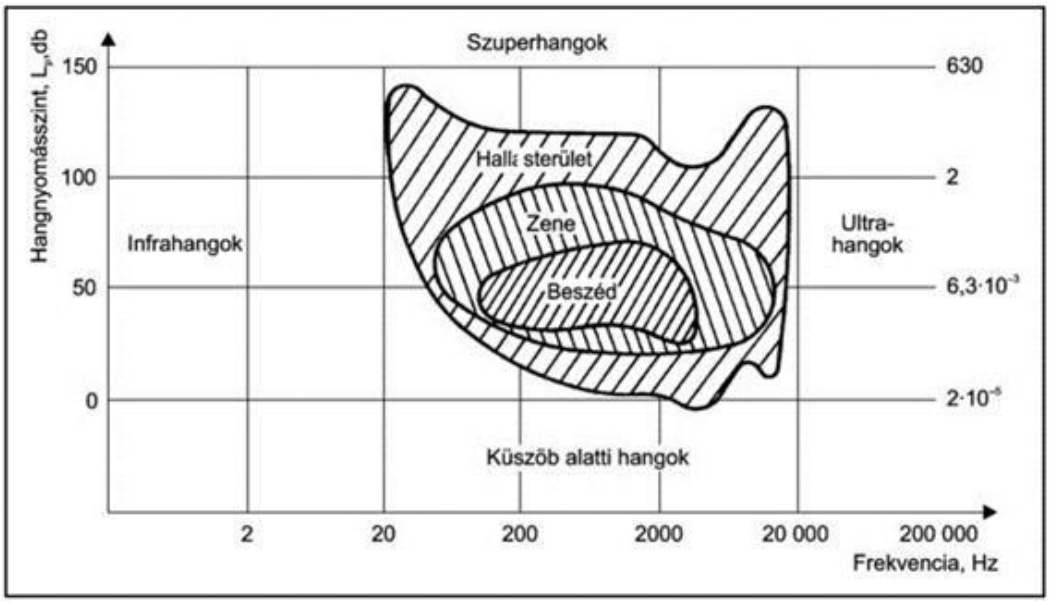
1. ábra *A fül vázlatos metszete*

A külső fül a fülkagylóból és a hallójáratból áll, amelyet a dobhártya zár le. Feladata a levegő rezgéseinek összegyűjtése, azok erősítése, és hozzájárul a hangforrások irányának meghatározásához, ezzel növelve tájékozódó képességünket.

A középfül a dobhártyából és a hallócsontokból áll (üllő, kalapács és kengyel). A dobhártya kör alakú, sugárirányú rostokból épül fel, a feszítőizmok pedig megfeszítve tartják. A bejövő hanghullámok hatására a dobhártya elmozdul a nyomásingadozás függvényében ez mechanikusan továbbítja a hangot, a belső fül a levegő rezgéseit folyadékrezgéssé alakítja. A nagyobb sűrűségű folyadék nagyobb érzékenységet tesz lehetővé. Ez által a középfül erősítőnek is tekinthető. A dobhártya fontos jellemzője, hogy 500 és 3500 Hz között veszteségmentesen visz át a hangot, a többi sávban pedig gyengíti azt [35].

A számítógépes beszédfelismerésnél a legfontosabb szerepe a belső fül vizsgálatának van. Ugyanis itt alakul a mechanikai rezgés ingerületté a belső fület az aggyal összekötő hallóidegeken. Ez határozza meg, hogy mi az, ami az agyban észleléssé válhat, mi az, amit jelentéssel tudunk felruházni. A külső fül iránykarakterisztikájától, és a dobhártya sávateresztő jellegétől eltekintve itt dől el az, hogy a hallott hang mely komponensei jutnak el a központi idegrendszerhez (1. ábra). A belső fülben található a csiga, amely a hallócsontokon keresztül kapcsolódik a dobhártyához, és felelős a hang dinamikai és frekvenciatartománybeli tulajdonságainak analizálásáért. A járatok kicsit, vagy egyáltalán nem működnek közre a hallásban [42].

A csiga a hossza mentén (3-4 cm) különféle frekvenciákra érzékeny. Ebből adódik az, hogy a hallás frekvenciatartományát 20 Hz és 20 kHz közé teszik a technológiai tervezés során (2. ábra).



2. ábra *A zene és a beszéd érzékelési tartománya a teljes hallási tartományon belül*

Mai napig elfogadott – és anatómiailag alátámasztott – elképzelés, hogy miképp alakul az észlelt rezgés frekvenciafüggő ingerületté. Ez az elgondolás egy magyar kutató nevéhez fűződik. Békésy György Nobel-díjas fizikus alkotta meg a vándorhullám-

elméletet, mellyel sikerült megmagyaráznia a frekvenciaérzékelés helyfüggését a csigában. Az elmélet szerint *„a kengyel bizonyos frekvenciájú mozgása az alaphártyán végigvándorló hullámmozgást vált ki, amely (…) az alaphártya anatómiai adottságaiból adódóan frekvenciától függően a membrán kezdeti vagy későbbi szakaszán okoz maximális elmozdulást”* [35].

Ahhoz, hogy kiemelhetővé tegyük a hang fontosabb paramétereit az elő-feldolgozás során, meg kell határozni, hogy a fül milyen fizikai paraméterekre milyen arányok szerint érzékeny. Az első megfontolandó paraméter a hang intenzitása. Korábban már említettem, hogy a fül körülbelül a 20 Hz-től 20 kHz-ig terjedő tartományban érzékeny az akusztikai hullámokra (2. ábra), ám az érzékenység egyáltalán nem azonos a különböző frekvenciákon. Számos kísérleti alanyon elvégzett kísérletek alapján írták fel a phon-skálát, ami a szubjektív hangosságérzet frekvenciafüggését adja meg [55]. Definíció szerint az 1000 Hz-es szinuszos hang hangosságértéke megegyezik a decibelben mért hangintenzitás értékkel, ettől eltérő frekvenciákon pedig a görbékről olvasható le a szubjektív érzet. A görbék az azonos hangosságúnak érzékelt intenzitásokat kötik össze. Körülbelül a 3-4 kHz frekvenciák környékén a legérzékenyebb a hallószervünk, és innen a kis és nagy frekvenciák felé haladva csökken az érzékenység. Ebből következik, hogy a hallásküszöb, azaz a legkisebb, ingert kiváltó intenzitás is 3-4 kHz környékén a legkisebb (10-13 dB).

A frekvenciaérzékelés a csiga mentén kb. 20 Hz-től 500 Hz-ig lineáris, felette logaritmikus. Kis frekvenciákon sokkal jobb a hallás felbontása, mint nagy frekvenciákon. A fenti tulajdonságokat megfontolva írható fel a Mel-frekvencia skála, melynek kiindulási alapja az 1 kHz-es, 40 dB intenzitású hang, melyhez az 1000 Mel értéket rendelték. Kis frekvenciákon kisebb a lépcső, amíg például 200 Hz-nek 283 Mel érték felel meg, és 400 Hz-nek 509 Mel (~1.8-as szorzó), addig 2000 Hz-nek 1520 Mel, 4000 Hz-nek pedig 2146 Mel (~1.41-es szorzó) [27].

Az akusztikai elő-feldolgozás során fontos a szerepük van a Fletcher által 1940-ben meghatározott kritikus sávoknak. Ha egy adott frekvenciájú szinuszos hangot hallunk, és vele egy időben szélessávú fehér zajt (itt a sáv nem annyira széles, hogy kiterjedjen az egész hallható tartományra, körülbelül 100-200 Hz-es sávokra kell gondolni), és a szinusz frekvenciája távol esik a zaj frekvenciatartományától, a két hangot függetlenül, tisztán elkülönülve érzékeljük. Azonban ha a frekvenciák fedik egymást - a hangintenzitások és a frekvenciaértékek függvényében - csökken a szinuszos hang hangosságérzete, és zaj akár teljes mértékben el is fedheti azt. Zajos környezetben nagyobb intenzitás szükséges hanginger kiváltásához, mint csendes környezetben (a phon-skála zajmentes környezetben mért értékeket mutat). A csiga felépítéséből adódóan a kritikus sávok a magasabb frekvenciák felé szélesebbek, aszimmetrikus jelleget mutatnak.

Ha a fülünket egyszerre több hang éri, és ezek egy kritikus sávba esnek, azok intenzitása összegződik, de a hallásunk nem tudja azokat megkülönböztetni. A kritikus sávok alapján íródott fel a bark-skála [58]. A skála a hallható tartományt (15500 Hz-ig) 24 sávra bontja, a sávok között átfedés van. A bark szám két hang közti

frekvenciakülönbséget megadó pszichoakusztikai jellemző. Azt mutatja meg, hogy a két vizsgált hang között hány kritikus sáv van. A bark-skála sávjai, a hallás frekvenciatartománybeli felbontásának nemlinearitása miatt, alacsony frekvenciákon keskenyebbek, 0 és 1600 Hz közé ugyanannyi kritikus sáv esik, mint 1600 és 16000 Hz közé. Érdekes pszichoakusztikai megfigyelés, hogy amíg például a hallásküszöb, és a hallható tartomány felső határa egyénről egyénre igen eltérő lehet, addig a kritikus sávok szinte mindenkinél ugyanolyan értékekkel adhatók meg.

Kiemelten fontos akusztikai elő-feldolgozási (lényegkiemelési) módszer a perceptuális lineáris predikció (PLP, Perceptual Linear Prediction). A módszer a kritikus sávok szerinti feldolgozást és a lineáris prediktorral való beszédkódolást ötvözi [24].

1.2 A nyelvi tudás szintjei

Nyelvtanilag helyes mondatok alkotásához, kimondásához és megértéséhez több szintű tudásra van szükségünk. Az artikulációs és percepciós bázis a kérdéses nyelv hangjaira vonatkozó agyi beidegződés. Az erre épülő következő szint a lexikai tudásbázis, ami a szavak ismeretét jelenti. E fölött a szintaktikai ismeretek állnak. Ez a halmaz az alapja, hogy értelmes mondatokat alkossunk. A szemantikai szint segítséget nyújt a szavak egyéni vagy együttes hatásukon keresztül megalkotott mondat jelentésének értelmezésében. Ez az összefüggés kiterjed a mondatok közötti térre is. A szemantikai szintű modellezés a mesterséges intelligencia kutatások egyik kedvelt területe. A legfelső szint a pragmatikai szint, azaz a kontextus (párbeszéd szereplőinek szándékai, a társadalmi környezet ismerete stb.) hozzájárulása a beszédhez, ami szintén fontos szerepet játszik az értelmezésben. Ezen szintek együttes ismerete nyelvi kompetenciánk részét képezik [42].

1.3 A beszéd szupraszegmentális szerkezete

A beszéd elméletileg szegmentális és szupraszegmentális részegységekre bontható. A szegmentális szerkezethez a beszéd létrehozásához szükséges alapvető komponensek tartoznak (beszédhangok és hangkapcsolódások szerkezeti elemei, a hozzájuk tartozó nyelvi időtartamok és arányok, a hangok egymáshoz viszonyított hangintenzitás-különbségei stb.). Ezek a beszédképzés alapvető elemei, így többnyire akaratunktól függetlenül jönnek létre. Ezen szerkezeti elemek felhasználásával már érthető beszéd hozható létre. A szupraszegmentális jellemzők alapvető tulajdonságai a beszédnek, amelyek hozzájárulnak a beszédmegértéshez. A szupraszegmentális tulajdonságok összességét prozódiának szokták nevezni, melynek megvalósítása inkább akaratfüggő [42]. A szupraszegmentális tényezőkre megfogalmazott definíciók közül Markó adta meg talán a legszabatosabb megfogalmazást [38]: *„a szupraszegmentális szerkezet a beszédprodukciós folyamat által létrehozott komplex beszédjelnek az a vetülete, amely az idő, a frekvencia és az intenzitás folyamatváltozásaiként írható le, és amelynek észlelése állandó viszonyításban lehetséges”*.

Ezek a tulajdonságok nemcsak a mesterséges beszéd természetességét javítják, hanem a gépi beszédfelismerés hatékonyságát is. Segítségével a beszélő érzelmeket, szintaktikai és pragmatikai információt stb. fejezhet ki [22]. Szupraszegmentális jellemzők:

- a beszédtempó;
- a szünet;
- a ritmus;
- a hangszínezet;
- a hangerő;
- a hanglejtés;
- és a hangsúly.

1.4 A halláskárosodás definíciója, mértéke és kihatása a beszéd elsajátítására

A siketek állapotát évszázadokig a hallás hiánya felől közelítették meg, s jobb esetben fogyatékosoknak, mintegy betegeknek könyvelték el őket, rosszabb esetben pedig emellett értelmi képességeiket is megkérdőjelezték. (A kialakult negatív vélemény azt a felfogást tükrözte, hogy – miután a nyelvet a hangos beszéddel azonosították – a beszéd hiánya a gondolkodás hiányát jelenti.) A siketoktatás 18. századi kezdeteit követően az utóbbi nézet jelentősen visszaszorult. Napjainkban azonban annak a szemléletnek is változnia kell, amely a hallás hiányára, illetve korlátozott voltára – tehát végső soron a fogyatékosságra – koncentrál. Ez az ún. orvosi-gyógypedagógiai szemlélet át kell, hogy adja helyét az antropológiai szemléletnek, amely az embert és fennálló képességeit tekinti mérvadónak, s értékeli például a jelnyelvi kommunikációt [29].

1.4.1 A halláskárosodás meghatározása és mértéke

A hallássérülés tág fogalom, biológiai, orvosi és pedagógiai szempontból eltérő értelmezéssel és kategorizálással definiálható. Biológiai és orvosi szempontból ide sorolandók a hallószerv bármely részének veleszületett vagy szerzett sérülései, esetleg fejlődési rendellenességei, melyek eredményeképpen a hallásteljesítmény az éptől eltér [46]. Gyógypedagógiai szempontból a hallási fogyatékosság zártabb fogalom, olyan hallási rendellenességet jelent, ahol a sérülés időpontja, mértéke és minősége miatt a beszédbeli kommunikáció spontán kialakulása zavart [10].

„A hallássérülés gyógypedagógiai fogalma (hallási fogyatékosság) elsősorban a beszédértéshez szükséges hallásterületen közepes vagy annál súlyosabb fokú nagyothallást, siketséggel határos vagy siketségnek diagnosztizált hallásveszteséget jelent. Más megközelítésben a hallássérültek pedagógiája a hallássérült kifejezést olyan hallászcsökkenésre alkalmazza, amelynek következményeként a beszédfejlődés nem indul meg, vagy a beszéd oly mértékben sérült, hogy a beszéd megindításához, korrekciójához speciális beszédfejlesztő módszerek alkalmazására van szükség” [13].

A hallássérülésnek két fő oka lehet. Egyik az örökletes hallássérülés, ami általában mindkét oldalt érintő károsodás (egy oldalt érintő károsodás az esetek kb. 10%-ban fordul elő). A károsodás lehet domináns vagy recesszív jellegű, lehet vezetéses, idegi

vagy kevert típusú. Kisebb százalékban egyéb tünetekkel is járhat (látássérülés, központi idegrendszeri zavar stb.). A másik oka/típusa a szerzett hallássérülés, ami lehet méhen belüli károsodás (dohányzás, alkoholfogyasztás, fertőzések stb.), okozhatja koraszülés, fertőző betegség, ototoxikus antibiotikum vagy zajártalom [23].

A halláskárosodásnak több szintjét különböztetjük meg:
- teljes hallásvesztés (süket), a hallást nem lehet kimutatni;
- süketséggel határos halláscsökkenés nagyobb, mint 90 dB;
- súlyos halláscsökkenés: 81-90 dB;
- nagyfokú halláscsökkenés: 61-80 dB;
- közepes halláscsökkenés: 41-60 dB;
- kisfokú halláscsökkenés: 26-40 dB;
- nem jelentős halláscsökkenés: 25 dB-ig.

Megnövekedett azoknak a cochleáris implantáció, hallásjavító műtéten átesett hallássérülteknek a száma, akik beültetett készülékkel megközelítőleg vagy teljesen elérik a hallás normál övezetét. Ezért manapság a hallássérültek 3 fő csoportját különítjük el [13]:
- siketek;
- nagyothallók;
- cochleáris implantáltak.

1.4.2 A halláskárosodás kihatása a beszéd elsajátítására

A hallássérülés közvetlen következménye a beszéd elsajátításának zavara vagy akár a beszéd teljes hiánya. A kifejező beszéd zavarai az alábbi területeken és módon fordulnak elő [10]:
- az artikuláció (a helyes kiejtés) és a szupraszegmentális elemek hibáiban;
- a szókincs hiányosságaiban;
- olvasási nehézségekben;
- a grammatikai hibákban;
- pragmatikai és szintaktikai pontatlanságokban.

A hallás fontossága nem csak beszédkommunikációs szempontból hangsúlyozandó. A felsorolás teljes körű leírást ad a hallás jelentőségéről [11], [12], [17], [26]:
- minden irányból közvetít;
- távolabbi eseményekről is közvetít;
- permanens ingerközvetítő;
- irányítja a vizuális észlelést;
- kíváncsiságot kelt;
- előkészít bekövetkező eseményekre;
- beszéd-belső beszéd befolyásolja a magatartást;
- hangulatokat közvetít;
- kapcsolatfelvétel-kapcsolattartást tesz lehetővé.

Az alábbi felsorolás a hallás részleges vagy teljes hiányának következményeire világít rá [11], [12], [17], [26]:

- a valóság szaggatott (mozaikszerű) információkból áll össze;
- hiányoznak a távolabbról érkező, figyelmet irányító információk;
- a személyiség merevebbé válik;
- társadalmi szokások, magatartási szabályzók hiányai;
- kapcsolatok beszűkülése;
- gondolkodás-, viselkedés- és személyiségbeli változások állnak be.

A bevezetőben megnevezett kutatási projekt és az azon belül végzett kutatásaim szempontjából kiemelt jelentőségű az artikuláció és a szupraszegmentumok (1.3 fejezet). Ezek tekintetében kiemelten fontos megjegyezni, hogy azok a beszédhangok alakulnak ki késve, vagy hibásan, amelyeket a hallássérült (gyermek) akusztikusan nem jól érzékel. Mivel a helytelen képzésről nincs, vagy gyenge a visszacsatolás, ezért a súlyos fokban hallássérültek általában nem tudják kijavítani azokat külső segítség nélkül. A projektben résztvevő szurdopedagógusok tapasztalatai szerint elmondható, hogy leggyakrabban a magas frekvencián hallható sziszegő hangok, a *ty-gy* zárhangok, valamint a *c-cs* affrikáták képzése hibás, de gyakori az ó-ő és az *ú-ű, ű-ő* hangok cseréje is. Általánosságban elmondható, hogy minél nagyobb a hallásveszteség, annál több beszédhangot érint a hibás ejtés [18]. A hallássérültek beszédére szintén jellemző a zöngés-zöngétlen zárhangok megkülönböztető képességének hiánya, helytelen használata vagy cseréje. Általában a hallásveszteség súlyosságával egyenes arányú a zöngétlen ejtés gyakorisága. A zöngés ejtéshibáknak a korrigálását is az auditív kontroll zavara nehezíti, bár ritkább jelenség [10].

A beszéd érthetőségét, minőségét elsősorban a szupraszegmentális tulajdonságok befolyásolják (1.3 fejezet). A beszélt nyelv „szegmentumai”, a magánhangzók és a mássalhangzók, melyek szótagokká, szavakká és mondatokká állnak össze, így alkotják a beszéd verbális aspektusát (1.2 fejezet). Miközben azonban ezeket a szegmentumokat képezzük, kiejtésünk egy más szempontból nézve változó. Eltérő magasságú hangok széles sorát használjuk, ami különböző módon változtatja meg a mondottak jelentését. Ezek azok a hanghatások, amelyek a szupraszegmentális elemzés számára adatokat szolgáltatnak. A hang magasságát és erősségét a sebesség és a ritmus megkülönböztető használatával együtt a nyelv prozódiai jegyeinek hívjuk [8].

A szupraszegmentális hiányosságok kiterjednek a normálistól eltérő tempóra és ritmusra, a hanglejtés változtatásának problémájára, a hangmagasság korlátozott terjedelmére, és a „siket hangminőségre”, ami mind a beszélt nyelv olyan aspektusa, amelyet a hallás, az auditív visszacsatolás és az önirányítás révén lehet megszerezni.

1.4.3 Hallássérültek oktatása

A jelnyelv az utóbbi évtizedek kutatásainak köszönhetően megtalálta helyét a nyelvek nagy családjában. A jelnyelveket – önálló nyelvekként – a nyelvi kisebbségek nyelvhasználatához kapcsoljuk, a siket közösségek évszázadok során kialakított kultúráját pedig az interkulturális kommunikációhoz. A jelnyelveknek van egy igazán különleges jellemzőjük az, hogy nem az auditív-akusztikus, hanem a vizuális-gesztikuláris modalitású nyelvek közé tartoznak. Ez azt jelenti, hogy – mivel a siketek nem hallanak – az auditív helyett a vizuális csatornát kell kommunikációs csatornának

kialakítaniuk. Ezen az üzenetek a két kéz mozgásai – gesztusai – révén érkeznek. A kommunikációs partner ugyanígy, a „kezek nyelvén” válaszol. Ezt a nyelvet jelnyelvnek, használatát jelelésnek nevezzük. Mint bármely más kisebb létszámú népcsoport, törzs tagjai, a siketek is nyelvük és a rá épülő kultúra révén különülnek el a többségi nyelvhasználóktól.

A hazai siket közösséget napjainkban leginkább megmozgató kérdések között az ún. bilingvális oktatás 2017. szeptembertől várható bevezetése áll az első helyen. Magyarországon a Jelnyelvi törvény 2009-es elfogadása óta eltelt fél évtized a jelnyelv elismertségét illetően hozott sikereket. Pedig éppen a bilingvális oktatás bevezetése azért is várat még most is magára, mert – szemben például a tévéműsorok feliratozásával vagy a tolmácsszolgálat kibővítésével – a leghosszabb előkészítő munkát kívánja. Hogy csak kettőt említsek a feladatok közül: tanárokat – halló és siket tanárokat is – kell felkészíteni mind magának a jelnyelvnek az oktatására, mind pedig tantárgyaknak jelnyelven történő tanítására. A bilingvális oktatás elsődlegesen a jelnyelv iskolai használatát kívánja megvalósítani, amelytől sokan remélik, hogy a siket gyerekek tudásának a színvonala emelkedni fog. A külföldi példák nyomán ebben bízhatunk is, ld. pl. Kárpáti Árpád leírását a Svédországban tapasztaltakról [31]. A bilingvális oktatás bevezetésének, mint fontos célnak az elérésével a másik nyelvnek, azaz a magyarnak az ismerete sem szorulhat háttérbe. A hazai siket közösség tagjainak a jövőben is tudniuk kell magyarul is kommunikálni. A bilingvális oktatásról olvasható a törvényben: *„A bilingvális oktatási módszer: olyan oktatási módszer, amely a beszélt magyar nyelv mellett a magyar jelnyelvet is alkalmazza az oktatás során”* [S3].

2 Felhasznált technológiák

2.1 Rejtett Markov-modellek alkalmazása a beszédfeldolgozásban

A Markov-lánc megalkotása a 20. század elejére tehető, amely Andrej Markov orosz matematikus nevéhez fűződik. Az első eredményeket 1906-ban e folyamatok tekintetében kizárólag elméleti szinten fektette le. A Markov-lánc diszkrét sztochasztikus folyamatot ír le. Az a fogalom, hogy valami Markov-tulajdonságú azt jelenti röviden, hogy adott jelenbeli állapot mellett, a rendszer jövőbeni állapota nem függ a múltbeliektől. Másképpen megfogalmazva, ez azt is jelenti, hogy a jelen leírása teljesen magába foglalja az összes olyan információt, ami befolyásolhatja a jövőbeli helyzetét a folyamatnak [57].

A "rejtett Markov-modell" kifejezésben a "rejtett" jelző arra utal, hogy mi csak a modell működésének az eredményét, a kimenetet (azaz a generált szekvenciát) ismerhetjük, a modell maga és a paraméterei számunkra ismeretlenek. Így mi csak a kimenetből következtethetünk a modell felépítésére és a működését leíró paraméterekre (az átmeneti és a kibocsátási valószínűségekre) [57].

A szótár minden egyes eleméhez tanulással – approximációs eljárással – létre kell hozni egy-egy Markov-modellt, majd a felismerés során a kiejtett (felismerendő) elemhez ki kell számítani minden modell esetén azt a valószínűséget, amilyen valószínűséggel a modell a felismerendő elemet ilyen kiejtéssel generálhatta. Ha ezek között a valószínűségek között van pontosan egy kiemelkedő, akkor a felismerés sikeres, és a kiemelkedő valószínűséghez tartozó szótári elem lesz az eredmény. (A rejtett Markov-modell érzékeny a túltanulásra.) Tehát az ilyen modellekre épülő beszédfelismerés tisztán statisztikai alapú. A HMM előnye, hogy elég egyszerűen kiterjeszthető nagyszótáras, folyamatos beszéd felismerésére, viszont ebben az esetben célszerűbb kisebb egységekből építkezni (triádokból, diádokból, hangokból). Ezek összekapcsolásából kaphatjuk meg a szavak modelljeit (az összekapcsolások általában valamilyen nyelvtani szabályrendszer alapján történnek), majd végül ezeket körbekapcsolva egyetlen nagy modellt is kaphatunk.

Az alábbi képletek a rejtett Markov-modellekkel való beszédfelismerési feladatok megoldását mutatják be.

$$w_r = \text{argmax}_{w \in W}\{P(w|X)\} \tag{1}$$

Vagyis azt a w_r szótári elemet (szó, hang, diád stb.) keressük, amelyre az X adott akusztikai megfigyelés-sorozat valószínűsége a legnagyobb. Mivel az X megfigyelés-sorozat számunkra ismert, ezért Bayes tételét alkalmazva:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)} \tag{2}$$

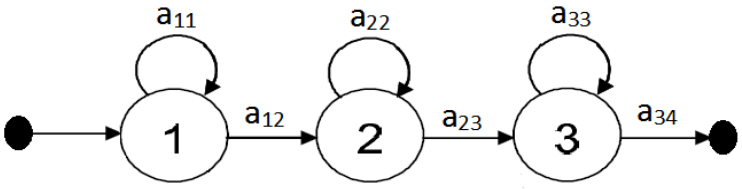
átalakíthatjuk az (1) összefüggést és az alábbiak szerint írhatjuk fel [57]:

$$w_r = \text{argmax}_{w \in W}\{P(X|w) * P(w)\} \quad (3)$$

Ebben az alakban a $P(X)$ tag elhagyható a nevezőből. A $P(X|w)$ valószínűséget az akusztikai, a $P(w)$ valószínűséget pedig a nyelvi modell határozza meg. Az akusztikai modell alapján arról informálódhatunk, hogy az adott akusztikai megfigyelés (felismerendő minta) az egyes szavakra (szótári elemekre) milyen valószínűségű, a nyelvi modell pedig az egyes szavak előfordulásának becsült valószínűségét szolgáltatja.

Beszédhangok esetén általában három állapotú lineáris struktúrájú modellt (ún. balról – jobbra) szokás választani (3. ábra). Magát a modellezést a diádok (olyan fonémakapcsolatok, amelyek az első hang felétől a második hang feléig tart) esetén szintén három állapot végzi, valójában azonban két további szélső állapotot is találunk, amelyek az egyes beszédelem-modellek összefűzését biztosítják.

Felismeréskor a rendszer számára minden keret érkezésekor két lehetőség áll fent, vagy állapotot változtat, vagy helyben marad, bizonyos valószínűséggel. Ezeket nevezzük állapotátmeneti valószínűségeknek, melyek becslése a tanítás során történik. Ez a mechanizmus biztosítja az időbeli illesztést a modell és az aktuális keret között. A rendszer az adott (belső) állapotból két keret érkezése között egy megfigyelést bocsát ki, mely tulajdonképpen egy hasonlósági mérték az adott állapotra jellemző jellemzővektor-eloszlás és az aktuálisan érkezett, a külső megfigyelést reprezentáló jellemzővektor között. Lényegében azt mondhatjuk, hogy e hasonlósági mérték a mérőszáma a megfigyelt jellemzővektor és a modellállapot spektrális illeszkedésének. Egy állapotra jellemző jellemzővektor-eloszlást általában sűrűségfüggvényével adunk meg, amelyről feltételezzük, hogy normális (Gauss) eloszlások lineáris kombinációjából áll elő. Ezt szokás kibocsátási valószínűségnek is nevezni [46].



3. ábra *3 állapotú lineáris modell*

A munkám során felhasznált HMM modellekre optimalizálást végeztem, így a végső modellek, amiket alkalmaztam, hang alapú felismerésre készültek, 7 állapotúak, nem tartalmaznak Gauss eloszlást és PLP együtthatókat alkalmaztam a betanításukhoz 10 ms-os időkerettel, kiegészítve az energia komponenssel.

2.2 Mesterséges neurális hálózat alkalmazása a beszédfeldolgozásban

Napjainkban a rejtett Markov-modelles megoldások mellett a mesterséges neurális hálózatok (Artificial Neural Network, ANN) is népszerű megoldások a beszédfeldolgozásban. A mesterséges neurális hálózatok valójában, nem lineáris elemeket tartalmazó numerikus eljárások, egyaránt alkalmazhatóak szimbolikus és nem

szimbolikus problémák megoldására is [20]. A neurális hálózatok olyan számítási modellek, melyek előnyei abból erednek, hogy jól párhuzamosíthatóak, sokféle tanítási algoritmus létezik hozzájuk, és a nem lineáris elemek bevezetése után a csak lineáris elemeket tartalmazó modellekhez képest új, bonyolultabb problémák megoldására is alkalmazhatóak [5]. A hálózatok leginkább a következő problématerületeken jelentenek kiemelkedő megoldási alternatívákat [20]:

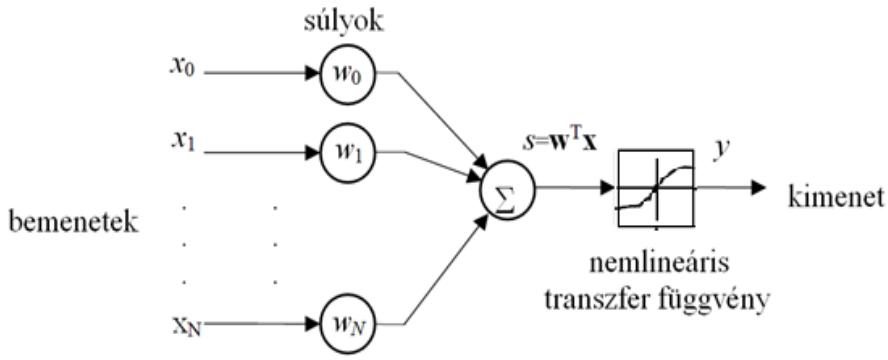
- mintázat-felismerés;
- függvényapproximáció;
- független komponensekre való bontás;
- operációkutatás;
- intelligens irányítások;
- inverz problémamegoldás (amikor nem a bementekből és a rendszer átviteli tulajdonságaiból kívánjuk a kimenetet meghatározni, hanem ismert kimenetek és rendszer alapján a bemenetet, vagy ismert kimenetek és bemenet alapján a rendszert identifikálni) [5].

A beszédfelismerésben a neurális hálózatokat, mint mintázatfelismerőket alkalmazzák, hiszen a beszédhez generált spektogramban vagy lényegkiemelés révén létrejövő jellemző (feature) vektorsorozatokban az egyes fonémákra vagy akár beszédakusztikai osztályokra jellemző mintázatok alapján kategorizálhatók a feldolgozási egységek.

2.3 A mesterséges neurális hálózatok architektúrája

A mesterséges neurális hálózatok megalkotását az idegrendszer, pontosabban az emberi agy felépítése ihlette, ám a mesterséges neurális hálózatok nem feleltethetőek meg egyértelműen az agyi idegsejtek és kapcsolati hálózatuk számítástechnikai reprezentációjának. A mesterséges neurális hálózat leginkább egy számításelméleti fogalomnak tekinthető. Gondoljunk csak a McCulloch-Pitts neuronra [20]: ez olyan neuron, amely bár nagy vonalakban hasonlít a biológiai neuronokra, de annyira absztrakt a struktúrája, hogy az valójában nem vesz figyelembe sok biológiai szempontot. A neurális hálók tehát nem foghatóak fel az emberi gondolkodás modelljeinek, sok olyan tényezővel nem foglakoznak, melyek egy biológiai rendszerben elsőrendűek. Például egy neuron esetén a kimenet aktív állapota rengeteg biofizikai kisülést, impulzust jelent, azonban mesterséges neurális hálózatoknál sokszor csak egy olyan állapotnak felel meg, mint egy tranzisztor állapota vagy egy logikai változó igaz vagy hamis értéke. A mesterséges neurális hálózatok valójában az agyi neurális hálózatok struktúráját követve igyekeznek a számítástechnikában kihasználni azok előnyeit: a gyorsaságot, a párhuzamos feldolgozást, a robosztusságot, a zajtűrést, az alkalmazkodást újszerű környezetekhez, a hiányos, vagy hibás bementek rugalmas kezelését, valamint a kis energiaigényt.

A mesterséges neurális hálózatokban, hasonlóan a biológiai hálózatokhoz, az általános felépítési egység a neuron. A mesterséges neuronok tulajdonképpen egy függvény hozzárendelést valósítanak meg. Adott *n* db bemenethez, egyetlen kimentet rendelnek. További paramétereik az ún. aktivációs függvény, tüzelési küszöb, valamint az egyes bementekhez rendelt súlyérték. Az 4. ábra egy általános neuron modelljét mutatja [2].



4. ábra *Az általános neuronmodell*

Az ábrán a következő neuron paramétereket láthatjuk:

- $x = [x_1, \dots, x_n]$a bemeneteket tartalmazó vektor;
- minden x_k bemenethez hozzárendel a modell egy W_k súlyt, azaz a neuronhoz tartozik egy $W = [W_1, \dots, W_n]$súlyvektor;
- Σ összegző függvény;
- θ tüzelési küszöbérték, egy aktivációs függvény (a 4. ábrán nincs külön szimbólummal ábrázolva, sok helyen $a(\ )$-val jelölik);
- valamint egyetlen y kimeneti érték.

A neuron működése a következő módon írható le. Minden *t* időpillanatban a Σ függvény szerint összegzi a bemenetek súlyozott értékeit. Jelölje R az összegző függvény kimenetét:

$$\mathrm{R} = \sum x_i * W_i \tag{4}$$

mátrixos alakban:

$$R = WT * X \tag{5}$$

Ez az R érték kerül majd az $a(\ )$ aktivációs függvény bemenetére, a θ tüzelési küszöbértékkel eltolva, azaz a neuron kimenete a következő alakban írható föl:

$$y = a(R - \theta) \tag{6}$$

Az $a(\ )$ összegző függvény többféle matematikai függvény lehet, a legelterjedtebbek:

- lineáris függvény;
- ugrásfüggvény;
- szigmoid függvény;
- valamint a tangens-hiperbolikusz függvény [5].

A neurális hálózatok paramétereit nem optimalizáltam, mivel a betanítás a nagyméretű adatbázison több napot vesz igénybe. A fonetikai osztályok és a hangok felismeréséhez feedforward neurális hálózatot választottam one-step secant backpropagation tanítási algoritmussal. A hangsúlyozáshoz a patternet neurális hálózatot választottam (legfőbb eltérése a feedforward hálózattal szemben, hogy a kimeneti rétege tangens szigmoid átviteli függvényt tartalmaz a lineárissal szemben). A tanítási algoritmusnak a resilient backpropagationt választottam. A neurális hálózatok tanítását és tesztelését a MATLAB szoftverrel valósítottam meg.

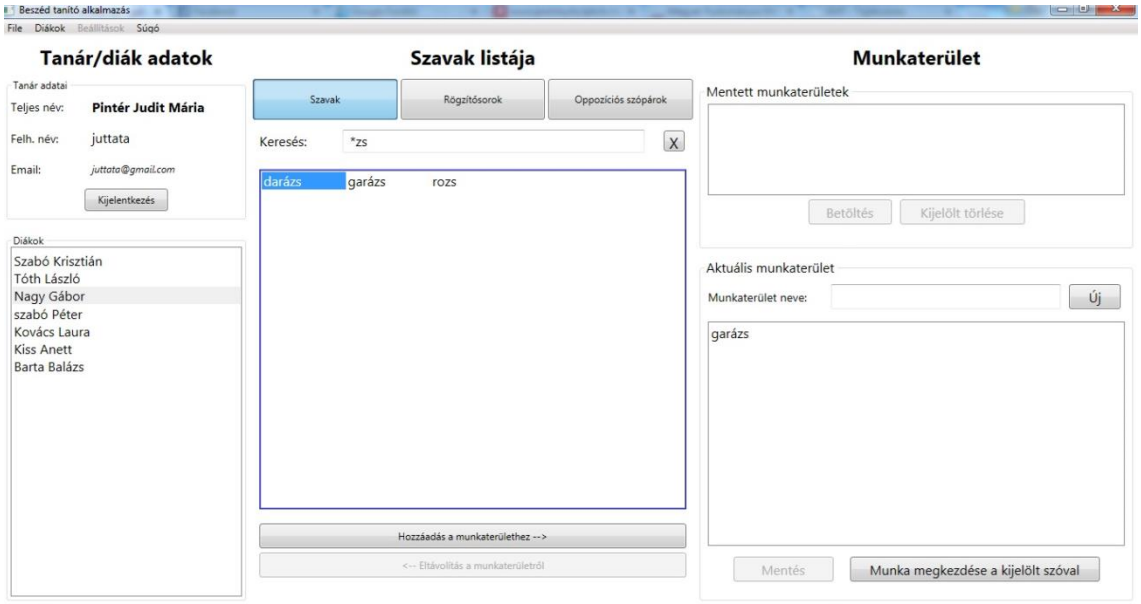
3 Kutatási projekt hallássérültek internetes beszédfejlesztésére

Az „Alap- és alkalmazott kutatások hallássérültek internetes beszédfejlesztésére és az előrehaladás objektív mérésére” címet viselő projekt a siket és nagyothalló személyek számára – az eddigi eszköztár részbeni megújításával – a sikeres beszédtanulás egyik kulcsát nyújthatja. A projekt gyakorlatban is hasznosítható célja egy komplex rendszer létrehozása, mely a beszédfolyamat audiovizuális megjelenítését szolgáltatja, egyrészről a beszéd hangképeinek másrészről az artikulációnak a vizuális megjelenítésével, egy oktatási keretrendszerbe foglalva. Ezek mellett számos olyan funkciót tartalmaz a rendszer (prozódia megjelenítés, automatikus minősítés, tudásalapú rendszer implementálása), amely a későbbiekben lehetővé teszi az egyéni gyakorlást nem csak számítógépen, hanem mobil eszközön is. A kifejleszteni kívánt technológia audiovizuális transzkódolását végző modulja nyelvfüggetlen, a beszélő fej és az automatikus minősítés újabb neurális hálók betanításával nyelvfüggetlenné tehető.

3.1 A beszédasszisztens koncepció

A rendszer tesztelése 2013. szeptemberben kezdődött 14 szurdopedagógus (nagyothallókkal és siketekkel foglalkozó pedagógus) részvételével és eltérő korosztályú és fejlettségi szinten álló gyerekekkel. A rendszer felhasználásának módszertana folyamatosan változik és bővül, mivel a pedagógusok szabad kezet kaptak a keretrendszer tanórán belüli alkalmazására, hogy tapasztalatokat gyűjtsenek és ajánlást tegyenek a rendszer továbbfejlesztésére. A továbbiakban a jelenleg használt rendszer főbb komponenseit és azok működését mutatom be.

A rendszer használatának feltétele a regisztráció, ami csak alap adatok megadásával jár, és ezek után teljes körű felhasználhatóságot kapunk a beszédasszisztenshez. Bejelentkezés után a szurdopedagógusoknak lehetőségük van, hogy kiválasszák az 5. ábrán látható kezdőfelületen, kivel szeretnének foglalkozni és milyen szavakat akarnak gyakoroltatni. (A szóadatbázis összetételét a későbbiekben fogom részletezni.) A kezdőfelületen keresztül lehetséges újabb diákok felvétele egyedi azonosítóval és megjegyzésekkel, valamint régebbi diákok törlése a rendszerből. Már korábban elmentett munkaterületek is újból meghívhatók. Ha kiválasztottuk a diákot, akivel gyakorolni szeretnénk, és a gyakorlandó szót, csak ekkor fog a beszédasszisztens továbblépni a gyakorlófelületre (7. ábra). A gyakorlás során létrejövő minták automatikusan feltöltődnek és tárolásra kerülnek a szerveren hallgatókhoz dedikáltan későbbi vizsgálatok és kutatások elvégzésének céljából. Hallgatók törlésekor a minták nem törlődnek automatikusan a szerveren, hanem csak a beszédasszisztens rendszeren belül kerülnek eltávolításra a hallgatók listájából.

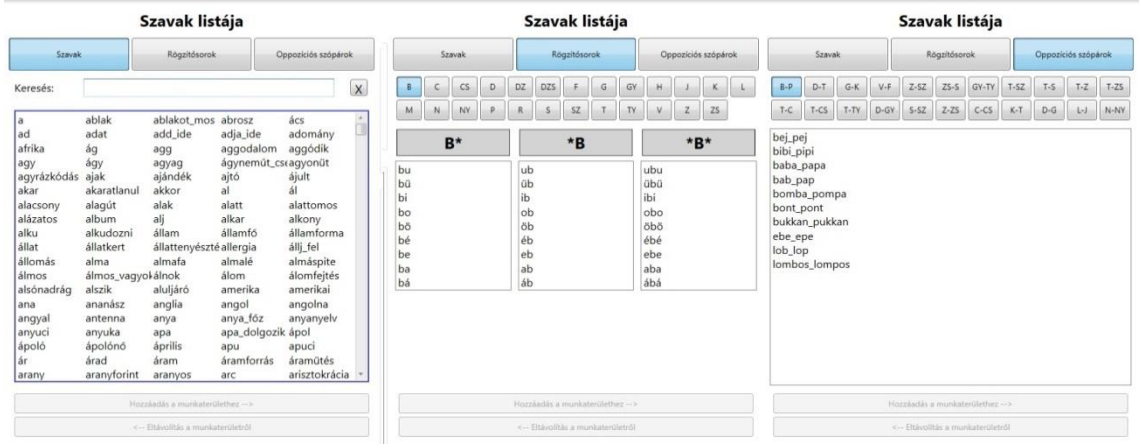


5. ábra *A beszédasszisztens rendszer kezdőfelülete*

3.1.1 Referencia beszédadatbázis

A rendszer a szurdopedagógusok tapasztalati és ajánlásai alapján 3 fő részből összetevődő szóadatbázist tartalmaz (6. ábra). A kezdeti fő adatbázist, ami 3000 szóból tevődik össze, a Miskolci Egyetem egyik kutatócsoportjának tagjai rendszerezték több szempont alapján (szófaji besorolás; témaköri besorolás; szótagszám alapján; hangok száma alapján; magánhangzó – mássalhangzó képlet alapján stb.). A pedagógusok tapasztalatai a rendszer használatával eredményezték a szókészlet bővítésének igényét az alábbi két főcsoporttal:

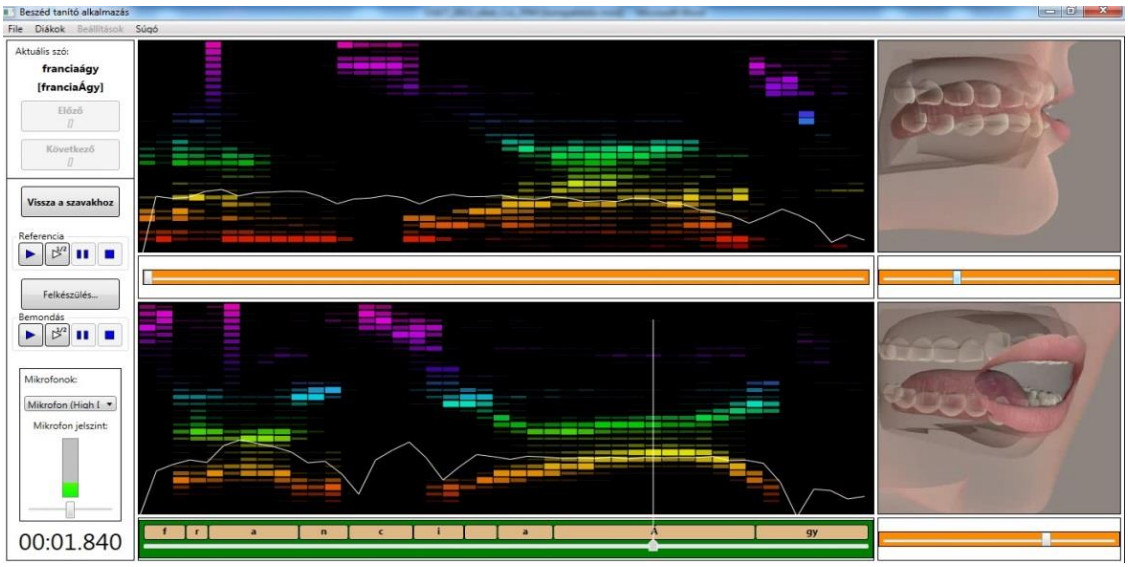
- rögzítősorok, amelyek a hangkapcsolatok begyakorlását segítik elő, valamint a mássalhangzók szó eleji, szóvégi vagy intervokális pozícióját;
- oppozíciós szópárok, amik csak egy hangban eltérő szókapcsolatok és különböző szembenállások gyakorlását segítik elő.



6. ábra *A referencia beszédadatbázis összetétele*

3.1.2 Audiovizuális transzkódolás

Az audiovizuális transzkódolás lényege a hangok absztrakt képi jelekké alakítása a frekvencia-összetevők és a hangerősség szerint kódolt formában. A grafikus szimbólumok oszlopok, amelyek mérete sávenergia-függő módon kisebb-nagyobb. A frekvencia-összetevők hangmagasságtól függően szín kódoltak: a mély hangok a nagyobb hullámhosszúságú színekkel (vörös) a magas hangok a rövid hullámhosszúságú színekkel (kék) jelennek meg. A beszéd szempontjából fontos frekvenciatartomány a 125 Hz – 8000 Hz [9]. Ezt 30 fix frekvenciasávra bontjuk, amelyek oszlopos formában kerülnek megjelenítésre. A beszéd frekvencia komponense tehát nem csak színkódolt, hanem pozíciókódolt is. Ez már nyújt akkora hangképi különbözőséget a szóképek között, hogy az egymáshoz hasonló hangalakú szavak megkülönböztetése lehetővé váljék. A 7. ábrán látható az audiovizuális megjelenítés, az alsó kép a referencia bemondást, míg a felső a saját aktuális bemondást reprezentálja.



7. ábra *A beszédasszisztens rendszer gyakorló felülete*

3.1.3 Tanulás és gyakorlás a beszélő fejjel

A 7. ábra jobb oldali részén látható a beszélő fej transzparens arccal két eltérő szögben. Hallássérültek beszélni tanítását támogató rendszerben az artikuláció bemutatása, a vizuális modalitás a szájmozgás, a nyelv és fogak láthatósága mellett egyéb jellemzők hozzáadása segíti a beszéd megértését [6], [40].

A siket emberek a hallókkal való kommunikációjuk során csak az ún. szájról olvasásra hagyatkozhatnak. Csakhogy míg a beszéd során 39 fonémát különböztet meg anyanyelvünk használója, addig a beszédhangokra vonatkozó, a látószervvel is érzékelhető eltérés a legjobb esetben is csupán 15 sorakoztatható fel, amiket vizémáknak nevezünk. A siket emberek számára nem ismerhető fel az azonos zártsági fokkal képzett veláris és palatális magánhangzópárok tagjainak különbsége. Így főként az ún. félig zárt és a zárt magánhangzók, tehát az *o–ö (ó–ő)*, illetve az *u–ü (ú–ű)* hangpár esetében nem megkülönböztethető a magánhangzók képzése [S3]. Vagyis nem tudni, hogy a megszólaló a *torok* vagy a *török*, a *szó* vagy a *sző*, a *fut* vagy a *fűt*, a *túr*

vagy a *tűr* szavakat ejtette-e. A magánhangzó-hosszúságnak sincsenek egyértelmű jelei az ajkak mozgását illetően, ugyanis, ahogy a szakirodalom fogalmazza: *„Az időtartamban vagy intenzitásban eltérő hangok is azonos artikulációs mozgásokkal jelennek meg az arcon”* [42]. Emiatt is kérdéses, hogy melyik szót mondta a beszédpartner. Szemünk elől rejtve marad a mássalhangzó-képzés jó része is. Támpontot egyedül a társalgás témaköre jelenthet.

A szájüreg tehát a hangképzés során valóságos „fekete dobozként” viselkedik a szemlélő számára, hiszen nyelvünk különböző irányú elmozdulásait – így a magánhangzó-képzést kísérő horizontális és vertikális mozgásokat, a nyelvnek a mássalhangzó-képzés során felvett kiinduló vagy végső helyzetét – nincs módunk megfigyelni. De a lágy szájpadlás és a nyelvcsap végezte apróbb mozgásokat sem látjuk (sőt az esetek igen nagy részében nem is tudatosítjuk). (Már nem is szólva a gégében lévő hangszalagok működéséről.) Az ép hallású embereknek ezek egyike sem okoz gondot, annál nagyobb problémaként jelentkezik az auditív csatorna működését nélkülözni kénytelen személyek, így elsősorban a siketek és a súlyosan nagyothallók számára. A beszélő fej transzparens, azaz átlátszó arca révén azonban tanúi lehetünk a nyelv sokféle mozgásának. A beszélő fejnek természetesen vannak korlátai a beszéd láthatóvá tételében: például nincs módja minden részletet bemutatni, így nem illusztrálja a száj-, illetve orrüregi képzés különbségeit, s nincs mód rámutatni vele a zöngésség-zöngétlenség „okozójára” sem. Ugyanakkor a képernyőn megjelenő audiovizuális transzkódolás révén létrejött oszlopos megjelenítés lehetővé teszi az utóbbi különbségek megfigyelését és gyakorlását is.

A vizuálisan megjelenő képet utánozva, továbbá törekedve az akusztikai képek mind nagyobb hasonlóságának elérésére (a referencia és saját bemondás összevetése nyomán) várható a saját kiejtés javításának lehetősége és igénye. Ezek tudatosabb hangképzéshez, jobb beszédteljesítményhez, nem utolsó sorban pedig javuló írásbeli nyelvhasználathoz vezetnek. A két nyelv, azaz a magyar jelnyelv és a magyar beszédnyelv eltérő nyelvtipológiai sajátosságaiból következően is – és a hallási probléma fokától sem függetlenül – maradnak el például sokszor a ragok a szavak végéről. A magyar írott nyelv tökéletesebb elsajátítása korunk kikerülhetetlen követelménye és a bilingvális oktatási programnak is az egyik célkitűzése.

A kezdeti rendszerben az egész arc látható volt, de a tapasztalok azt mutatták, hogy elegendő a száj megjelenítése. A teljes arc esetén az artikuláció elvész, így azonban a diákok könnyebben tudják értelmezni. A szurdopedagógusok egyedi igénye volt a beszélő fej dupla megjelenítése, így jobban tudják szemléltetni két hang vagy szó képzésének különbözőségét.

3.2 Az automatikus minősítés és a minősítési skála létrehozása

A 3. fejezetben bemutatott beszédasszisztens rendszer egyik szolgáltatása az automatikus minősítés és visszajelzés, hogy a hallássérült diákok önállóan gyakorolhassák a mintaszavak kimondását. A tanulás során a referencia kiejtést a szerver vagy a tanár produkálja. A diák ezt igyekszik utánozni az ő aktuális bemondásával. Ezzel rokon probléma merül fel a beszéd gépi felismerésénél. Előre

(modellezés segítségével) eltárolt, valóságos beszédből származó beszédrészletek (hang, hangátmenet, szó, stb.) közül kell a felismerendő beszédrészlethez leghasonlóbbat megtalálni, és ha a hasonlóság elég nagy, akkor a beszédrészlet felismertnek tekinthető. A hallássérültek beszélni tanításánál a hasonlóság automatikus ellenőrzése és a visszajelzés generálása alapkutatás, amely megköveteli egy hasonlósági mérték kidolgozását. A hasonlósági mértéknek monoton összefüggésben kell lennie a hallássérült és halló bemondók által kiejtett hangok, hangkapcsolatok, szavak szubjektív (épen halló emberek által végzett) tesztek átlagos megítélésével (MOS = Mean Opinion Score). A különböző lényegkiemelési és távolság számítási módok elemzésével kidolgozható a szubjektív értékelésnek megfelelő hasonlósági mérték. Ez az alapja az előrehaladás értékelésének és a visszajelzés generálásának. Az értékelés nyilvánvalóan a korábbi eredményekkel összevetve alakítható ki, hiszen ugyanaz a kiejtés egyik tanulónál siker, a másiknál kudarc lehet. Az automatikus értékelés verifikálása érthetőség vizsgálattal történhet.

A célként létrehozandó rendszerhez hasonló fejlesztés az, amelyet olyan személyeknél alkalmaztak, akiknek gégerák miatt eltávolították a gégéjüket, vagy olyan gyerekeknél, akik ajak- és szájpadhasadékkal születtek. Ezeknél a személyeknél szignifikáns összefüggések érhetők el a szubjektív és az automatikus minősítés között. Az egyes betegségekhez tipikus, jól detektálható beszédhibák társulnak. Ezekhez a beszédhibákhoz a kutatók és fejlesztők rendelkezésére álltak minták, így megalkotható volt az automatikus minősítés. A PEAKS (Program for Evaluation and Analysisof all Kinds of Speech Disorders) egy rögzítő és elemző rendszer hangképzési és beszédzavarok automatikus vagy manuális minősítéséhez. A PEAKS rendszert több kórház is alkalmazza nem csak Németországban. A pedagógusok számára is segítséget nyújt az oktatásban. A félautomatikus módszer a mérsékelttől a jó értékelésig 60%-os korrelációs szintet ér el az összes fonetikai rendellenesség esetén. Beszélők szintjén a percepciós és az automatikus értékelés között 89%-os korrelációt értek el, teljesen automata rendszernél pedig 81%-ot. Ez a korrelációs eredmény az értékelők közötti korrelációs tartományba esik [36], [37]. Az általunk fejlesztett rendszer is hasonló tulajdonságokkal rendelkezik kivéve, hogy a hallássérültek beszédhibái nem definiálhatóak vagy csoportosíthatóak és szűkíthetőek le. Ezért a feladatom egy olyan automatikus minősítés megalkotása, ami független a beszédhiba típusától és az aktuális bemondás mérhető jellemzőin alapszik [3], [7], [28].

3.3 Szóadatbázis az automatikus minősítés megalkotásához

A már említett minősítési skála megalkotásához szükséges adatbázis mintáit eltérő beszédprodukciós fejlettségi fokú gyerekektől gyűjtöttem be. A mintákat laikus hallgatók valamint szurdopedagógusok minősítették. Az artikuláció helyessége a szépen beszélő ép hallóktól az alig érthetően beszélő hallássérültekig terjedt. A minták rögzítését az adott oktatási intézményen belül végeztem egy csendesnek mondható szobában a pedagógusok segítségével. A gyerekek a bemondások előtt egyszer átnézhették a felolvasandó szavakat, hogy a bemondást csak minimális mértékben befolyásolják az olvasási nehézségek.

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. Minden pedagógus csak a másik iskola diákjainak bemondását értékelte, hogy elkerüljük a beszélő felismeréséből eredő előítéleteket. A bemondást többször is meghallgathatták az értékelők és megjegyzéseket is fűzhettek a mintákhoz. Az eredményeket internetes alkalmazáson keresztül rögzítettük. A minősítés alapját a pedagógusok esetén az általuk meghatározott ötfokozatú 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 (az a beszédzavar, mely szerint egyes hangzók hiányosan képeztetnek, 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.

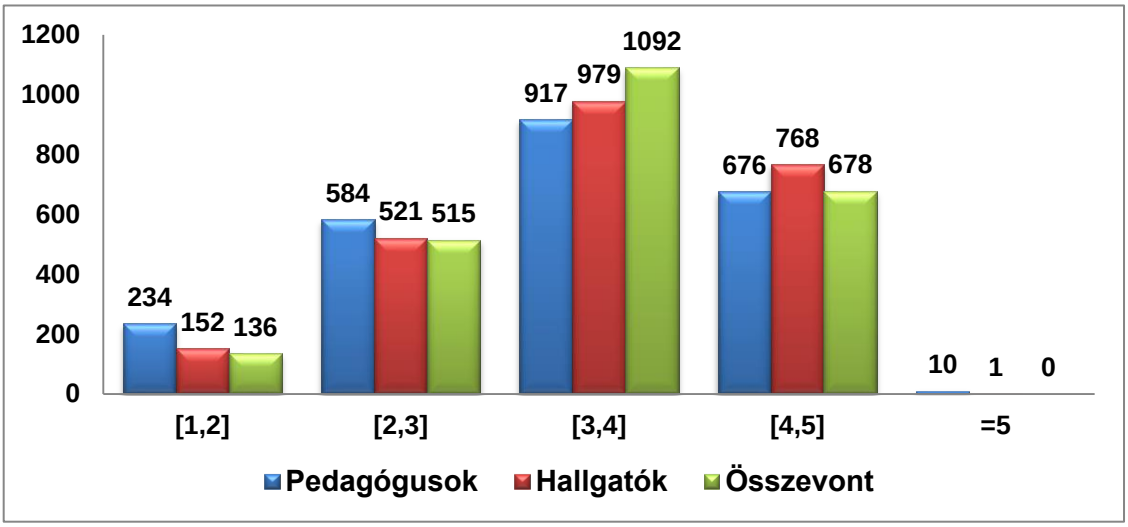
A laikus hallgatóknak a mindennapi nyelvhasználat alapján kellett 1-től 5-ig pontozniuk a bemondásokat.

1. táblázat *A 2421 szó minősítésének eredményei intervallumokra bontva*

	Értékelések		
	Pedagógusok	Hallgatók	Összevont
[1-2]	*234*	*152*	*136*
[2-3]	*584*	*521*	*515*
[3-4]	*917*	*979*	*1092*
[4-5]	*676*	*768*	*678*
=5	*10*	*1*	*0*

Az 1. táblázat és a 8. ábra a pedagógusok, a hallgatók és a két csoport összevont értékeléseit tartalmazza, illetve szemlélteti intervallumokra bontva. Az összevont értékelés egy adott hangminta összes értékelésének átlagát jelenti. Halló beszédével azonos szinten érthető kiejtésű szó a pedagógusok egyöntetű véleménye szerint csak 10 esetben fordult elő, míg a hallgatók szerint csak 1 esetben. Érthetetlen vagy ahhoz közeli minták száma sem kiemelkedő, az adatbázisban többségében közepesen érthető bemondások fordulnak elő.

A szegmentálási vizsgálatok elvégzéséhez ebből a 2421 szóból választottam ki 300 szót. A kiválasztott szókészlet elég változatos nem csak a szavak hosszúsága alapján, hanem a hangkapcsolatok előfordulásának szempontjából is, ami az egész szóadatbázisra is jellemző. A rögzítésre került minták szókészletét a szurdopedagógusok készítették elő körültekintően figyelembe véve az egyes diákok aktív szókészletét. A fejlesztésben többségében 7. és 8. osztályos tanulók vesznek részt, de van néhány olyan diák is, akik 1. és 2. osztályos korukban kerültek be a programba. Ezekben az esetekben nem várhattuk el, hogy olyan szavakat mondjanak fel, amelyek nem képezték eddigi tanulmányaikat így a pedagógusok személyre szabottan alkották meg a felmondott szavak készletét.

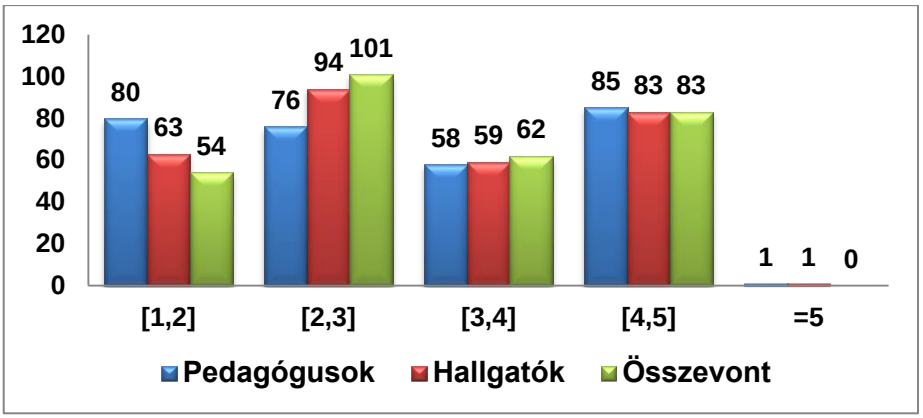


8. ábra *A 2421 szó eloszlása az értékelések alapján*

A 2. táblázat és a 9. ábra a 300 szót tartalmazó adatbázis minősítéseit mutatja be. Azért is érdemes külön szemléltetni az értékelések eloszlását, mert a hallgatók az esetek 63 %-ban jobbra értékelték a bemondást a szurdopedagógusokkal szemben. (Ez az érték mind a 2421 szót figyelembe véve is 60 %). A minták kiválasztásánál törekedtem arra, hogy az intervallumok függvényében egyenletes eloszlást mutassanak a minták, ami az ötösre és egyesre értékelt minták esetén nem volt lehetséges, tekintve azok számát.

2. táblázat *A 300 szó minősítésének eredményei intervallumokra bontva*

	Értékelések		
	Pedagógusok	Hallgatók	Összevont
[1,2]	*80*	*63*	*54*
[2,3]	*76*	*94*	*101*
[3,4]	*58*	*59*	*62*
[4,5]	*85*	*83*	*83*
=5	*1*	*1*	*0*



9. ábra *A szavak eloszlása értékelések alapján*

A minősítési skála és az automatikus kiértékelés megvalósításához végzett vizsgálatok során a hallgatók és a pedagógusok általi összevont értékelést tekintem referenciának és a minősítések átlagát fogom felhasználni a további vizsgálatok során.

4 Gyenge minőségű beszéd szegmentálása

A beszédtempó beszélőről beszélőre kiejtésről kiejtésre változik. Ezek a nemlineáris megnyúlások és rövidülések nem feltétlenül számítanak hibás ejtésnek. A hallássérültek az átlagos beszédtempónál általában lassabban beszélnek. Az egyes hangok kiejtésének minősítéséhez össze kell párosítani a referencia alakzat és az aktuális kiejtés időszegmenseit. A referencia és az aktuális hullámforma azonos hosszúságúvá tétele lineáris nyújtással, illetve zsugorítással elvégezhető. Ez azonban nem biztosítja az egyes hangok időbeli párhuzamát, mert a kiejtés ritmusa eltérhet a referenciától. Egyes hangokat hosszabban másokat rövidebben ejtve a lineáris vetemítésnél nem azok a hangszegmensek kerülnek fedésbe, amelyekre hasonlítaniuk kell így az összevetés hamis eredményre vezet. Különösen jellemző a hallássérültek beszédére az egyes hangok megszokottól eltérő idejű artikulációja. A referencia és a vizsgált beszéd összehasonlításához tehát dinamikus idővetemítésre van szükség, amire a számítógépes beszédfeldolgozásban kidolgozott eljárások és algoritmusok állnak rendelkezésre. Ezek a módszerek jó minőségű beszédre, a mindennapi kommunikációban elfogadható kiejtésre megfelelően működnek. A torz hangokra, a rendkívül elnyújtott, akadozó beszédre gyenge eredményt szolgáltatnak. A beszéd minősítésének kulcskérdése a helyes szegmentálás. Egyik kutatási célom a szegmentálásra szolgáló módszerek továbbfejlesztése annak érdekében, hogy a szinte érthetetlen beszédre is használható szegmentálási eredményeket kapjak.

4.1 A felhasznált beszédadatbázis

Az beszédadatbázis hanganyagaival tanítottam be a későbbiekben ismertetett szegmentáláshoz és automatikus minősítéshez felhasznált neurális hálózatot. Az adatbázis olvasott szövegeket tartalmaz, speciális fonetikai elvárásoknak megfelelően került összeállításra és általános felhasználói környezetben (irodákban, laboratóriumokban, lakásokban) rögzítették, ezért megfelelően alkalmazható volt a célzott vizsgálatok elvégzéséhez [S12].

Az adatbázis összefoglaló műszaki adatai:

- magyar nyelvű, olvasott szövegű, személyi számítógépes környezetben felvett adatbázis;
- 16 bites, 16 kHz-es mintavételezéssel;
- 332 beszélő közvetlenül a számítógépbe rögzített hanganyaga;
- beszélőnként 12 mondat és 12 szó;
- a felvételek többféle mikrofonnal, hangkártyával, PC-kel készültek;
- környezet változó zajosságú irodahelyiség, laboratórium, otthoni környezet;
- az adatbázis teljes anyaga annotált;
- az adatbázis harmada (100 beszélő) kézileg szegmentált és címkézett.

4.2 Dinamikus idővetemítési módszerek

Akkor beszélhetünk ideális időbeli összehasonlításról, ha a két mintát az egyes beszédhangok mentén illesztjük egymáshoz. Ezt a bevált módszert alkalmazzák gépi beszédfelismerésnél, ám ez a módszer nyelvfüggő [33,52].

Egy korábban kifejlesztett akusztikai-fonetikai (AF) hangosztályokra épülő vetemítési módszer célja a prozódia (a dallam, a kiejtés sebessége, a ritmus, a hangsúlyozás, a hangerő és a hangszínezet együttese) összehasonlítása, ami több nyelvre is alkalmazható. Olyan vetemítési eljárásra volt szükség, amely szigorúan csak a beszédhangok mentén illeszti össze a két mintát és csak azokon belül valósít meg lineáris skálázást. Az eljárás három alapvető módszert alkalmaz. Első lépés a szegmentálás, majd ez után következik a mintaillesztés és a vetemítés az aktuális és a referencia bemondás között. Ennél a módszernél az újdonságot a gépi szegmentálás megvalósítása jelentette, ugyanis ehhez általános akusztikai hangosztályokat használtak fel, amik nyelvtől független artikulációs konfigurációkat határoztak meg [33]. Ebből következett a több nyelvre való alkalmazhatóság is. Ebben az esetben a fejlesztők feltételezték, hogy az aktuális és a referencia minta között az eltérés minimális (a bemondó kooperatív), ezért háromféle eltéréssel számoltak: beszúrás, kihagyás, másképpen ejtés. Az AF gyenge minőségű beszédre nem adaptált szegmentálási eljárás.

A dinamikus idővetemítésnek más eltérő szabályrendszeren alapuló változatai is léteznek. Egy másik dinamikus idővetemítési (Dynamic Time Warping – DTW) eljárás az optimális időillesztést, mint minimális hosszúságú illetve súlyú út keresését tekinti egy adott gráfban. Tételezzük fel, hogy a felismerendő x szó l darab szegmensből áll és az i-ediket (i=1,2,….,l) jellemző adatokat az x_i vektorban foglaljuk össze. Ezen elemi vektorokat azután egyetlen ”hosszú” vektorrá fűzzük össze az osztályozási algoritmusban. Tehát a bejövő szót az x_1, x_2,….x_l vektorsorozat jellemzi. Hasonlóképpen jellemezze az y_1, y_2, ….y_r vektorsorozat azt az y szótárelemet, amellyel a beérkezett szót éppen össze akarjuk hasonlítani. (Feltételezzük, hogy az egyes szegmenseket jellemző elemi vektorok mindig azonos dimenziószámúak).

A cél, hogy az x_1, x_2,…x_l vektorsorozatból némelyek megismétlésével és mások elhagyásával egy olyan $x_1, x_2, \ldots x_r$ (r hosszúságú) vektorsorozatot állítsunk elő, amelyre

$$D = \sum_{i=1}^{r} d(x_i, y_i) \qquad (7)$$

”távolság” minimális. Itt a *d(x,y)* egy tetszőleges adott távolságfüggvény. Az $x_1, x_2, \ldots x_r$ vektorsorozat előállításakor az alábbi mellékfeltételeket kell betartani:

- 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 [16].

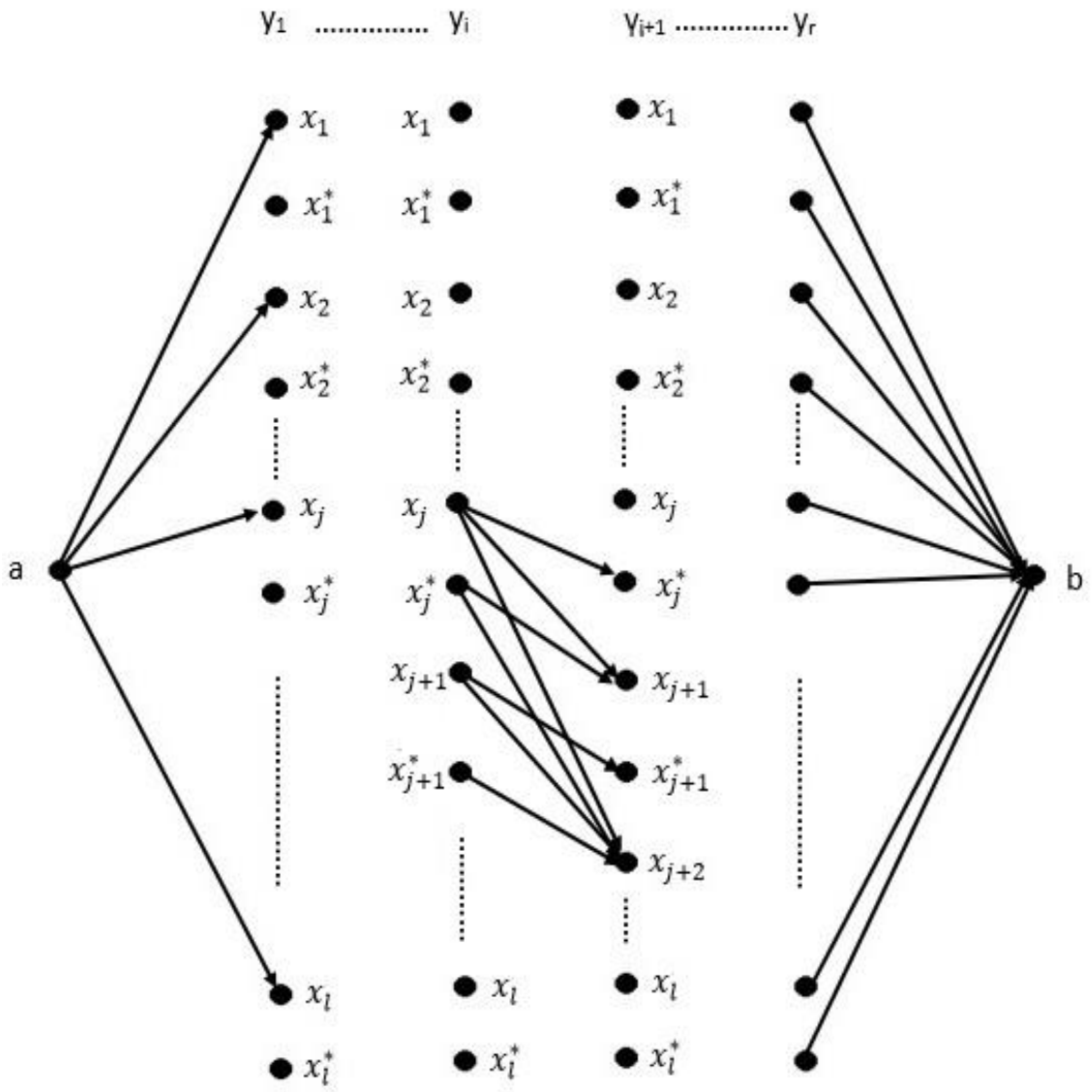
A 10. ábrán látható a megoldó algoritmus működése. Konstruálunk egy gráfot, amelyben az ábrán az y_i jel alatti pontok mindegyiket (tehát az i-edik oszlop) az x_i vektor egy lehetséges értékének felel meg. Itt az x_j^* jelű pont az x_j vektor megismételt változatának van megfeleltetve. Minden ponthoz hozzárendelünk egy súlyértéket is, mégpedig az y_i oszlopban álló, x_j-vel vagy x_j^*-gal címkézett pontokhoz a $d(x_j, y_i)$ értéket.

A gráf éleit az ábra közepén látható módon húzzuk be. Az i-edik oszlopban levő x_j –t reprezentáló pontot az i+1 – edik oszlopban azokkal a pontokkal kötjük össze, amelyeknek megfelelő szegmens állhat a vetemített szóban az i+1 – edik helyen, ha az i – edik helyen x_j áll. Ez lehet x_j^* (ha x_j –t megduplázzuk), x_{j+1} (ha lineárisan haladunk), valamint x_{j+2} (ha x_{j+1} –et elhagyjuk). Hasonlóan az i –edik oszlop x_j^* pontját összeköthetjük az i+1 –edik oszlopban x_{j+1} –el vagy x_{j+2} –vel, de például x_j –vel vagy x_j^* -gal nem, hiszen ha az i –edik helyen x_j^* áll, azaz x_j második előfordulása, akkor az i+1 –edik helyen már nem állhat ugyanaz, mert ez már x_j megháromszorozása lenne.

Végül hozzávesszük a gráfhoz a zérus súlyú (”fiktív”) *a* és *b* pontokat a 11. ábra szerint. (Ezek csupán az algoritmus szemléletesebb megfogalmazásához szükségesek.)

Ezután könnyen belátható, hogy a gráfban minden *a* –ból *b* –be vezető út kölcsönösen egyértelműen megfeleltethető az *X* szó egy *R* szegmensre normált és egyben vetemített változatának. Az ebben szereplő vektorok az úton levő csúcsokhoz rendelt vektorok, az út *a* –ból *b* –be való bejárásának sorrendjében. Az adott vetemítéshez tartozó *D* távolság pedig éppen az út pontjainak összege.

Ezután az optimális időillesztés problémája ekvivalens azzal a feladattal, hogy keressünk a leírt és megkonstruált gráfban minimális súlyú *a* –ból *b* –be vezető utat. Erre pedig ismeretes rengeteg gyors és hatékony algoritmus.



10. ábra *A megoldó algoritmus működésének szemléltetése*

4.3 Az idővetemítés szabályainak módosítása a gyenge minőségű beszéd szegmentálására

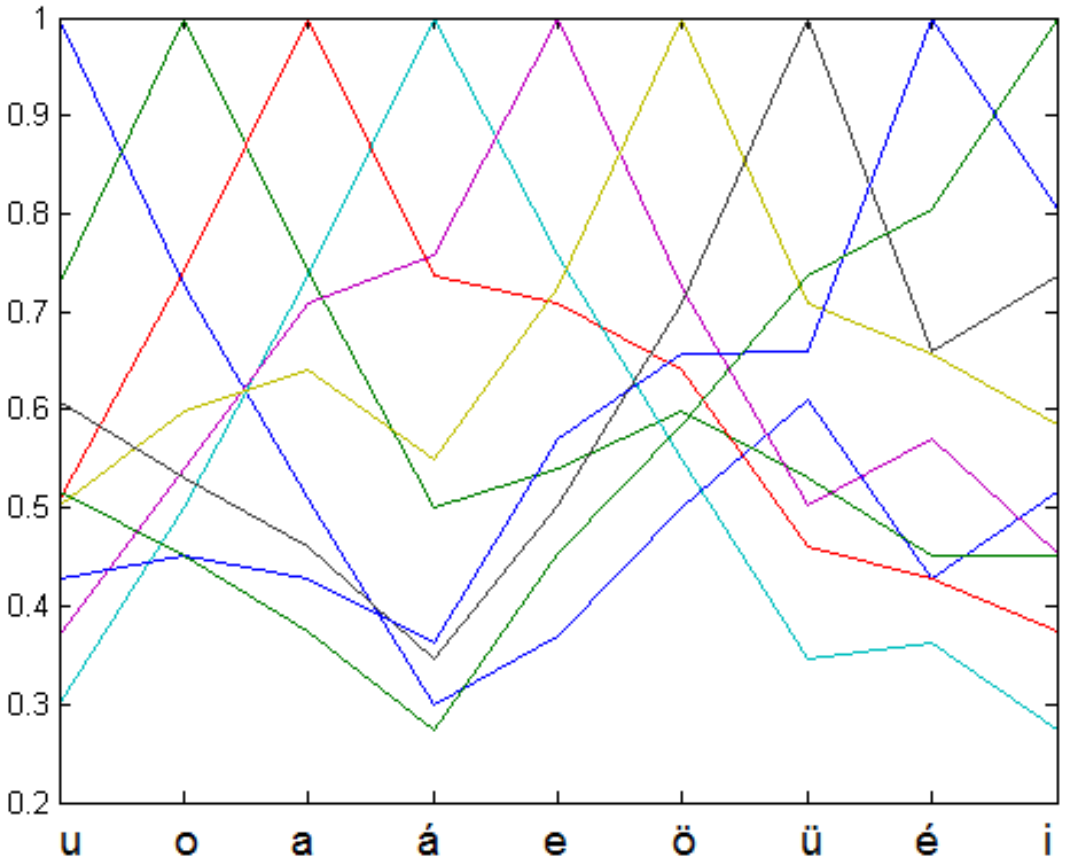
Az ismertetett AF és DTW eljárásnál leírt szabályok alapján egy referencia szóhoz az idővetemítés nem bizonyult sikeresnek. A módszereket alkalmazva hallás alapján értékeltem a szegmentálási eredményeket, amik rendkívül torz eredményt mutattak. Felhasználva a bemutatott két szegmentálási eljárás alapszabályait megalkottam egy a gyenge minőségű beszéd szegmentálására alkalmas adaptált dinamikus idővetemítési eljárást (Adapted Dynamic Time Warping – ADTW). Az eljárás módszerét és annak vizsgálatát a következő alfejezetek mutatják be.

4.3.1 A referenciagenerálás

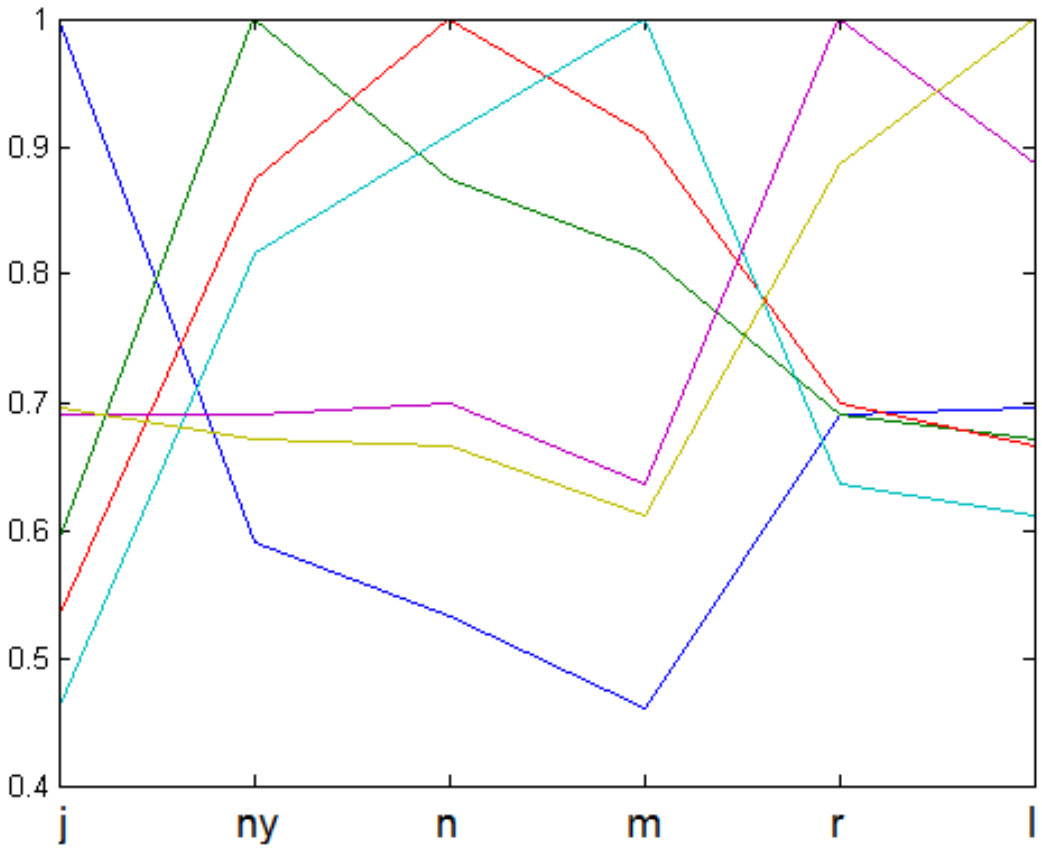
A dinamikus idővetemítésnél a keresett szót egy referencia bemondással vetjük össze és keressük egyes időkeretek megismétlésével, illetve kihagyásával a referencia bemondáshoz leginkább hasonló ütemezést. A hallássérült gyerekek bemondásai közül

a nehezen érthetők nem alkalmasak arra, hogy a referencia bemondáshoz valamilyen hasonlósági mérték szerint eléggé hasonlítsanak. Próbálkoztam férfi, női, gyerek bemondáshoz és szintetizált hanghoz is vetemíteni a keresett szavakat, ezek azonban nem voltak sikeresek. Egy konkrét bemondást referenciaként használva a szegmentálás sikertelenségét az individuális különbségeknek tulajdonítottam. A sok bemondóval tanított neurális hálózat statisztikai alapon jobban visszatükrözi az egyes hangokhoz mérhető hasonlóságot.

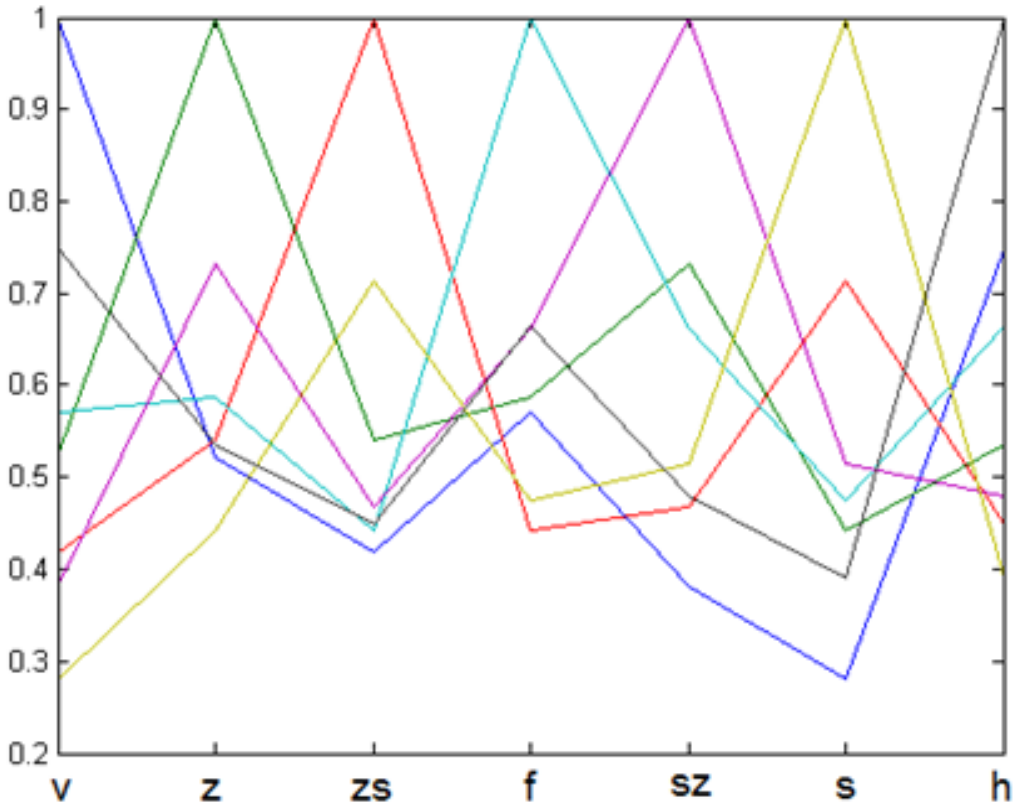
Egy 300 bemondótól származó 4 és fél órás hangadatbázis alapján PLP lényegkiemelést alkalmazva meghatároztam az egyes hangok stacionárius szakaszaihoz tartozó együtthatók átlagát. Majd Euklideszi távolságot képeztem a magyar beszédhangok átlagai között [15]. A szünet és a négy osztály valamint a 32 hanghoz tartozó outputot használtam a távolság meghatározására. A normalizált távolság megfordításával hasonlóságmértéket képeztem az egyes hangok között. Az alábbi ábrákon szemléltetem a neurális hálózat osztályaihoz tartozó hangok egymáshoz számított hasonlósági mértékeit (11. – 14. ábra).



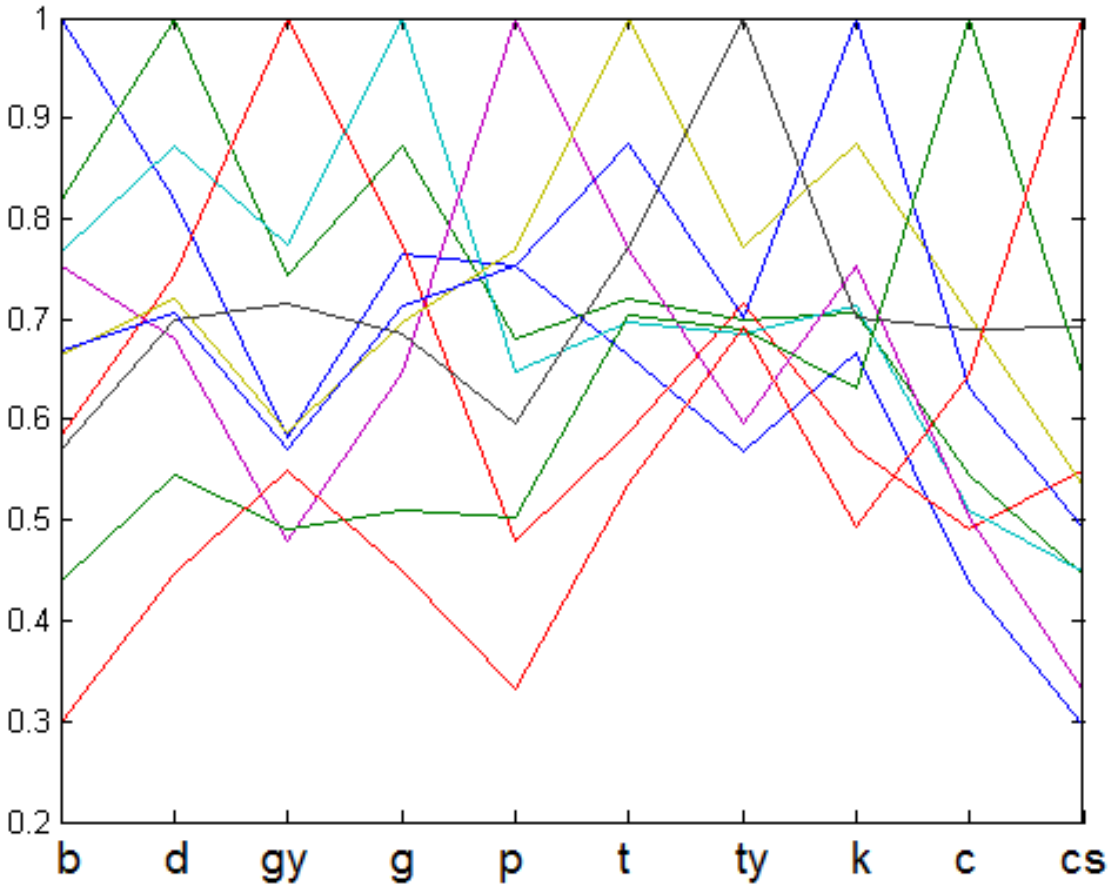
11. ábra *Magánhangzók hasonlósági mértéke*



12. ábra *Félmagánhangzók hasonlósági mértéke*

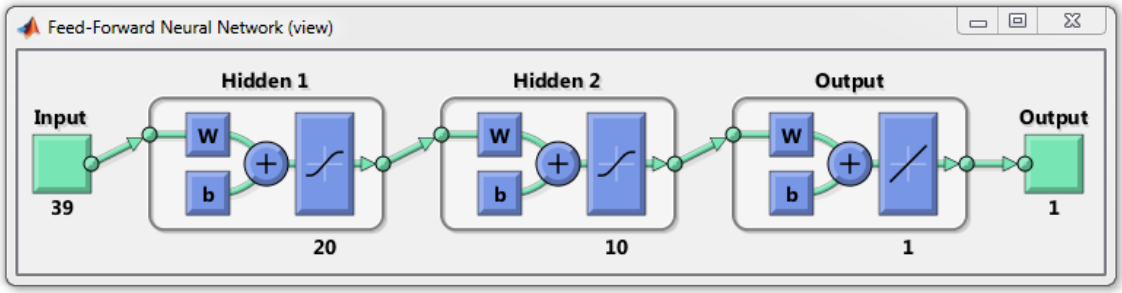


13. ábra *Réshangok hasonlósági mértéke*



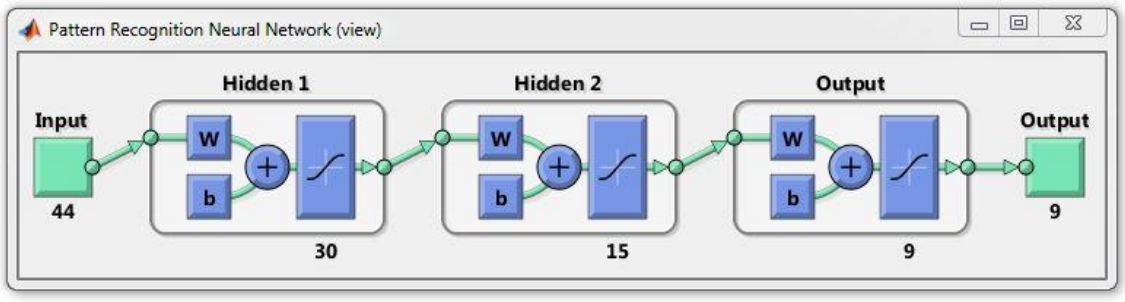
14. ábra *Zárhangok hasonlósági mértéke*

A referenciát úgy alakítom ki (ebben a feladatban nem a szó felismerése a cél, hanem a bemondás illesztése, ezért rendelkezésre áll a vizsgált szó fonetikus leírása), hogy a hangok átlagos időtartamával számolva az adott hanghoz tartozó kimenetet és a hangot magába foglaló csoport kimenetét aktívvá teszem.



15. ábra *Az akusztikai hangosztályt meghatározó neurális háló modellje*

A 15. ábrán látható neurális hálózat 2 rejtett réteggel és egy kimenettel rendelkezik, aminek a kimenete akkor aktív, ha a bemeneti jel az adott hangosztályba sorolható. Az első rejtett réteg 20 neuront, a második 10 neuront tartalmaz. A 16. ábra a magánhangzók osztályához tartozó neurális modellt szemlélteti szintén 2 rejtett réteggel 30 illetve 15 neuronnal. A kimenet 9 állapotú a magánhangzók számának megfelelően.



16. ábra *A magánhangzók akusztikai hangosztályának neurális háló modellje*

Az egyes hangok referencia időtartamát az adott beszédhang átlagos hosszára állítottam be [42]. 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 hozzárendelem. 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 4.3. fejezetben ismertetett szabályaival a vetemítés jobb eredményt mutatott, mint azokban az esetekben, amikor referenciaként bemondott szavakat használtam.

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

Ezért:

- minden hang után beiktattam 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 találkoztam.

Ezért:

- egy időkeret kétszeri ismétlése is megengedhető, ezzel egy időintervallum háromszorosára nyújtható.

4.3.2 Az alkalmazott lényegkiemelés

Összefüggésben a 4.3.1. fejezetben tárgyalt referenciagenerálással olyan lényegkiemelési módot kellett választanom, amely alkalmazkodik a mesterségesen generált referenciához. A szegmentáláshoz alkalmazott neurális hálózat bemenetét PLP jellemzők képezték. A lényegkiememlés során az aktuális 40ms-os keret mellé a megelőző 80 ms-os szakasz két keretének átlagát és a következő 80 ms két keretének átlagát veszem. $3*13$jellemző írja le a 200 ms-os intervallum közepére eső hangot. A fonetikai osztályozásra szánt 5 neurális hálózat betanítása ezekkel a paraméterekkel történt. A fonetikai osztályokon belüli hangok felismerésére tanított neurális hálózatok a

39 PLP jellemzőn túl az osztályozó neurális hálózatok kimeneteit is megkapták inputként. 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*).

A zárójelekben az osztályokhoz tartozó külön neurális hálózat kimeneteihez tartozó hangokat soroltam fel.

4.4 A létrehozott speciális idővetemítési eljárás összevetése más módszerekkel

Az előzőekben ismertetett ADTW eljárást összehasonlítottam az AF szegmentálási eljárással és egy HMM modelleket alkalmazó forced aligment (”kényszerillesztés”) szegmentálási módszerrel, amihez PLP lényegkiemelést használtam. Több lényegkiemelési módszert is megvizsgáltam mielőtt a perceptuális lineáris predikcióra esett a választásom. A teszteléshez a 5.5 fejezetben bemutatandó beszédadatbázisból használtam fel mintákat. Az 3. táblázat és 17. ábra mutatja be azokat a szegmentálási eredményeket, amiket az alábbi lényegkiemelési módszerekkel sikerült elérnem:

- MFCC – Mel-Frequency Cepstral Coefficients, Mel-frekvencia kepsztrális együtthatókat alkalmazva 13 komponens képezi az alkalmazott eljárás alapját [24];
- PLP – Perceptual Linear Prediction (lásd 1.1.1 fejezet);
- MEL: A 125 Hz – 8 kHz közötti frekvenciatartományt a mel-skála alapján felosztjuk 30 részre (lásd 1.1.1 fejezet).

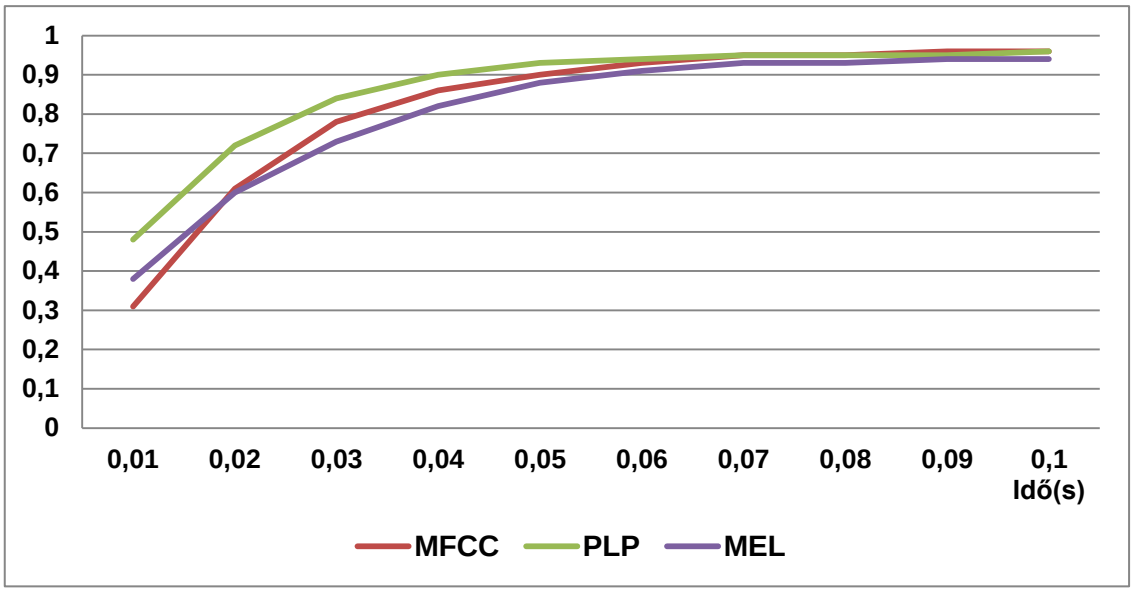
A táblázat oszlopaiban az egyes lényegkiemelési módszerek eredményei láthatók az alapján, hogy az adott mintaszám hanyadrészénél kisebb az időbeli eltolódás az adott időkorlátnál.

3. táblázat *A szegmentálás pontossága különböző lényegkiemelési módszerek esetén*

	Lényegkiemelési módszerek		
Tolerancia [s]	***MFCC***	***PLP***	***MEL***
<= 0,01	*0,31*	*0,48*	*0,38*
<= 0,02	*0,61*	*0,72*	*0,6*
<= 0,03	*0,78*	*0,84*	*0,73*
<= 0,04	*0,86*	*0,9*	*0,82*
<= 0,05	*0,9*	*0,93*	*0,88*
<= 0,06	*0,93*	*0,94*	*0,91*
<= 0,07	*0,95*	*0,95*	*0,93*
<= 0,08	*0,95*	*0,95*	*0,93*
<= 0,09	*0,96*	*0,95*	*0,94*
<= 0,10	*0,96*	*0,96*	*0,94*

Összevetve az MFCC, PLP és MEL lényegkiemelési módszereket a PLP bizonyult a legjobbnak, ezért a szegmentálási és energia vizsgálatok során ezzel a lényegkiemelési módszerrel betanított rejtett Markov-modelleket és neurális hálózatokat használom fel. A PLP lényegkiemelést a 12 komponenshez hozzáadott energia, ezek első és második deriváltjai alkotják. A HMM modelleket előzetes tesztelések alapján 7 állapotúnak választottam meg. A betanítást a HTK toolkit keretrendszerrel, 10 ms-os időkerettel és hang alapú felismerési egységgel végeztem. (A diád alapú betanított beszédfelismerők a vizsgálatok során nem teljesítettek olyan jól, mint a hang alapú felismerők.)

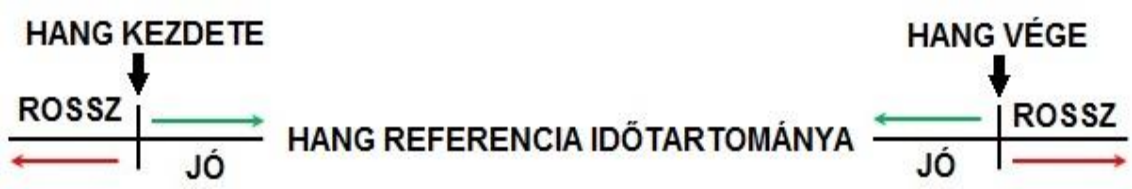
A HTK toolkit keretrendszer elsősorban rejtett Markov-modell alapú beszédfelismerők fejlesztésére szolgál, ami megfelelő nagyságú infrastruktúra hátteret biztosít ezen feladat számára, de egyéb beszédfeldolgozási vizsgálatok számára is népszerű eszközként szolgál [57].



17. ábra *A szegmentálási hibák toleranciája különböző lényegkiemelkiemelési módszerek esetén (MFCC,PLP,MEL)*

A betanított modellek létrehozásánál a felismeréshez módosítottam a nyelvtan fájlt, így megengedi, hogy az egyes hangok között akárhányszor előfordulhasson szünet. A PLP eljárás így részlegesen adaptáltnak tekinthető gyenge minőségű beszédre. Az eredményeket kézileg szegmentált eredményekkel vetettem össze, amiket kineveztem referenciának. Az alábbi 4.-6. táblázat a vizsgált eljárások eredményeit szemléltetik a referenciához viszonyítva oly módon, hogy az első oszlop az egyes hangok kezdetét akkor tekinti rossznak, ha az hamarabb kezdődik, mint a referencia és akkor jónak, ha később, mint a referencia. Az egyes hangok végén pedig fordítottan, azaz az eredmény jónak tekinthető, ha az hamarabb végződik, mint a referencia és rossznak, ha később, mint a referencia. A sorok az egyes eltolási tűréshatárok figyelembevételével tartalmazzák az eredményeket. Mivel a hangok közepét keressük, ezért nem tekinthető hibának az, ha az adott időérték a hang tartományába esik a kezdetnél és a végnél egyaránt. Az értékek azonban nem veszik figyelembe azt az esetet, ha például egy hang kezdeti ideje még a végénél is később kezdődik, azaz teljesen kívül esik a hang időtartományán. Erre vonatkozóan a 7. táblázat tartalmaz olyan kiértékeléseket, ahol az

egyes szegmentálási eljárásokra 0 ms-os tűréssel vizsgáltam az ilyen jellegű hibákat. A teszteket a 2421 szóból kiválasztott 300 szavas tesztadatbázison végeztem el. A könnyebb érthetőségért a 18. ábrán szemléltetem, hogy mi tekinthető jó és rossz eredménynek.

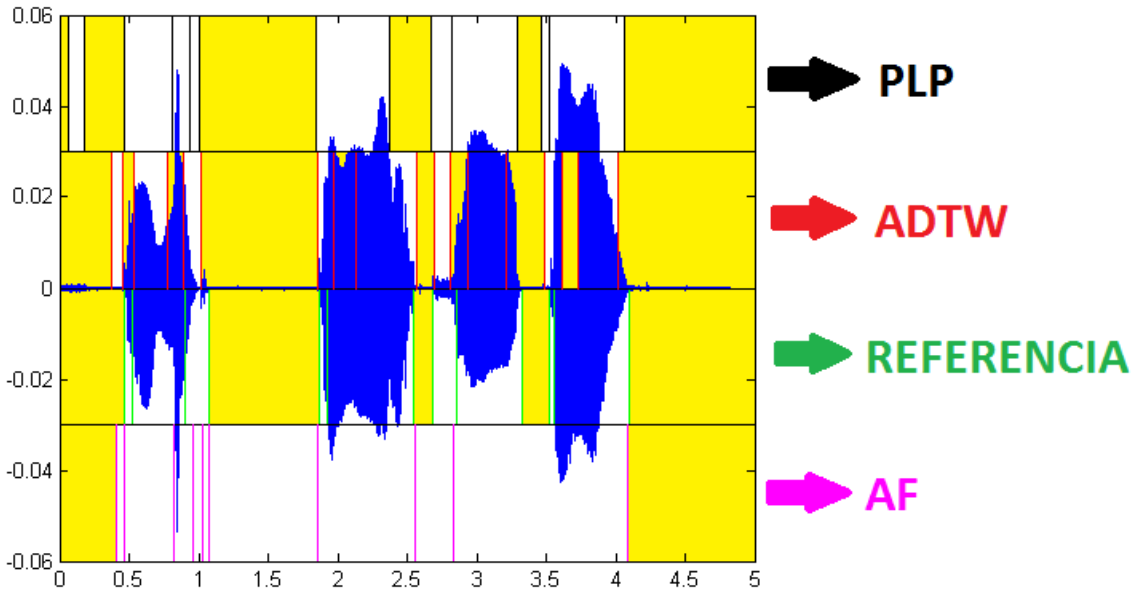


18. ábra *A szegmentálási eredmények osztályozása*

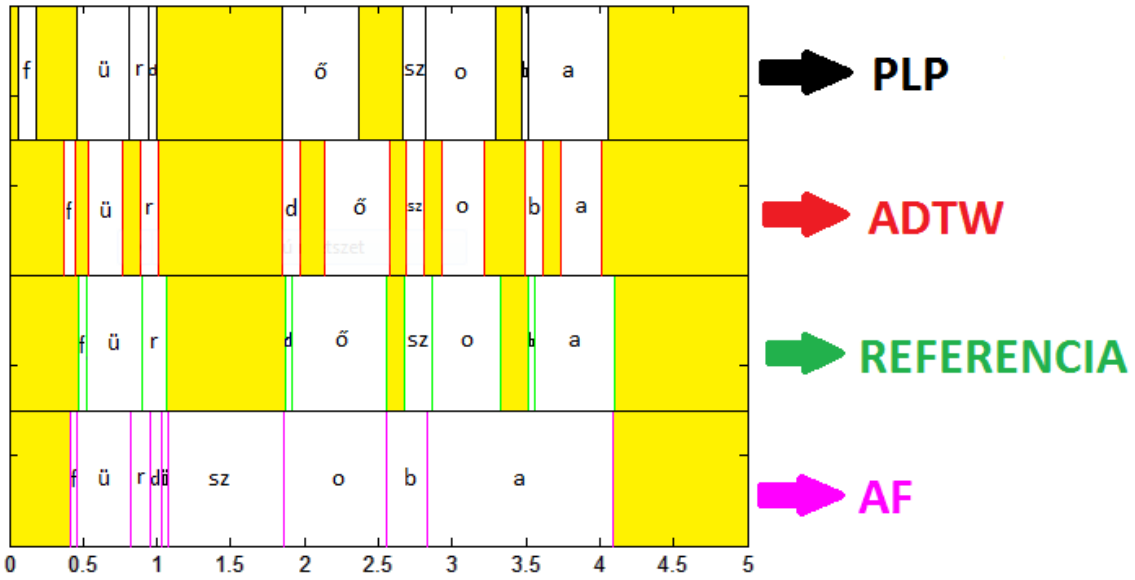
A 19.-20. ábrán a *fürdőszoba* szó hullámformája látható és a vizsgált szegmentálási eljárások eredményei. Az ábra kiemeli azt a lényegi különbséget, hogy az AF szegmentálási eljárás mivel jó minőségű beszédet feltételez bemenetként, ezért nem enged szüneteket beszúrni az egyes hangok közé a másik két eljárással ellentétben. A bemondást a pedagógusok egyesre értékelték. A szüneteket sárga színnel jelöltem mindkét ábrán.

A bemondás az alábbi linkeken meghallgatható:

http://mazsola.iit.uni-miskolc.hu/~pinter/furdoszoba.wav



19. ábra *A fürdőszoba bemondás szegmentálási eredményei I.*



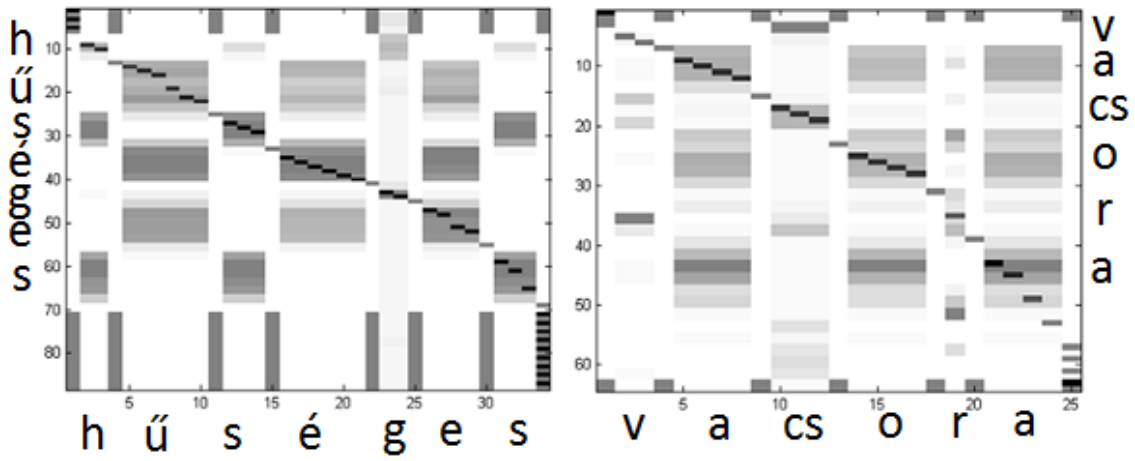
20. ábra *A fürdőszoba bemondás szegmentálási eredményei II.*

A 21. ábra a *hűséges* és a *vacsora* szó dinamikus vetemítését mutatja be a specializált módszerrel. Az első szó 5-ös a második szó pedig 1-es minősítést kapott a pedagógusok és a hallgatók értékelése szerint. A vízszintes tengelyen az automatikusan generált referencia minta látható időkeretenként, függőlegesen pedig az aktuális bemondás. A téglaterületek világossági értéke az egyes keretek egymáshoz való hasonlóságát (azaz az egyes hangok hasonlósági mértékét) fejezi ki. Minél sötétebb a terület annál nagyobb a hasonlóság. Az ábra tetején és alján megjelenő keskeny sávok a minden egyes hang közé beszúrt szüneteket szemléltetik. A fekete átlós sávok az illesztési pontok. Annál jobb a felismerés, minél több fekete sáv esik a legsötétebb területekbe.

A bemondások az alábbi linkeken meghallgathatóak:

http://mazsola.iit.uni-miskolc.hu/~pinter/vacsora.wav

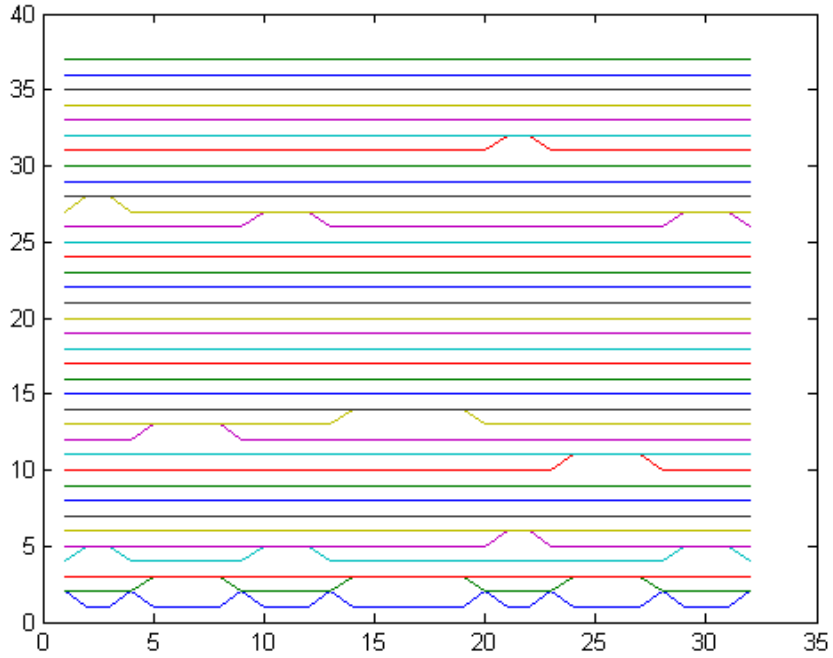
http://mazsola.iit.uni-miskolc.hu/~pinter/huseges.wav



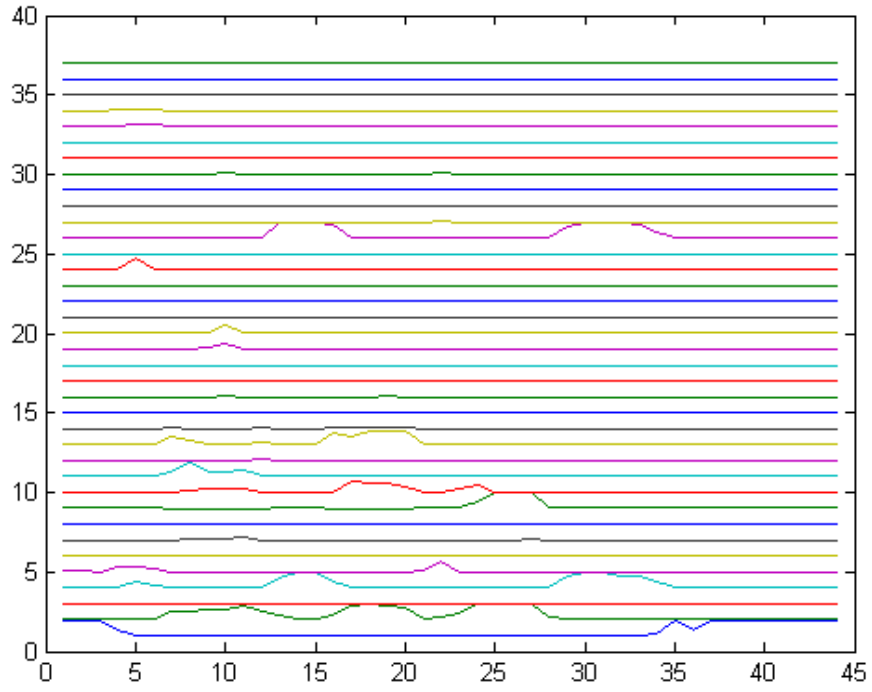
21. ábra *A hűséges és vacsora szó dinamikus vetemítése az ADTW módszerrel*

A 22.-25. ábrán a *hűséges* és a *vacsora* szó illesztésekor a kimenetek az ábrákon látható aktivitási értékeket veszik fel. A 22. és 24 ábrán a referencia generáláskor aktivizált kimenetek láthatók, a 23. és 25. ábrán pedig a bemeneti minta alapján generálódott aktivitások. Az egyes kimenetek fentről lefelé haladva rendre: *cs, c, k, ty, t, p, g, gy, d, b, h, s, sz, f, zs, z, v, l, r, m, n, ny, j, i, é, ü, ö, e, á, a, o, u* hangokhoz tartoznak. A legalsó öt kimenet pedig az öt akusztikai hangosztályhoz tartozik: zárhang; réshang; félmagánhangzók; magánhangzók; szünet.

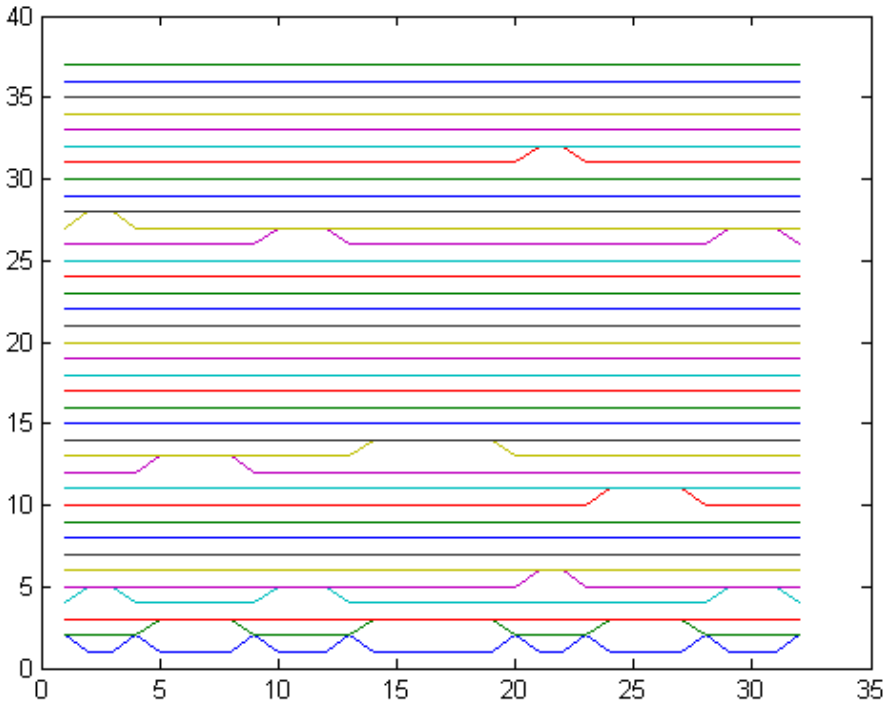
Szembetűnő a különbség a két különböző minősítésű szó között. Az ötösre minősített *hűséges* szó kimenetei sokkal intenzívebbek, mint az egyesre értékelt *vacsora* szó kimenetei.



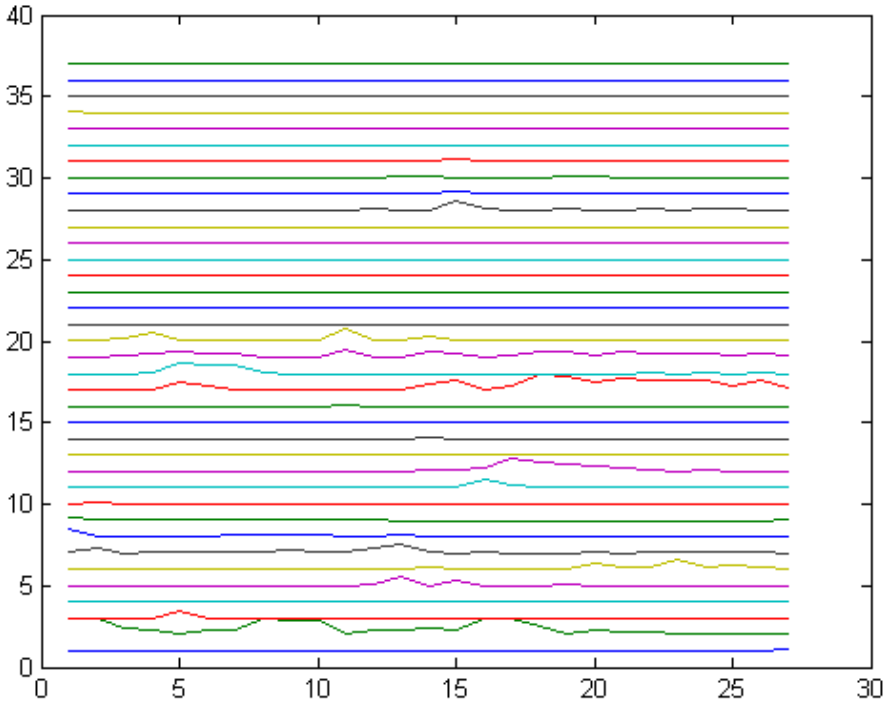
22. ábra *A kimenetek aktivitása a hűséges szó generálásakor*



23. ábra *A kimenetek aktivitása a hűséges szó illesztésekor*



24. ábra *A kimenetek aktivitása a vacsora szó generálásakor*



25. ábra *A kimenetek aktivitása a vacsora szó illesztésekor*

4. táblázat *Az AF szegmentálási eljárás eredményei*

AF szegmentálási eljárás				
	Kezdő		Vég	
Tűrés	Rossz	Jó	Jó	Rossz
0 ms	1785	62	1782	65
20 ms	1747	100	1795	52
40 ms	1667	180	1801	46
60 ms	1500	347	1805	42
80 ms	1340	507	1811	36
100 ms	1187	660	1812	35
200 ms	692	1155	1825	22

5. táblázat *A PLP szegmentálási eljárás eredményei*

PLP szegmentálási eljárás				
	Kezdő		Vég	
Tűrés	Rossz	Jó	Jó	Rossz
0 ms	1327	528	1569	286
20 ms	620	1235	1669	186
40 ms	296	1559	1734	121
60 ms	190	1665	1761	94
80 ms	142	1713	1780	75
100 ms	122	1733	1792	63
200 ms	66	1789	1824	31

6. táblázat *Az ADTW szegmentálási eljárás eredményei*

ADTW szegmentálási eljárás				
	Kezdő		Vég	
Tűrés	Rossz	Jó	Jó	Rossz
0 ms	137	1718	1717	138
20 ms	97	1758	1753	102
40 ms	80	1775	1776	79
60 ms	64	1791	1791	64
80 ms	54	1801	1803	52
100 ms	46	1809	1815	40
200 ms	25	1830	1838	17

7. táblázat *Az egyes szegmentálási eljárások 0 ms-os tűréssel, a hang időintervallumán kívül eső határok száma*

	Kezdő	Vég
	Eltolt	Eltolt
AF szegmentálási eljárás	15	1258
PLP szegmentálási eljárás	135	42
ADTW szegmentálási eljárás	39	77

Az említett 4.-6. táblázatban összefoglalt eredmények alapján az is látható, hogy az AF szegmentálási eljárásnál a szünetek hiánya miatt a hangok eleje jelentős számban helytelen, míg a hangok végének detektálása az 4. táblázat alapján 96%-ban helyes. Ezek alapján az eljárás nem tűnne teljesen célszerűtlennek a gyenge minőségű beszéd szegmentálására, de nem hagyhatjuk figyelmen kívül a 7. táblázat eredményeit, ahol megmutatkozik az eljárás alkalmatlansága. 1258-szor fordult elő, hogy a hang végének időpontját az eljárás hamarabbra tette, mint magának a hangnak a valós kezdete. A PLP szegmentálási eljárás már 0 ms-os tűréshatár mellett is jól teljesít a specializált dinamikus idővetemítéssel egyetemben. A 8. táblázatban az eredmények százalékos alakulása látható mindhárom eljárás esetén, 0 ms-os tűréshatár mellett, figyelembe véve az eltolásokat.

8. táblázat *A szegmentálási eljárások százalékos eredményei*

	Kezdő	Vég
AF szegmentálási eljárás	*2,5 %*	*28,4 %*
PLP szegmentálási eljárás	*21,3 %*	*82,7 %*
ADTW szegmentálási eljárás	*90,9 %*	*88,8 %*

A 8. táblázatban látható százalékok megmutatják, hogy a specializált eljárás a 90,9%- os és 88,8%-os eredményével a PLP lényegkiemelést alkalmazó HMM modelleken alapuló szegmentálási eljárásnál is jobban teljesít. Szembetűnő a hangok kezdetének meghatározásánál a 69,6 százalékos különbség, de a vég meghatározásnál is közel 6 százalékkal jobb eredményt produkál.

Összevetve az eljárások eredményeit és figyelembe véve a 7. táblázat eredményeit az általam megalkotott ADTW eljárás teljesített a legjobban.

4.5 Tézis [S1], [S2], [S10], [S13]

I. *Megvizsgáltam a nemlineáris idővetemítés különböző módjait, és a gyenge minőségű beszédre módosítottam a dinamikus idővetemítés kapcsolódási szabályait, ezzel az általam vizsgált eljárásoknál lényegesen több határérték esett az egyes hangintervallumok belsejébe.*

4.5.1 Újdonság

Jó minőségű beszédre kidolgozott eljárások nem megfelelően működnek nagyon torz és akadozó beszédre. Az eljárás újdonsága a referencia generálás és az alkalmazott kapcsolódási szabályok.

4.5.2 Mérések

A szegmentálás pontosságát különböző mutatók alapján vizsgáltam, és a gyenge minőségű beszédhez nem adaptált rendszereknél pontosabb eredményeket kaptam.

4.5.3 Érvényességi korlátok

Az alkalmazás korlátai: Az alkalmazott neurális hálózat nyelvfüggő, a tanításra használt beszédadatbázis magyar nyelvű. Más nyelvekre a neurális hálózatot be kell tanítani.

4.5.4 Konklúzió

A gyenge minőségű beszédre kidolgozott szegmentálási eljárás alapját képezi az automatikus minősítésnek, hiszen kulcsszerepe van a referencia hang és az elemzés alatt álló hang megfeleltetésében.

5 Hangsúlydetektálás relatív intenzitás alapján

Mivel a siketek nem hallják a saját hangjukat, külön nehézséget okoz a szupraszegmentális jellemzők megtanulása. A beszédminősítés egyik fontos eleme a hangsúly detektálása. A hangsúlyos szótagnál a hangmagasság és a hangerő emelkedik. A szótagon belül általában a magánhangzó a legnagyobb energiájú, ezért az intenzitás mérésekor célszerűnek tűnik a magánhangzó energiájának vizsgálata. Ha megmérjük az egyes magánhangzók átlagos energiáit, egy sok beszélős hosszú hangmintákat tartalmazó adatbázison, kiderül, hogy a magánhangzók igen eltérő átlagos energiával rendelkeznek. A hangsúlydetektálásnál az energia az egyik vizsgált jellemző, de egy hangsúlytalan nagy átlagenergiájú hang (pl. *a, e*) energiája jelentősen meghaladhatja a hangsúlyos gyengébb hangok (pl. *i, u*) pillanatnyi energiáját. A hangsúlydetektálást célzó vizsgálataim során ezért a magánhangzó intenzitását nem egyszerűen az energiával azonosítom, hanem relatív intenzitás értéket határozok meg a magánhangzó átlagenergiájához viszonyítva a pillanatnyi energiát. A módszer eredményességét a hangmagasságot is figyelembe vevő hangsúlydetektálásra betanított neurális hálózattal verifikálom.

5.1 A hangsúly

A beszédkutatás egyik legnehezebb területe a hangsúlyozás. Egységhez kapcsolva megkülönböztetünk szóhangsúlyt, szakaszhangsúlyt (szószerkezetek esetében) és mondathangsúlyt. Mivel a 3. fejezetben bemutatott beszédasszisztens rendszerben a gyakorlást végző személyek szavakat és mondatokat gyakorolnak, elsősorban ezeknek a hangsúlyozási kérdéseivel foglalkoztam.

A hangsúly valamely szó egy szótagjának kiemelése, megkülönböztetése a többi szótagtól. A nyelveknek két csoportját különböztetjük meg hangsúlyozás szempontjából, a kötött (a hangsúly mindig a szó egyértelműen azonosított szótagjára esik) és a kötetlen (pl. az angol és német nyelvekben a hangsúly kötetlen, sőt, az angolban a hangsúly jelentéselkülönítő szerepű is lehet) vagy szabad hangsúlyozású nyelvek csoportját. A magyar nyelv kötött, mivel mindig az első szótagon realizálódik (szerepe tisztán a közlés lényeges elemeinek kiemelésére és a közlés logikai tagolására szorítkozik). Erős érzelmek kifejezésekor a hangsúly a kötött hangsúlyozású nyelvekben is eltolódhat, illetve akár egy szó minden szótagján is megjelenhet. A hangsúly létrehozásában három fő tényező együttesen vagy egyedileg játszik szerepet adott nyelvtől függően, ahol szabályszerűségek figyelhetők meg [32]. Ezek a tényezők:

- az alapfrekvencia kiemelkedése a hangsúlyos szótagon;
- a hangsúlyos szótag nagyobb intenzitással való kiejtése;
- és a hangsúlyos szótag magánhangzójának időtartambeli hosszabbodása.

Összetett szavaknál és jelzős szerkezeteknél fő- és mellékhangsúlyokat is el tudunk különíteni. Magyarban a magánhangzók megnyúlása döntően érzelmeket fejez ki [22].

Több nyelvész véleménye szerint a magyar nyelvben a hangsúly elsősorban nyomatéki, vagyis intenzitástöbbletből ered, azonban Kassai [32], illetve Szaszák [49] tapasztalatai alapján is meghatározó a hangsúlyban az alapfrekvencia szerepe is. A hangerő emelkedése fiziológiai okokból kifolyólag automatikusan maga után vonja az alapfrekvencia emelkedését is, mivel a megnövekedett szubglottális nyomás a hangszalagokat gyorsabb rezgésre készteti [32], így a hangintenzitás és az alapfrekvencia menete között összefüggés fedezhető fel. További probléma, hogy a szegmentális szerkezet jelentősen befolyásolja az intenzitást, míg az alapfrekvencia esetében ez csak a zöngés-zöngétlen különbségtételre igaz. Kassai javaslata szerint érdemes figyelembe venni a hangsúly vizsgálatánál mind az alapfrekvencia, mind az intenzitás, mind az időtartam alakulását, mert a hangsúly sokkal bonyolultabb viszonyban van ezekkel a tényezőkkel, mint az intonáció, amelyet tekinthetünk az alapfrekvencia által meghatározottnak. Olaszy [43] részletesen vizsgálja az alapfrekvencia és a hangsúly kapcsolatát magyar nyelvre, a hangsúlyozáshoz szorosan kapcsolódóan két megállapítást érdemes megjegyezni:

- Kiemelten hangsúlyos szótagon belül (pl. eldöntendő kérdés utolsó előtti szótagjában), a fókuszpozícióban (mondat esetén a leghangsúlyosabb szó és annak pozíciója - magyarban ez jellemzően az ige előtti pozíciót jelenti), a frekvencia meredek esése figyelhető meg. Ez a jelenség egységesen jellemző a magyar nyelvben. Az alapfrekvencia csúcsa minden esetben a magánhangzóban található meg. A meredek esés kétféleképpen valósulhat meg a hangkörnyezettől függően:
    - ha a magánhangzó előtti hang zöngétlen gerjesztésű, akkor a magánhangzóban az alapfrekvencia a csúccsal indít és meredeken esik;
    - ha a magánhangzó előtt zöngés mássalhangzó áll a szótagban, akkor az alapfrekvencia a megelőző hangban magasról indul, majd enyhén tovább emelkedik, csúcsát a magánhangzóban éri el.
- A hangsúlyozás által megkövetelt alapfrekvencia-emelkedés elmarad, ha a nyelvi szerveződésben magasabban elhelyezkedő intonáció ezt megkívánja: gyakori jelenség tagmondatok végén álló rövidebb szavak esetében. Ekkor a szó első szótagján a legalacsonyabb az alapfrekvencia értéke, majd ezután fokozatosan emelkedik, csúcsát az utolsó szótag magánhangzóján éri el.

5.2 Az alapfrekvencia

Az alapfrekvencia (F0) nem más, mint a hangszalagok pillanatnyi rezgésszáma. Az alapfrekvencia csak akkor értelmezhető, ha zöngés (kváziperiodikus) gerjesztés jelen van a beszédben. Az alapfrekvencia mérése összetett feladat és számos algoritmus készült a meghatározása. Ezen módszerek közül csak az általam felhasznált eljárást fogom részletezni.

Az alapfrekvencia detektálás egyik lehetséges és elterjedt módja az autokorrelációs függvény maximumainak meghatározásán alapul [21], [48]. A legjobb illeszkedést a beszédjel eredeti és az önmagához képest eltolt függvénye között akkor kapjuk, ha az eltolás mértéke éppen a periódusidővel egyezik meg. Tehát az autokorrelációs függvény

zöngés beszédszakaszra is majdnem periodikus. Az autokorrelációs függvény periódusideje csúcskereséssel jól meghatározható. Leggyakrabban az autokorrelációs függvény helyett egy rokon függvényt, az átlagos magnitúdó különbség függvényt, AMDF (Average Magnitude Difference Function) használják [21], [48]. Az AMDF függvény a beszédjel alapperiódusának megfelelően nem maximumokat, hanem minimumokat keres/ad. Az AMDF függvény $D_n\,(k)$ például az alábbi összefüggéssel definiálható:

$$D_n\,(k) = \frac{1}{N} \sum_{i=n-N+1}^{n} |\,x_i - x_{i-k}| \qquad (8)$$

Az *x* beszédjel *i* diszkrét időpontbeli értéke x_i, n az az időindex, amelyre az AMDF függvény értékét szeretnénk számítani, *N* pedig az ablakszélesség, amelyre átlagolunk. A minimumokat a k változó szerint keressük.

A beszédjelből kinyert alapfrekvencia-értékeket felhasználásuk előtt előfeldolgozásnak célszerű alávetni. Az előfeldolgozás leggyakoribb célja az alapfrekvencia kontúrjának (görbéjének) simítása az ingadozások eltüntetésére és az alapfrekvencia interpolációja a zöngétlen helyeken, mivel számos esetben nehezítené a feldolgozást, ha az alapfrekvencia-menet szaggatott lenne.

Az alapfrekvencia detektálási lehetőségei:

- Nagy hibák: oktáv vagy még nagyobb tévesztés az alapfrekvenciában. Jellemzően a gyorsan halkuló vagy hangosodó szakaszokon fordul elő, elsősorban szó elején vagy végén.
- Kis hibák: apróbb pontatlanságok az alapfrekvenciában. Jellemzően a vegyes gerjesztésű hangoknál fordul elő (zöngés mássalhangzók).
- Zöngés-zöngétlen tévesztés: jellemzően ez is a vegyes gerjesztésű hangoknál fordul elő.

5.3 Az energia

A beszédjel energiájának számítása a legalapvetőbb jelfeldolgozási műveletek közé tartozik.

$$E_n = \sum_{i=n-N+1}^{n} x_i^2 \qquad (9)$$

Az összenergia gyakran használatos, mint a szupraszegmentális jegyek akusztikai korreláltja. Az energia számításánál kiemelten fontos az átlagoláshoz megválasztott minták száma, másképpen az időablak nagysága.

5.4 Hangsúlydetektálási módszerek

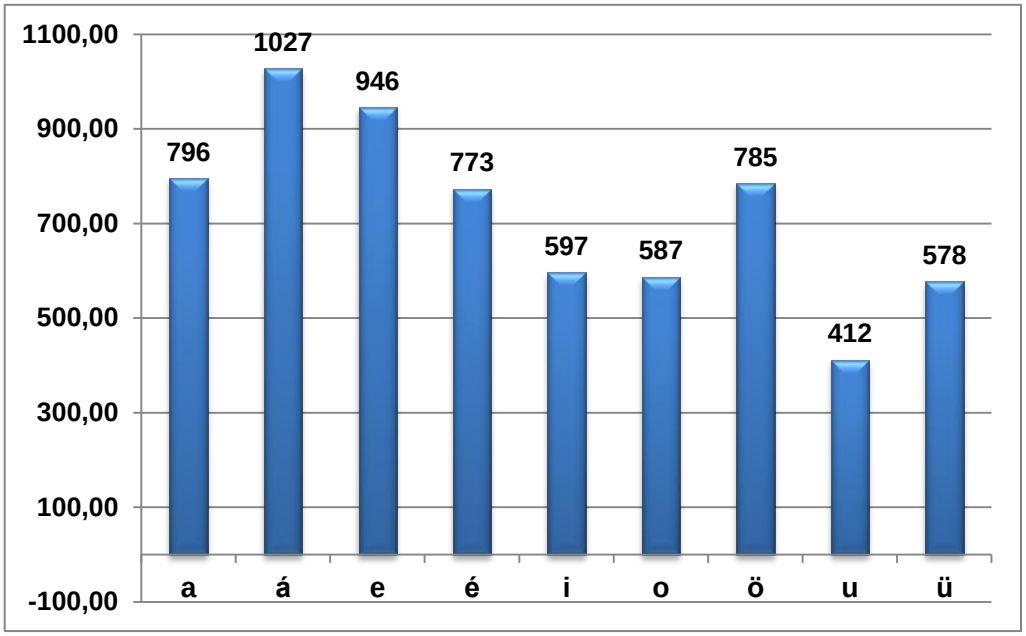
A hangsúly detektálását általában az energia és az alapfrekvencia alapján végzik. Megvizsgálták a hossz, az amplitúdó és a spektrális változások különböző módokon normalizált értékeit [51]. Több esetben mély neurális hálók betanításával és alkalmazásával valósították meg az automatikus hangsúlydetektálást angol nyelvre [41],

[51], [52]. Angol nyelvet tanító szoftver fejlesztése során, amely a nyelvtanulók által produkált hangsúlymintázatokat vizsgálja, és adott esetben kijavítja azokat, egyesek arra az eredményre jutottak, hogy a hangsúly legmegbízhatóbb jelzése a hossz és az amplitúdó információ kombinációja [56]. A magyar nyelvben a hangsúly, illetve áttételesen a hangsúlyt meghatározó akusztikai-prozódiai jellemzők, az alapfrekvencia és az energia alapján lehetséges a szóhatárok detektálása is [53]. Cseh nyelvre a szavak hangsúlyozását detektáló rendszer a beszéd 3 fő jellemzőjét használja fel a felismeréshez:

- a relatív szóhosszúságot;
- a relatív intenzitást;
- és a bemondás alapfrekvenciáját.

A tesztek alapján a legnagyobb hatékonyságot a szavak relatív hosszúságának alkalmazása jelentette [39].

Megfelelő lényegkiemelési módszert választva az egyes hangok átlagenergiája (26. ábra) nagy különbségeket mutat, ezért a kis energiájú hangsúlyos magánhangzók energiája nem éri el a nagyobb energiájú hangsúlytalan magánhangzókét. (A PLP energia logaritmusát használtam a lényegkiemelés egyik jellemzőjeként, ebből exponenciális függvény alkalmazásával számoltam az átlagos abszolútérték összeget.) Módszeremben a magánhangzók pillanatnyi energiáját az adott magánhangzók átlagenergiájához viszonyítom így mutatom ki hangsúlytalan vagy hangsúlyos jellegét.



26. ábra *A magánhangzók átlagos abszolútérték összege PLP lényegkiemelési módszer esetén*

5.5 A felhasznált hangsúlyadatbázis

Magyar nyelvű beszédtechnológiai kutatásokban és az oktatásban is nagy igény mutatkozott egy referenciaként használható, helyes hangsúlycímkéket tartalmazó mondatgyűjteményre. Az első magyar hangsúlyadatbázist Olaszy Gábor és Abari Kálmán alkotta meg [1], [44] a magyar mondatok helyes hangsúlyozási mintázatainak

bemutatása írott és hangzó formában is egyaránt, melyekhez képi megjelenítések is társulnak. A hangsúlyadatbázis 1869 kijelentő mondatot tartalmaz és webes lekérdező felületen keresztül érhető el. A könnyen értelmezhető bináris felépítésű hangsúlymintázat jelentése: az írott mondat szavai elé tett hangsúly jelek sorozata egy mondatra vonatkoztatva az írott formába beágyazva. A hangsúlymintázat jeleinek száma megegyezik a mondatban szereplő szavak számával. Ha *H*=hangsúlyos, akkor az *"Aki tudja, csinálja, aki nem, tanítja, mondta sóvárogva."* mondat hangsúlytérképe: *-HH-HH--*. Vizsgálataimhoz 10 személytől egyenként 50 bemondott mondatot használtam fel. A hangsúlyos és hangsúlytalan bejegyzés nem az aktuális bemondásra vonatkozik, hanem a mondat értelmezése alapján hangsúlyosnak és hangsúlytalannak ítélt szavak kapták a *H* illetve – jelzést. A bemondók nem mindig az elvárt hangsúlyozási minta alapján olvasták fel a mondatokat, így felkértem egy nyelvészeti szakértőt, hogy elemezze a mondatokat egyenként és alkossa meg a mondatok hangsúlyképletét. Az elemzés során a szakértő külön jelölte a hangsúlyos, a félhangsúlyos és hangsúlytalan szavakat. A tanító és tesztelő mondatok kiválasztása véletlenszerű volt. Tesztelő mintáknak a rendelkezésre álló mondatok 25%-t választottam ki véletlenszerűen.

5.6 Az alkalmazott hangsúlydetektálási módszer

Az általam magyar nyelvre létrehozott hangsúlydetektálási módszer több komponensből tevődik össze. A detektálás egyik fő összetevője a relatív intenzitás. A módszer a magánhangzók átlagenergiájához viszonyítja a pillanatnyi energiát. A mondatok elejétől vége felé haladva csökkenő relatív intenzitást tapasztalható. A csökkenő tendencia korrigálására kiegyenlítem a mondatok átlagos amplitúdóját. Az értékeket a magánhangzó közepének 50 milisecundumos (800 minta) környezetére számoltam.

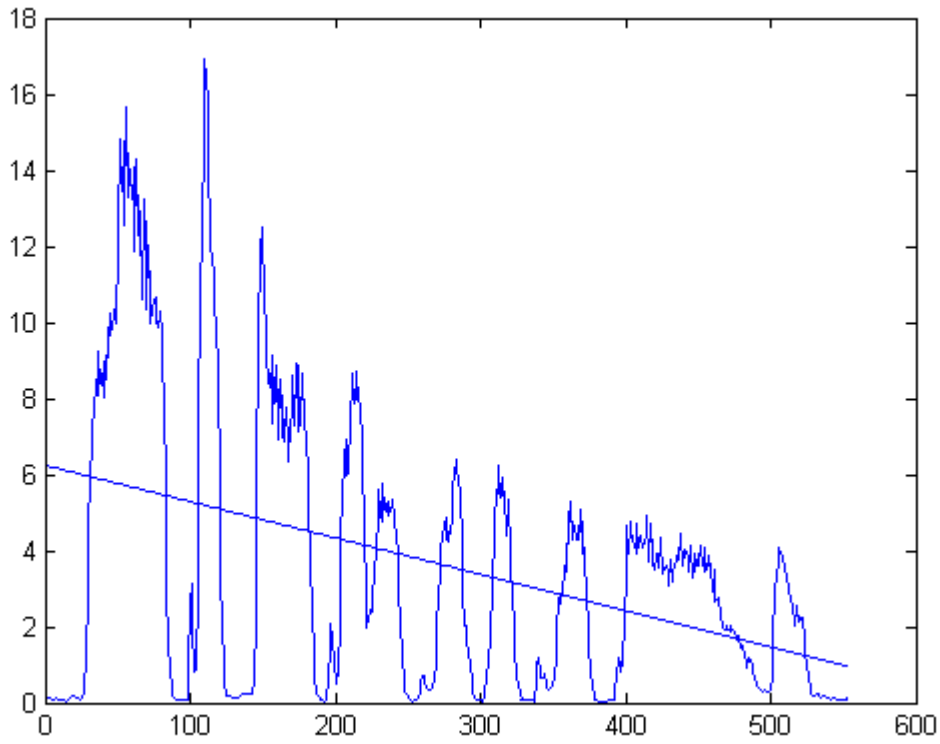
5.6.1 A relatív intenzitás és a kiegyenlítés módszere

A kiegyenlítés alapötlete, hogy a hullámforma abszolút értékére vonatkozó regressziós egyenest állapítok meg, amivel osztom a pillanatnyi amplitúdót. Az abszolút értékeket 10 milisecundumonként összegeztem.

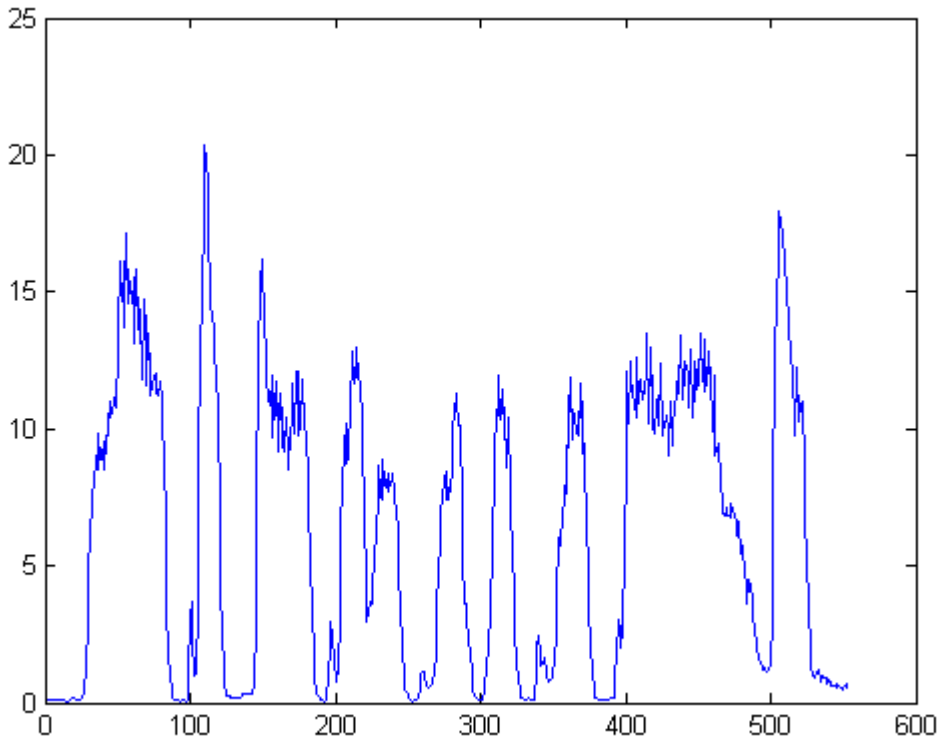
A 27. – 28. ábra a *" A lelkét nem keríthetik hatalmukba."* példamondat kiegyenlítését mutatja be a számolt regressziós egyenes alapján. A 27. ábrán a valós abszolút amplitúdó burkolója, a 28. ábrán a kiegyenlített abszolút amplitúdó burkolója látható.

Az alábbi linken meghallgatható a mondat:

http://mazsola.iit.uni-miskolc.hu/~pinter/a_lelket.wav



27. ábra *A példamondat abszolút amplitúdójának burkolója és a regressziós egyenes*



28. ábra *A példamondat kiegyenlített abszolút amplitúdójának burkolója*

Az energiát az alábbi képlet alapján határozom meg:

$$E_n = log \sqrt{\sum_{i=n-N+1}^{n} x_i^2} \qquad (10)$$

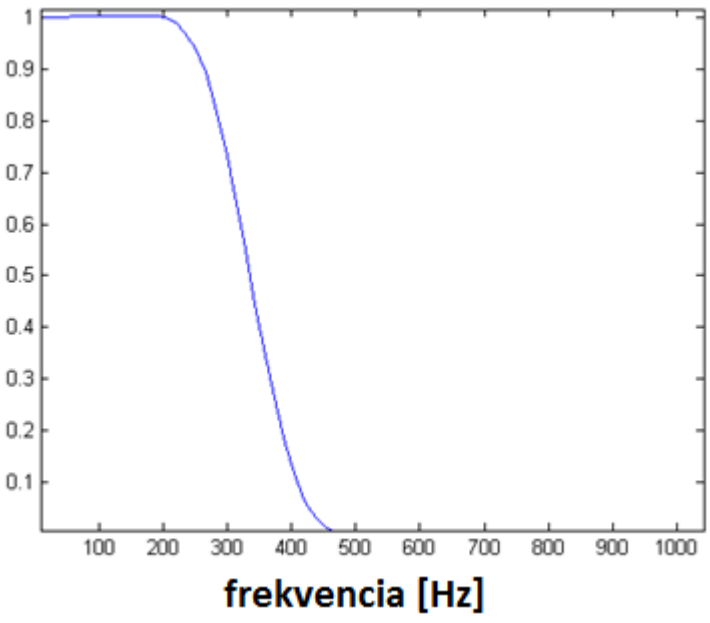
Az értékeket 800 mintánként, azaz 50 milisecundumonként számoltam. A pillanatnyi és az átlag energiát is a (10) képlet alapján számolom, a relatív intenzitást pedig a két logaritmus különbségeként kapom meg.

(Megjegyzendő, hogy a hangsúlyos szótagok pillanatnyi energiájának az átlaga 6,6 dB-el haladja meg a hangsúlytalan szótagok pillanatnyi energiájának átlagát.)

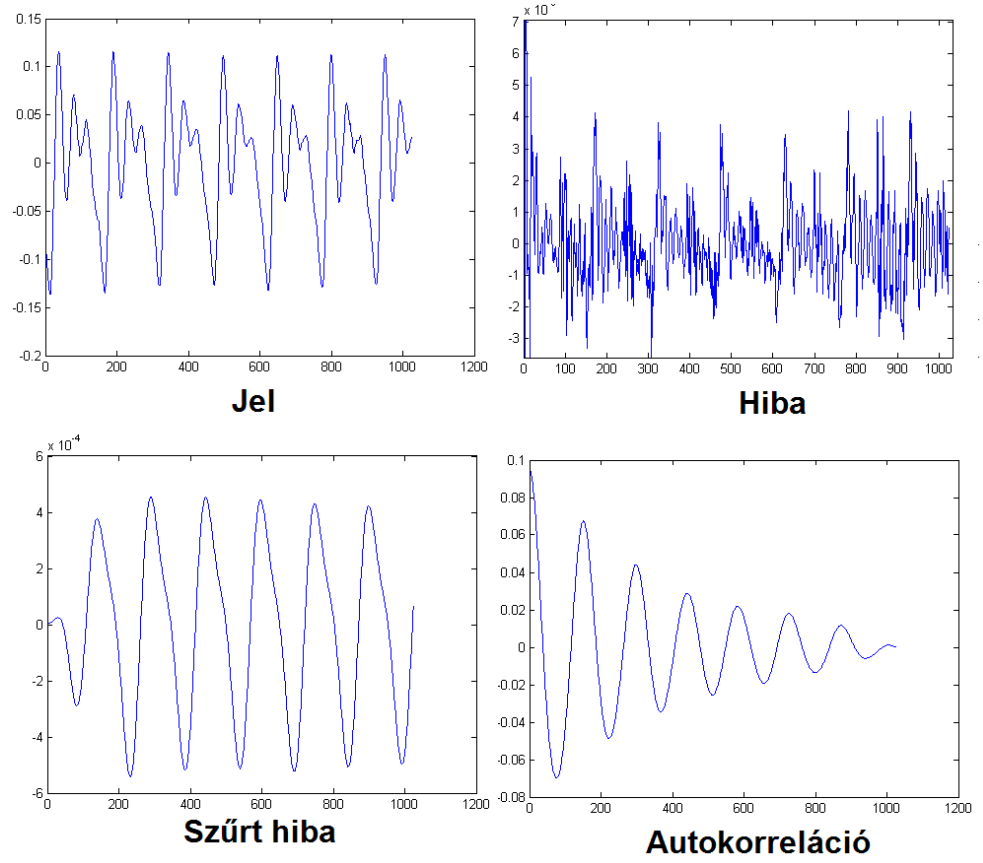
5.6.2 Az alapfrekvencia meghatározása

A másik hangsúlyt befolyásoló jellemző a beszéd alapfrekvenciája. A hangsúlyos szótagoknál az alapfrekvencia megemelkedik. A relatív intenzitás alkalmazásának a pillanatnyi energiájával szemben a verifikálása érdekében neurális hálózatot tanítottam be az alapfrekvencia (F0) és a relatív intenzitás felhasználásával. Az alapfrekvencia meghatározásához kipróbáltam az ismert beszédelemző alkalmazásokat (Praat, Wavesurfer, Opensmile). Egyes magánhangzókat – különösen a mondatvégi mélyebb és halkabb hangokat – ezek a rendszerek gyakran zöngétlennek mutatták. Az F0 alapján a hangsúlyos szótagok felismerése nem volt elég sikeres. Ezért saját alapfrekvencia meghatározó algoritmust fejlesztettem ki.

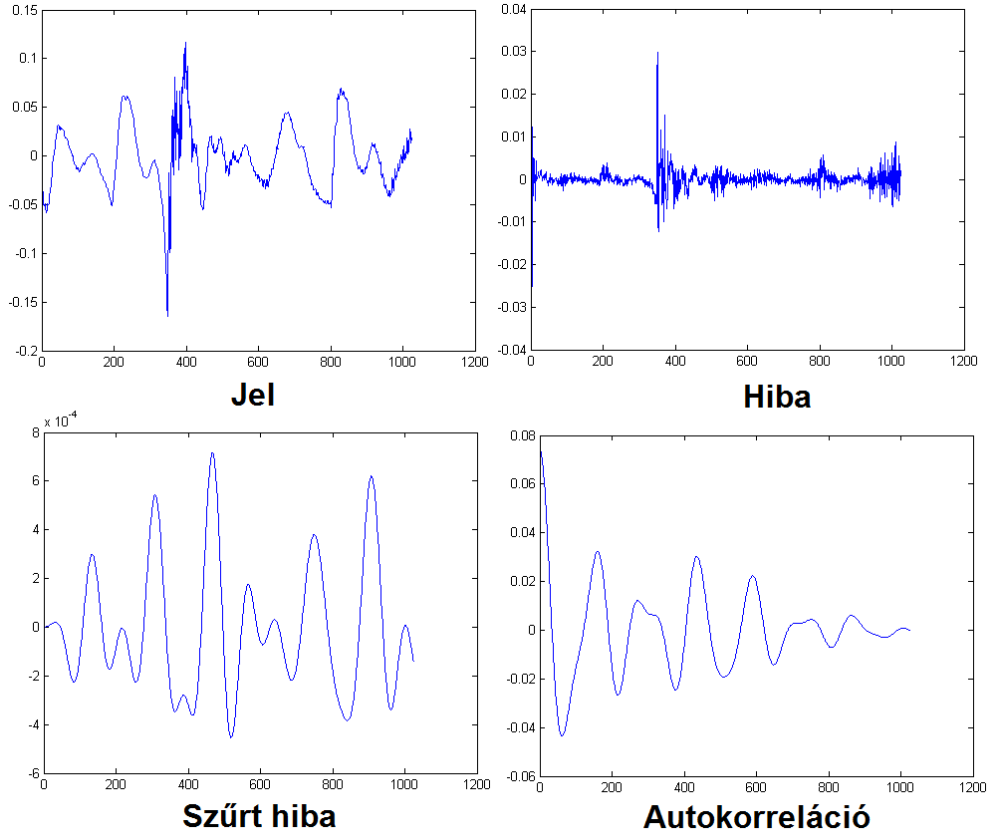
Ismert tény, hogy a lineáris predikció hibája a zöngés hangoknál a periódus elején kiugróan magas. A periódus kezdetén a predikciós hiba erős aluláteresztős szűrésével a zavaró hibák csökkenthetők. Az alkalmazott aluláteresztő szűrő karakterisztikája látható a 29. ábrán.



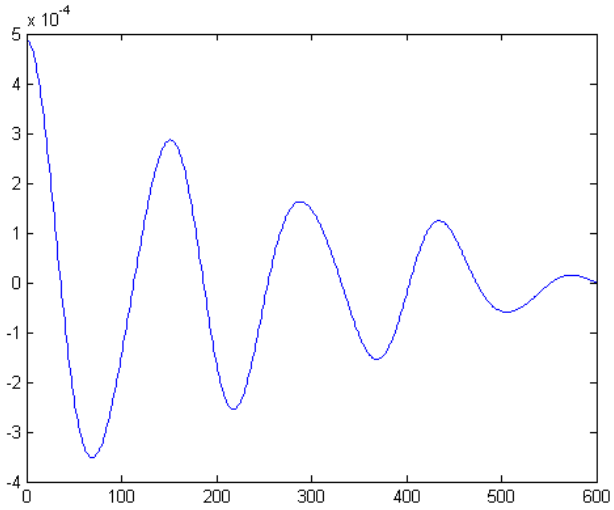
29. ábra *Aluláteresztő szűrő karakterisztikája*



30. ábra *Predikciós hiba szűrése autokorrelációs függvénnyel I.*



31. ábra *Predikciós hiba szűrése autokorrelációs függvénnyel II.*



32. ábra *Az autokorrelációs függvény autokorrelációs függvénye*

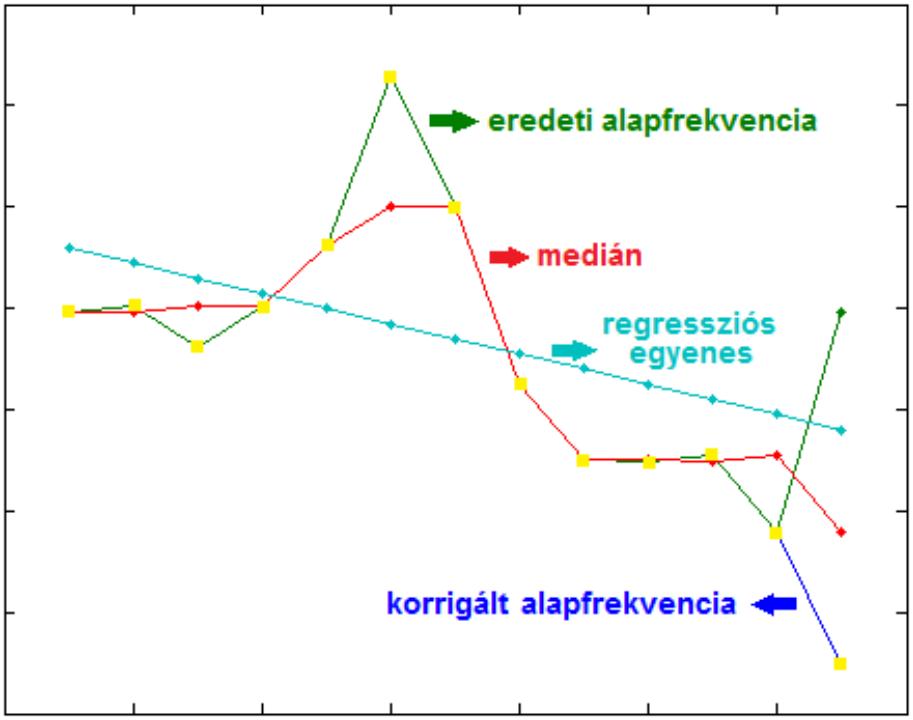
A 31. ábrán a 30. ábrán szereplő jellel szemben az autokorrelációs függvény lokális maximumai nem mutatnak meggyőző csúcsot. A periodikus jelleg és periódus idő kiemelésére újabb autokorrelációs számítást végezhetünk. Az autokorrelációs függvény autokorrelációs függvénye (32. ábra) jelelméleti szempontból nehezen értelmezhető, de kizárólag a periódusosság kiemelésére alkalmasnak tűnik.

A periódus idő (a minták száma $*$ mintavételi idő) az autokorrelációs függvény maximumának kijelölésével meghatározható.

A hang rekedtessé válása vagy az autokorrelációs függvényben egy felharmonikusnál jelentkező maximum oktávugrást okozhat az alapfrekvenciában. Ezt a hibát a mondat magánhangzóinak alapfrekvenciáit elemezve oktávszűréssel korrigálom. Az alapfrekvencia becslésére három érték átlagát használom fel:

1. az előző magánhangzó alapfrekvenciáját;
2. a mondat F0 menetének regressziós egyenesét;
3. az F0 mediánszűrését.

Amennyiben ehhez a becsült értékhez közelebb áll az adott magánhangzóra kapott alapfrekvencia fele vagy kétszerese, az F0 értékét az adott magánhangzóra a közelebb álló értékre változtatom. A 33. ábrán egy példamondat alapfrekvencia meghatározása látható oktávszűrést alkalmazva. Zöld színnel az eredeti alapfrekvenciát jelöltem, pirossal a medián értékeket, világoskékkel az F0 menetének regressziós egyenesét. A becslés végeredménye a sötétkék kiemelt jelzésű korrigált alapfrekvencia.



33. ábra *Az alapfrekvencia korrigálása oktávszűréssel*

5.7 Eredmények

A relatív energia és az alapfrekvencia felhasználásával neurális hálózatot tanítottam be, a hangadatbázis mondatainak megosztásával (5.5 fejezet) a tanítás a szakértői hangsúlyképlet alapján történt. Egy szótagot:

- relatív energiájával;
- alapfrekvenciájával;
- az aktuális szótag és a következő szótag aktuális energiájának különbségével (az utolsó szótagnál 0);
- az aktuális szótag és az előző szótag aktuális energiájának különbségével (az első szótagnál 0);
- az aktuális szótag és a következő szótag aktuális alapfrekvenciájának különbségével (az utolsó szótagnál 0);
- az aktuális szótag és az előző szótag aktuális alapfrekvenciájának különbségével (az első szótagnál 0) jellemeztem.

A tanításhoz és a teszteléshez az itt felsorolt 6 jellemző aktuális, az előző és következő szótagjának konkatenált 18 jellemzőjét használtam. Az első szótagnál a megelőző szótag jellemzőit nullával helyettesítem, az utolsó szótagnál a következő szótag jellemzőit helyettesítem nullával, hogy minden esetben 18 jellemzőt kapjak.

Az egyes mondatokhoz tartozó hangsúlyképletet, a tanító minták esetén úgy alkalmaztam, hogy ha a képlet alapján egy szó hangsúlyos, akkor az adott szó első szótagjának súlyozása 1 a többi szótagjának és a hangsúlytalan szavak összes szótagjának a súlyozása 0. A mondatokhoz szegmentált anyag is tartozik, így adott volt

az egyes hangok időzítése, ami alapján számolhatóak voltak az alapfrekvencia és relatív intenzitás értékek.

A tesztelés során a hangsúlymintázatok referencia szerepet töltöttek be. A tanítási folyamat után a tesztelésre kiválasztott mondatok eredményeit a hangsúlymintázataikkal vetettem össze, amikre a Pearson-korrelációs eredmény 60,1%-ra adódott. Amennyiben a relatív intenzitást a pillanatnyi energiával helyettesítettem, a korreláció 54,6%-ra csökkent [30].

5.8 Tézis [S1], [S10], [S11], [S12]

II. *Megvizsgáltam a hangsúlydetektáláshoz használt jellemzők hatékonyságát. Megállapítottam, hogy ha a magánhangzó pillanatnyi energiája helyett a relatív intenzitását használtam, mintegy 10%-kal nagyobb korrelációt értem el a vizsgált adatbázison a mondatok hangsúlyképletéhez viszonyítva.*

5.8.1 Újdonság

A hangsúlyadatbázisban az adatbázis megalkotói a mondat értelme alapján jelölték ki a hangsúlyos szavakat. Ezek általában nem esnek egybe a felolvasó hangsúlyozásával. Ugyannak a mondatnak a több bemondó által felolvasott változataihoz egyénileg határoztam meg a hangsúlyokat.

5.8.2 Mérések

A hangsúlydetektálás hatékonyságát a hangsúlyképlettel vetettem össze és Pearson-korrelációt számoltam. A hangsúlydetektálásánál általában használt pillanatnyi energiával szemben a magánhangzók relatív energiáját használva érdemi hatékonyság javulást értem el.

5.8.3 Érvényességi korlátok

A betanítás mindössze 10 beszélővel és beszélőnként 50 mondattal történt, bővebb adatbázis használatával a beszélő függetlenség és a hatékonyság javítható.

5.8.4 Következtetések

A szótag (magánhangzó) intenzitásának vizsgálatánál érdemes figyelembe venni az adott magánhangzó átlagos energiáját és ehhez viszonyítani az aktuális energiát, ezzel elérhető, hogy a kis átlagenergiájú hangsúlyos hangok is megfelelő nyomatékot kapjanak.

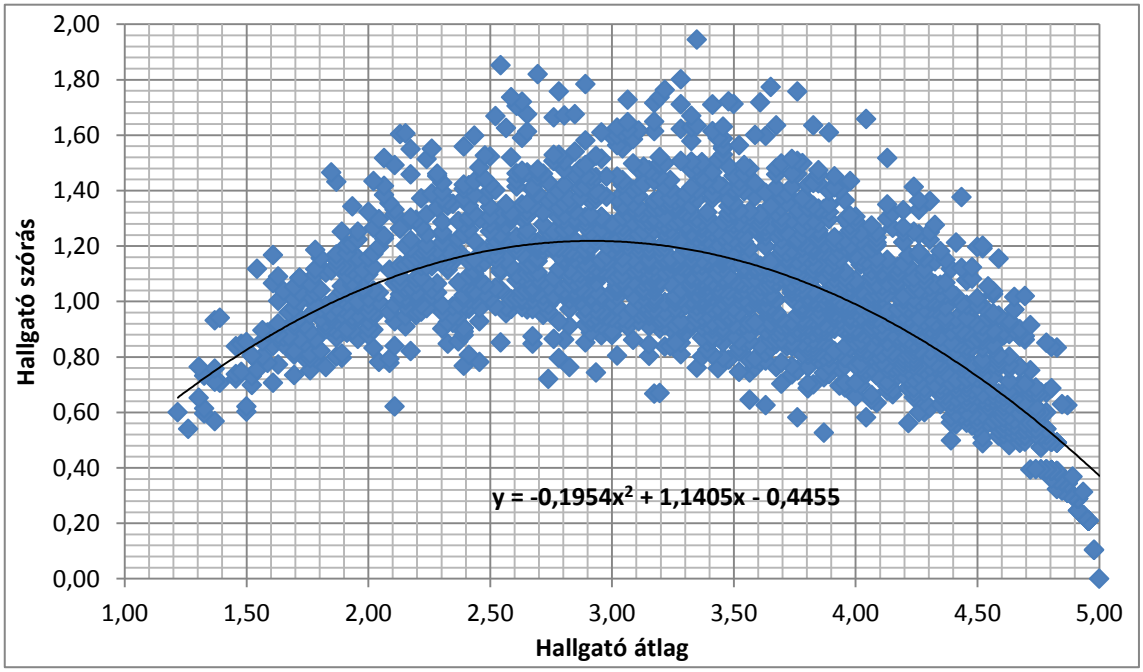
6 A minősítési skála megalkotása

Az automatikus minősítés megvalósításának következő lépése a minősítési skála definiálása. Ehhez elengedhetetlen az 3.3 fejezetben ismertetett hangadatbázis, aminek mintái a beszédprodukció különböző fokán álló hallássérült diákoktól kerültek rögzítésre és az érthetetlentől az észrevehetetlen kiejtési hibáig átfogják a beszédminőség különböző szintjeit.

A szurdopedagógusok szakmai szempontok szerint értékelték a bemondásokat. Nem szakértő (naiv) egyetemi hallgatók a hétköznapi nyelvhasználat szempontjából pontozták ugyanezeket a bemondásokat. Ezek a pontszámok képezik az egyes bemondások szubjektív minősítését. Az egyes szavak automatikus minősítésének célja, hogy a szubjektív értékelés eredményét minél jobban megközelítse.

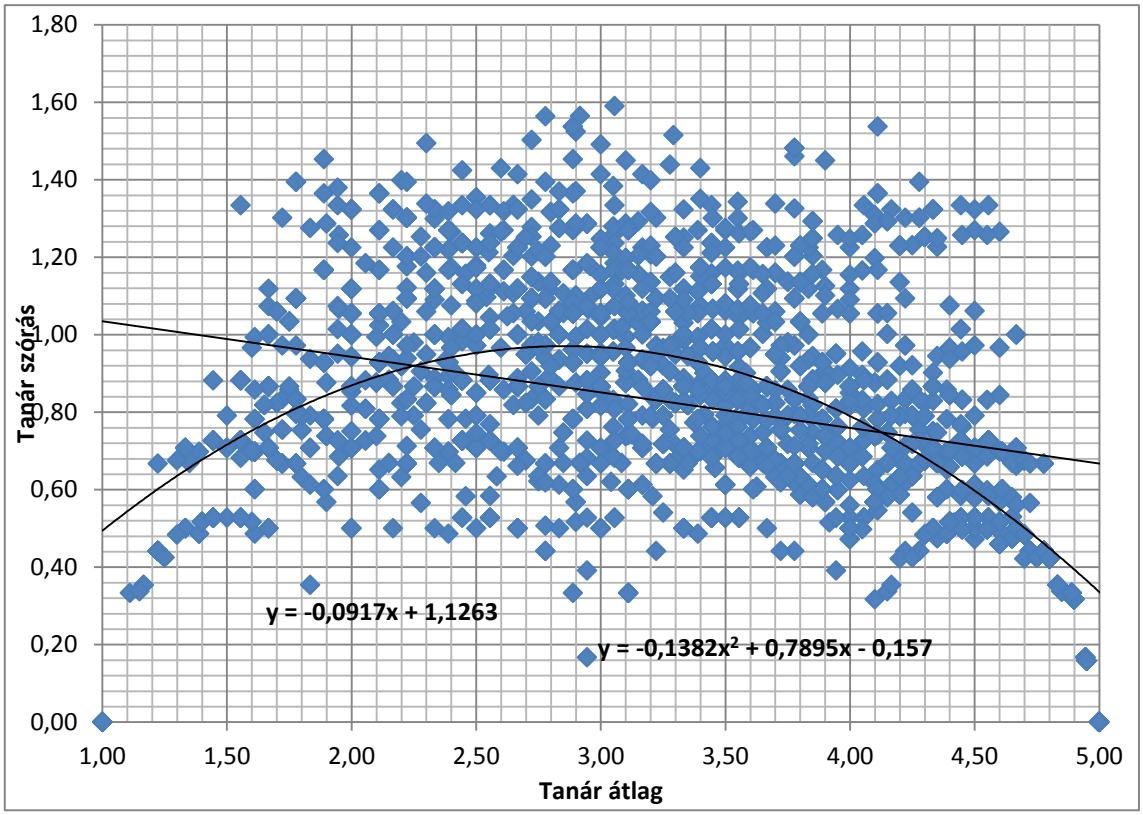
6.1 Hangadatbázis elemzése

Elvégeztem néhány elemzést és kiértékelést a teljes adatbázison, kiszámoltam a szórásokat és a minősítések átlagát. A 34. ábrán a hallgató átlag – hallgató szórás diagram szerepel. Jól látható, hogy a gyenge és a nagyon jó érthetőségű szavaknál kisebb a szórás. Érdekes, hogy itt a feltételezett lineáris trendvonal helyett a másodfokú bizonyult jobbnak.



34. ábra *A hallgatók értékelésének átlaga és szórása*

A tanár átlag – tanár szórás diagram (35. ábra) kiegyensúlyozottabb eloszlást mutat. Ez valószínűleg a szakmai hozzáértéssel magyarázható. Itt a szórások is szűkebb intervallumból veszik értékeiket.



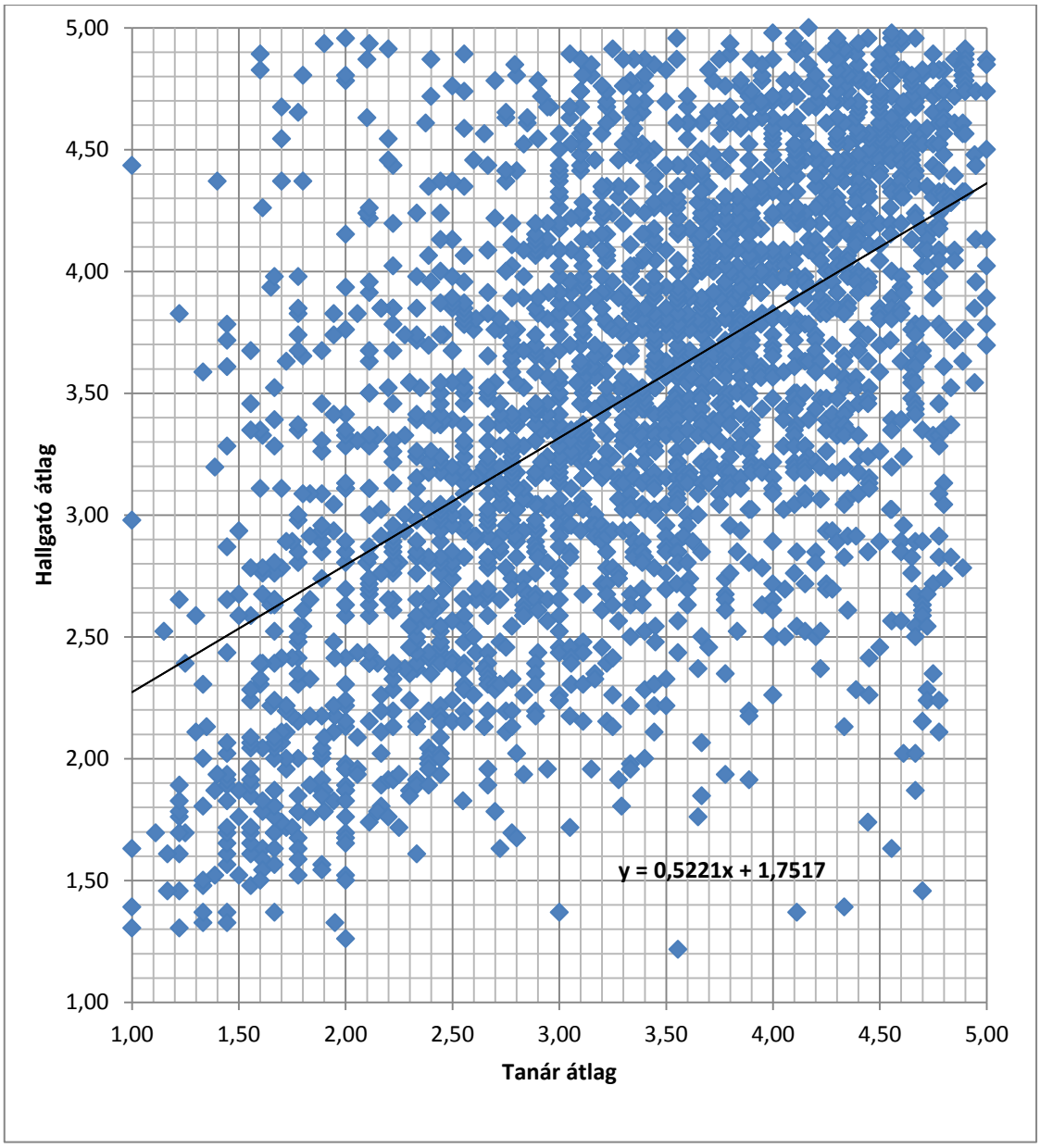
35. ábra *A tanárok értékelésének átlaga és szórása*

A tanár átlag – hallgató átlag diagram a 36. ábrán, a „gyenge” érthetőségi intervallumban mutat nagyobb eltéréseket. Összességében az eredményekből jól látható, hogy noha ugyanazon skála alapján pontozták a hallássérült gyerekek szavait a hallgatók és a pedagógusok, mégis saját csoportjaikon belül is rendkívül nagy szórást mutat ugyanannak a szónak a szubjektív megítélése. A pedagógusok értékelésének átlagos szórása 0,82, a hallgatók értékeléseinek pedig 1,01.

A pedagógusok és a hallgatók átlag pontszáma is nagy eltéréseket mutat. A 9. táblázatban a 298 szó megoszlása látható az lapján, hogy mekkora különbség van a hallgatói és a pedagógusi értékelések átlagai között.

9. táblázat *A pedagógusi és hallgatói értékelések átlagának különbségei*

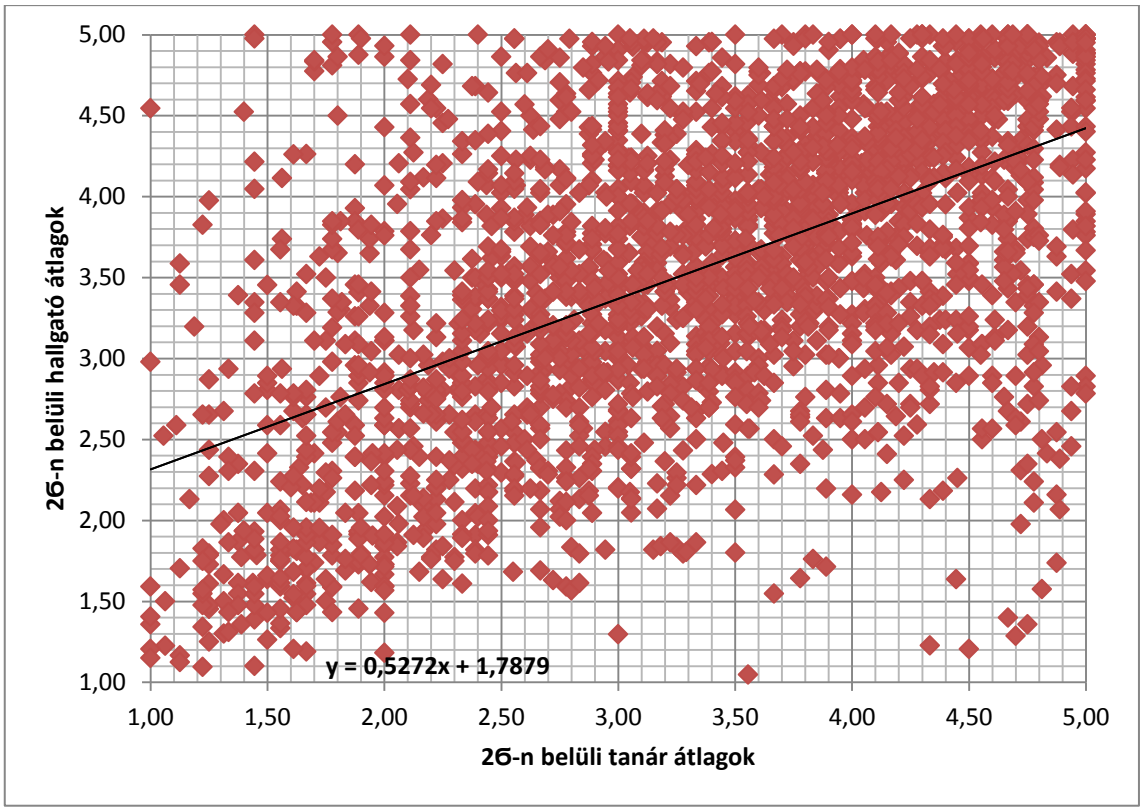
	Átlagok eltérésének nagysága
<= 0,1	*46*
<= 0,2	*84*
<= 0,5	*186*
<= 1	*269*
<= 1,5	*296*
<= 2	*298*



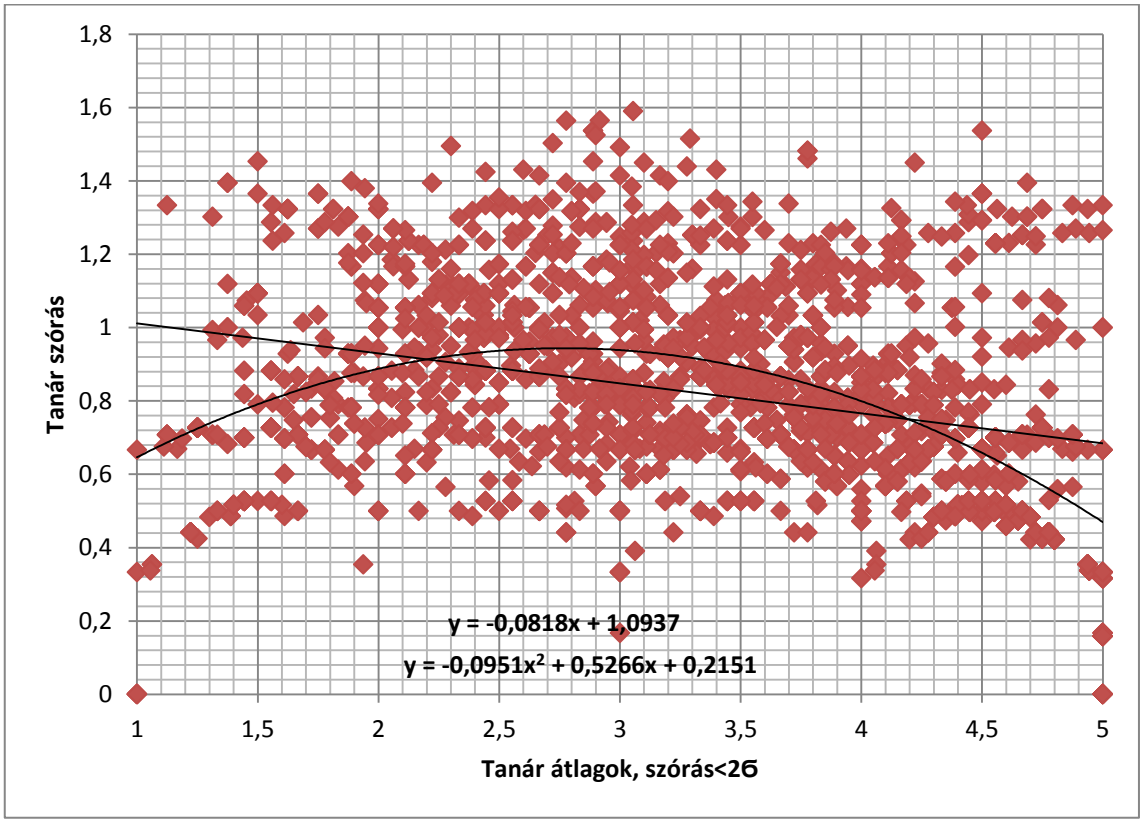
36. ábra *A tanári és a hallgatói átlagok ábrázolása*

Mivel az eredmények ilyen nagy szórást mutattak a mintákhoz definiáltam egy megbízhatósági tartományt és azokat a mintákat elhagytam, amelyek ezen a tartományon kívül estek remélve, hogy kisebb eltérés mutatkozik majd és ezáltal jobban értelmezhetőek lesznek az eredmények. A 2σ megbízhatósági tartomány nem más, mint a szórás kétszeresének tartománya. A 37.-39. ábrán a szűkített tartományon végzett elemzések láthatók, amik sajnos hasonlóan az előzőekhez nagy eltéréseket mutatnak.

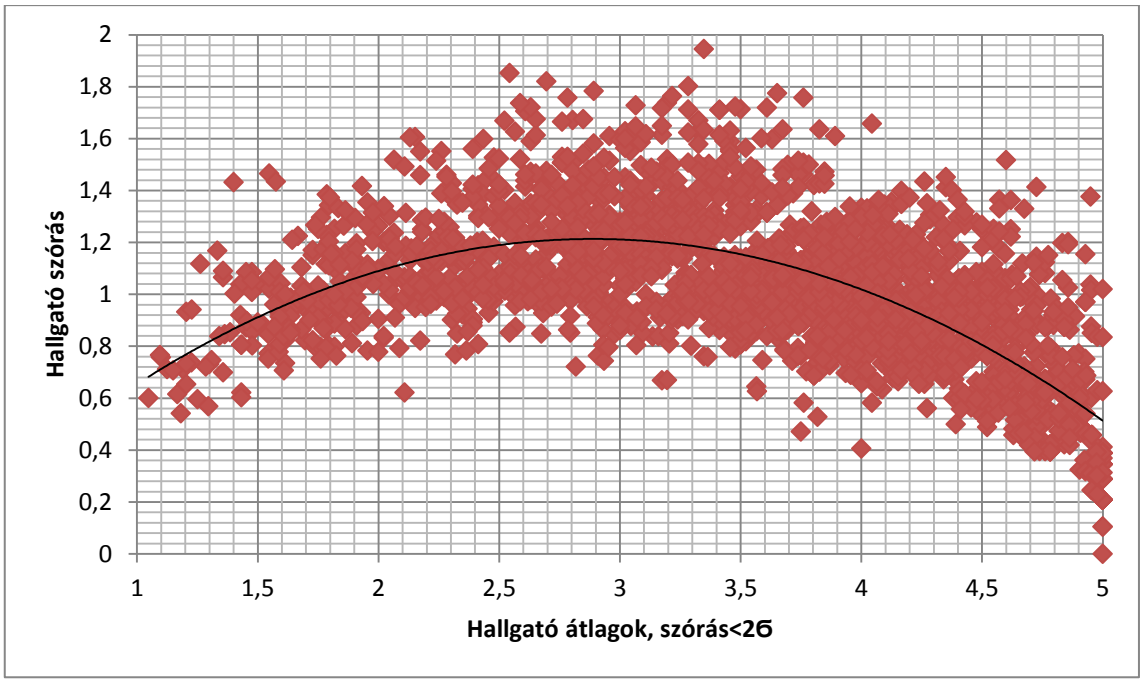
Összességében megállapítható, hogy az értékelések alapján a beszéd minőségének megítélésére nem fogalmazható meg olyan kritérium, amely alapján a minősítés egyértelmű lenne, ezért felkértem egy szakértőt, akinek a szakterülete a beszédfeldolgozás, hogy a minősített szavak egy szűkített csoportját elemezze szakmai szemszögből. A már korábbi tesztelésekre is kiválasztott 300 szó szakértői minősítését a következő fejezet taglalja.



37. ábra *A 2 σ belüli tanári és hallgatói átlagok ábrázolása*



***38.* ábra** *A tanárok 2σ belüli értékelésének átlaga és szórása*



39. ábra *A hallgatók 2σ belüli értékelésének átlaga és szórása*

6.2 A szakértői elemzés

Az kiértékelést végző szakembert megkértük, hogy elemezzen egy 300 szóból álló mintahalmazt. Az eredmények hitelességét a beszédfeldolgozás területén szerzett több évtizedes szakmai tapasztalata támasztja alá, és munkája során már korábban is dolgozott együtt hallássérültekkel.

A felkért szakértő célja az volt, hogy megbecsülje a szubjektív értékeléseket. Azt feltételezte, hogy a beszéd 5 fő tényezőjének együttese alapján születik meg a megítélt pontszám. Ezek a tényezők:

- a beszédtempó;
- a ritmus;
- a hangsúly;
- a dallam;
- és a hanghibák.

Tesztelései során a hanghiba és a ritmushiba súlyát próbálta meghatározni. A jó kiejtés referenciájaként a szavak PROFIVOX (szövegfelolvasó technológia) rendszerrel generált hanganyagát használta fel [42]. Elkészítette a már korábbi fejezetben részletesebben bemutatott 300 szó hanghullám fájljait ugyanazon paraméterekkel, mint amilyeneket a felcímkézésre kapott, majd ugyanúgy felcímkézte azokat, mint az eredeti felvételeken. A minták természetesen arányosan reprezentálták a nagyon jótól a nagyon rosszig a 2421 szóból álló hanganyagot. A címkefájlok adatait az eredeti és a szintetizált kimondásoknál átemelte egy-egy táblázatba szavanként. Ezekben kigyűjtötte az abszolút és relatív szóidőket, hangidőket minden szóra. A szakértő azt tapasztalta, hogy PROFIVOX mindig gyorsabban beszélt, ezért a teljes szóidő a szintetikusnál mindig meghaladta a tanulók kimondását.

6.3 A minősítési skála meghatározása

A legkisebb négyzetes hibát a hanghiba és a ritmushiba figyelembevételével a hanghiba súlyozó együtthatójára -2,78 és ritmushibáéra -0,51 adódott. Lineáris illesztést végeztem a szubjektív eredmények átlagára és a hanghiba és ritmushiba együttes alkalmazására, így az optimális együtthatók:

$$y = ax + bz + c, ahol\ a = -2{,}78\ \ b = -0{,}51\ \ c = 4{,}25 \qquad (11)$$

A negatív szorzók azt fejezik ki, hogy minél nagyobbak a hibák, annál gyengébb a minőség. A tapasztalatok és az eredmények azt mutatják, hogy hanghibára jóval érzékenyebb a szubjektív értékelő, mint a ritmushibára. Ha feltételezzük, hogy a ritmushiba és a hanghiba egyformán jól fejezi ki a vonatkozó hiba mértékét, akkor helytálló az a megállapítás, hogy a hanghiba 5-ször fajsúlyosabb, mint a ritmushiba.

6.4 Tézis [S1], [S2], [S3]

III. *Megvizsgáltam a hanghiba és ritmushiba hatását a szubjektív értékelés eredményeire. Meghatároztam a hanghiba és ritmushiba súlyozó együtthatóit az optimális lineáris illesztéshez.*

6.4.1 Újdonság

Hallássérült gyerekek beszédének elemzéséhez ilyen méretű adatbázist és ilyen részletes elemzést nem ismerek. A pedagógusok által meghatározott szakmai szempontok és a mérhető jellemzők összekapcsolására nem találtam irodalmi utalást.

6.4.2 Mérések

Gradiens módszerrel megállapítottam az optimális együtthatókat.

6.4.3 Érvényességi korlátok

A hangadatbázisban a nagyon rossz minőségű beszédre kevesebb példa található, mint a jobb minőségűekre. Bővítve a kevésbé érthető szavak listáját a megállapítás pontosítható.

6.4.4 Következtetések

Az automatikus minősítés kiindulópontja lehet a hanghibára és a ritmushibára vonatkozó súlyozó együttható.

7 Automatikus értékelés megalkotása

A 3. fejezetben bemutatott beszédasszisztens rendszer egyik szolgáltatása az automatikus minősítés és visszajelzés. A szolgáltatás célja, hogy a hallássérült diákok önállóan gyakorolhassák a mintaszavak kimondását. A 3.2. fejezet elején részletezett módon a tanulás során a referencia kiejtést a szerver vagy a tanár produkálja, a diák pedig ezt igyekszik utánozni az ő aktuális bemondásával. Az értékelés nyilvánvalóan a korábbi eredményekkel összevetve alakítható ki, hiszen ugyanaz a kiejtés egyik tanulónál siker, a másiknál kudarc lehet. Az automatikus értékelés verifikálása érthetőség vizsgálattal történhet.

Amikor egy hang kiejtését vizsgáljuk, az merül fel, hogy hasonlítsuk össze egy referenciával és egy távolság függvény alapján értékeljük a hang megfelelését a referencia bemondáshoz képest. Megvizsgáltam szokványos lényegkiemelési eljárásokat (MFCC, PLP, BARK) a hallássérült gyerekek bemondásainak elemzésére. A szegmentálási adatok alapján kijelöltem a hangok stacionárius szakaszát (amennyiben értelmezhető) és elvégeztem a lényegkiemelést. A stacionárius szakaszt a hanghoz tartozó időintervallum közepére helyeztem. Az aktuális jellemző vektorokat a 6.1. fejezetben kifejtett teljes adatbázisra kiszámított jellemző vektorok átlagával vetettem össze. Így minden hangra kaptam egy távolság értéket. A szó jellemzésére a hangokra kapott távolság értékek átlagát vettem. A szavakra kapott átlagokat a szubjektív értékelés alapján kapott a csoportba tartozó szavakra átlagoltam.

10. táblázat *A különböző lényegkiemelési módszerekkel számított hangokra adott távolságértékek átlaga a minősítési intervallumokra*

	MFCC	PLP	BARK	NN
[1-2]	*0,9961*	*0,2603*	*0,201*	*0,209*
[2-3]	*0,9978*	*0,2555*	*0,2053*	*0,29*
[3-4]	*0,9989*	*0,2368*	*0,1875*	*0,425*
[4-5]	*0,9971*	*0,2458*	*0,1796*	*0,521*

A táblázatból látható, hogy a kapott távolságok nem követik következetesen az osztályok monotonitását. Nincsenek monoton összefüggésben az osztályok minősítésével, nem következetesek. MFCC ezredekben tér el és nem is teljesül a monotonitás.

A több mint 300 beszélős 4,5 órás hangadatbázissal betanított HMM modell HTK implementációjából kiolvashatók, hogy egy adott hangot milyen valószínűséggel generál a hozzátartozó HMM modell. A hallássérült gyerekek bemondásainak felismerési eredményeiből kiolvasott valószínűségek és a szó szubjektív értékelése között nem fedeztem fel korrelációt. Ennek oka lehet, hogy a felismerő szegmentálása nem volt megfelelő.

A jellemzők átlagához viszonyított távolság elemzés sikertelensége után ismét a már betanított neurális hálózat kimeneti aktivitásait kezdtem el vizsgálni. A neurális hálózat

ideális esetben az adott hangra egységnyi, merőben eltérő hangra 0 kimeneti aktivitással válaszol. A helyesen artikulált hangra nagy, a hibásan artikulált hangra kis kimeneti aktivitást produkál. A hasonlósági mértéket az adott hanghoz tartozó outputtal azonosítom. Megvizsgáltam a tanulmányozott szavak egyes hangjaihoz tartozó kimenetek átlagát. A szubjektív tesztek minősítésével vizsgált Pearson-korreláció 55,3% lett az átlagra. A szakértői hanghiba és a szubjektív pontok korrelációja -0,7515. A szubjektív pontok és a ritmus hiba korrelációja -0,2648.

A szubjektív tesztekkel az összehasonlítást részben korrelációs részben a számított pontszámok különbsége szerint vizsgáltam. Az összehasonlításhoz lineáris illesztést végeztem a szubjektív tesztek pontjai és a hasonlóságmérték között. A hasonlóság 0 és 1 közötti kimeneteket produkál, a szubjektív tesztek pontjai 1 és 5 közé esnek. Gradiens módszerrel megkeresem azt a szorzót és eltolást, amellyel a hasonlóság értéket korrigálva a szubjektív pontszámokkal a legkisebb négyzetes hibát adja. Az automatikus minősítés jóságát a szakértői értékeléssel vetem össze. A szakértői értékelés a hanghibát és a ritmushibát jelölte meg a minősítés alapjaként. A hanghibára és a ritmushibára megállapítom az optimális együtthatókat az $y = ax + bz + c$ kifejezésben, ahol x a hanghiba, z a ritmushiba 0 és 1 között értelmezett az értéke. Az előző 9. fejezetből adódóan a=-2,78 b=-0,51. Mivel ezek a hibák annál nagyobbak minél gyengébb minőségű az artikuláció az a és b szorzók negatívra adódnak. A legkisebb négyzetes hibát eredményező együtthatók meghatározása után megvizsgáltam, hogy a szakértői minősítés korrigált pontszámai mennyiben térnek el a szubjektív minősítés eredményétől. A részletes elemzés alá vetett 294 szó közül megvizsgáltam, hogy hány szó pontszám különbsége hány szónál kisebb a 11. táblázat oszlopaiban szereplő értékeknél. Az eredmények átlagosan 88,5 százalékos egyezőséget mutatnak.

Mivel a szubjektív pontszámok elég nagy szórást mutatnak, azt is megvizsgáltam, hogy a szakértői és az automatikus minősítés pontszáma hány szónál esik a hallgatói és a pedagógusi pontszám átlagok közé.

11. táblázat *A szakértői és az automatikus pontszámok referenciához mért különbsége intervallumokra bontva*

	<=0,1	<=0,2	<=0,5	<=1	<=1,5	<=2
Szakértői	*21*	*44*	*131*	*253*	*285*	*291*
Automatikus	*21*	*35*	*101*	*207*	*268*	*287*
(Automatikus/Szakértői)	*100%*	*80%*	*77%*	*82%*	*94%*	*99%*

A szakértői hanghiba és a ritmushiba optimális illesztésekor a 294 szóból a hallgatói és a pedagógusi átlagok közé 54 szó esik. Ugyanez a neurális hálózatok kimeneteinek szavankénti átlagnál és a ritmushibánál, vagyis az automatikus minősítésnél 44 szó.

Az automatikus minősítéshez használt átlagos neurális hálózat kimeneti aktivitás és a ritmushiba együtthatói a legkisebb négyzetes hiba esetén: 2,92 és -0,76. Az együtthatókból megállapítható, hogy a hanghibánál kevésbé megbízható neurális hálózat outputtal párosítva a ritmushiba nagyobb súlyozó együtthatót kap.

7.1 Tézis [S1], [S2], [S3]

IV. *Több módszert megvizsgáltam a hallássérült gyerekek hangfelvételeinek elemzésére. Csak a hangfelismerésre betanított neurális hálózatok kimeneti aktivitására találtam differenciált és monoton eredményeket a különböző minőségi osztályokra. Módszeremmel a szubjektív értékelést a tolerancia tartományokban a szakértői becsléshez képest átlagosan 88,5 százalékos pontossággal közelítettem meg.*

7.1.1 Újdonság

Nem találtam irodalmi utalást arra, hogy a beszédfelismerésre betanított neurális hálózatok kimeneti aktivitását beszédminőség becslésre használták volna.

7.1.2 Mérések

Az automatikus értékelés eredményét a szakértői minősítéssel vetettem össze. Referenciaként a szubjektív értékelés pontszámait használtam.

7.1.3 Érvényességi korlátok

A betanított neurális hálózat nyelvfüggő a szegmentálás kritikus eleme az eljárásnak.

7.1.4 Következtetések

Az automatikus minősítés alkalmasnak látszik arra, hogy a beszédasszisztens rendszerben a gyakorló minták sorrendjének és nehézségi fokának meghatározásához inputként szolgáljon. Ugyancsak szándékomban áll a minősítés eredménye alapján audiovizuális visszajelzést generálni dicsérő és további gyakorlásra ösztönző üzenetek kiválasztására.

8 Összefoglalás

8.1 Összefoglalás és tervezett kutatási irányok

Kutatásaimat egy TÁMOP projekt keretében végeztem, amelynek célja a siket és hallássérült gyerekek beszélni tanítását segítő beszédasszisztens rendszer kidolgozása. Feladataim elsősorban a beszédprodukció minőségének automatikus értékelése köré csoportosultak. A beszédhangok helyes hangzásának értékeléséhez nélkülözhetetlen a hangzó beszéd szegmentálása beszédhangokra. Ez a feltétele annak, hogy a beszédnek azt a szegmensét hasonlítsuk össze egy referenciával, amely az illető hanghoz tartozik. A rejtett Markov-modell természeténél fogva kezelni tudja állapotainak többszöri ismétlődését, így egyes beszédintervallumok nyújtására és zsugorítására kiválóan alkalmas. A torz és gyakran érthetetlen beszéd hangjai azonban olyan távol esnek a Markov-modell állapotaitól, hogy a tiszta beszéddel tanított HMM felismerő nem volt képes a megfelelő pontosságú szegmentálásra. Amikor a HMM felismerőt neurális hálózat outputokkal tanítottuk, valószínűleg a sok ellentmondó adat miatt nem tudtuk betanítani a felismerőt. A beszédhangok fonetikai osztályaira és az osztályon belüli hangok megkülönböztetésére betanított neurális hálózatok kimeneti aktivitása lehetővé tette a torz beszédnél is a szegmentáláshoz elegendő hasonlósági értékek generálását. Az akadozó és ritmushibás beszéd szegmentáláshoz a dinamikus vetemítés algoritmusát célszerűen módosítottam. Ezzel a gyenge minőségű beszédre is jó szegmentálási eredményeket értem el.

Az automatikus minősítés referencia adatainak felvételéhez hangfelvételt készítettünk hallássérült gyerekek bemondásaival. A hangadatbázis minőségét a szurdopedagógusok által megadott szakmai szempontok alapján pedagógus és naiv értékelőkkel pontoztattam. Referenciának a szubjektív pontszámok átlagát tekintettem. A szubjektív minősítés nehézségét jelzi, hogy mind a szakértői mind a hallgatói pontszámok nagy szórást mutatnak. A pedagógus és a hallgatói pontszámok közötti különbség is gyakran jelentős. Szűkített szókészletre részletes elemzést adott egy erre felkért szakértő. Ezzel a több hetes munkával készült hang- és ritmushibát számszerűsítő értékeléssel vetettem össze az automatikus minősítés eredményét. Az automatikus minősítéshez megkíséreltem szokványos távolság mértékeket generálni, ezek azonban nem mutattak egyértelmű kapcsolatot a különböző minőségi kategóriákkal. A szegmentálásnál alkalmazott neurális hálózatok kimeneti aktivitása differenciált és monoton összefüggést mutatott a minőségi osztályokkal. A hiba négyzetes középértékét minimalizálva illesztettem az automatikus minősítés pontszámait a szubjektív minősítés pontjaihoz. Eredményeimet szakértői értékeléssel hasonlítottam össze.

A szegmentálási jellemzőkön túl a szupraszegmentális tényezők is befolyásolják a beszéd érthetőségét. A prozódia értékelésének egyik szempontja a hangsúly a megfelelő szótagra helyezése. A hangsúly automatikus detektálására az alapfrekvencia mellett a

magánhangzók relatív intenzitását használtam fel, amit az adott magánhangzó átlagenergiájához viszonyított pillanatnyi energiával értelmezek.

Munkám során, mivel a nagy adatbázissal tanított neurális hálózatok betanítása több órát vesz igénybe, így a neurális hálózat optimalizálását nem végeztem el. A fejlesztés során a hallássérült gyerekek bemondásairól felvételeket készítünk. Ezeknek a felvételeknek az elemzése a pedagógusi értékeléssel minősítve további tanító mintaként szolgálhat. Az automatikus minősítéshez a formáns struktúra elemzése hasznos kiegészítő lehet, ha a formánsok kinyerésére megbízható módszert találunk.

8.2 Tézisek

8.2.1 I. Tézis [S1], [S2], [S10], [S13]

Megvizsgáltam a nemlineáris idővetemítés különböző módjait, és a gyenge minőségű beszédre módosítottam a dinamikus idővetemítés kapcsolódási szabályait, ezzel az általam vizsgált eljárásoknál lényegesen több határérték esett az egyes hangintervallumok belsejébe.

8.2.2 II. Tézis [S1], [S10], [S11], [S12]

Megvizsgáltam a hangsúlydetektáláshoz használt jellemzők hatékonyságát. Megállapítottam, hogy ha a magánhangzó pillanatnyi energiája helyett a relatív intenzitását használtam, mintegy 10%-kal nagyobb korrelációt értem el a vizsgált adatbázison a mondatok hangsúlyképletéhez viszonyítva.

8.2.3 III. Tézis [S1], [S2], [S3]

Megvizsgáltam a hanghiba és ritmushiba hatását a szubjektív értékelés eredményeire. Meghatároztam a hanghiba és ritmushiba súlyozó együtthatóit az optimális lineáris illesztéshez.

8.2.4 IV. Tézis [S1], [S2], [S3]

Több módszert megvizsgáltam a hallássérült gyerekek hangfelvételeinek elemzésére. Csak a hangfelismerésre betanított neurális hálózatok kimeneti aktivitására találtam differenciált és monoton eredményeket a különböző minőségi osztályokra. Módszeremmel a szubjektív értékelést a tolerancia tartományokban a szakértői becsléshez képest átlagosan 88,5 százalékos pontossággal közelítettem meg.

9 Summary

9.1 Summary and future research directions

My research has been done within the framework of a project called TÁMOP-4.2.2.C-11/1/KONV-2012-0002 (Social Renewal Operational Programme), the aim of which has been the development of a speech-assistant program helping hearing impaired and deaf children to learn to speak. My main task has been to find a way for the program to be able to automatically evaluate the speech-reproduction of these children. For the proper evaluation the speech needs to be segmented into phonemes. This needs to be done in order to be able to compare the corresponding segment of the speech of the children to a reference speech. The hidden Markov-model can handle the continuous repetition of its states, which means that it is perfect for the elongation and shrinkage of the speech intervals. As the states of the Markov-model are very different from the malformed and – in most of the cases – incomprehensible speech, the HMM speech recognizer is not able to segment the speech to an adequate accuracy. When training the HMM speech recognizer with neural network outputs the contradictory data caused the training to fail. The output activity of the phoneme have been recognized by neural networks, which have been trained to be able to distinguish between the different phonetic classes of speech and the different phonemes in a given class, that has made it possible to generate similarity values for malformed speech patterns as well. The algorithm of the dynamic time warping has been modified in order to be able to segment the erratic and rhythm defective speech patterns as well. This has provided good segmenting results even in case of poor quality speech.

We have made voice records from speech of hearing impaired children to define the reference data for the automatic evaluation. The quality of the voice database has been classified with experts and non-experts considering technical aspects given from teachers. We considered the average of the subjective scores as a reference. Both the expert's and non-expert's scores show a large deviation which indicates the difficulty of subjective rating. The difference between the scores of the experts and non-experts are also often significant. I have attempted to generate standard distance scales for the automatic evaluation but these did not show unequivocal relationship with different categories. The output activity of the neural networks applied in the segmentation phase has shown differentiated and monotone correlation with the evaluation classes. I have framed the automatic evaluation scores into the subjective evaluation scores by using minimum mean square error estimator. I have compared the results with that of experts' reviews.

The legibility of speech is influenced not only by the segmentation properties, but by the supra-segmentation factors as well. One of the evaluation criteria of the prosody is to place the emphasis on the proper syllable. To be able to automatically detect the emphasis, beside the fundamental frequency, I have used the relative intensity of

vowels, which has been interpreted by comparing the average energy of a given vowel to its momentary energy.

As the training of neural networks with big databases takes several hours, the optimization of the neural network has not been performed. During the development phase voice recordings of hearing impaired children has been made. The analysis of these recordings, evaluated by pedagogues can be used as further training examples. The analysis of formant structure can be a useful additional tool for the automatic evaluation, if we find a reliable method for the extraction of formants.

9.1.1 I. Thesis [S1], [S2], [S10], [S13]

I have examined different methods of non-linear time warping and modified the connectivity rules of dynamic time warping for low quality speech by which significantly more limit values have fallen into each voice interval than in case of the methods I have examined.

9.1.2 II. Thesis [S1], [S10], [S11], [S12]

I have examined the efficiency of stress detection features. I have concluded that when using relative intensity of vowel instead of instantaneous energy, then regarding the test database I have achieved almost 10% correlation in relation to the emphasis formula of the sentences.

9.1.3 III. Thesis [S1], [S2], [S3]

I have analyzed the effect of word error and tone error on results of subjective evaluation. I have determined the weight factor of word error and tone error to the optimal linear fitting.

9.1.4 IV. Thesis [S1], [S2], [S3]

I have tested several methods for analysing the voice recordings of hearing impaired children. Only the output activity of the neural networks trained for phoneme recognizing has given differentiated and monotone results for the various quality classes. As regards subjective rating in tolerant ranges, my method has reached 88.5 % accuracy in relation to expert rating.

Az értekezés témakörében készített saját publikációk

Folyóiratcikkek

[S1] Dr. Czap László, Pintér Judit Mária: A beszédasszisztens koncepció, Multidiszciplináris tudományok- A Miskolci Egyetem közleménye, (2013) 3. kötet. 1 sz. pp. 241–250. HU ISSN 2062-9737

[S2] Bodnár Ildikó, Czap László, Pintér Judit: Kutatási projekt a hallássérültek internetes beszédfejlesztésére, Alkalmazott Nyelvészeti Közlemények 2014. VIII. évfolyam 2. szám, pp. 19–32. ISSN 1788-9979

[S3] Csetneki Sándorné Dr. Bodnár Ildikó, Czap László, Pintér Judit: Számítógéppel segített beszédfejlesztés, Modern Nyelvoktatás (2014) XX. évfolyam, 4.szám: pp. 75–86. ISSN 1219-628X

[S4] Dr. Czap László, Pintér Judit Mária: Az akusztikus és vizuális jel aszinkronitása a beszédben, Multidiszciplináris Tudományok- A Miskolci Egyetem közleménye (2014), 4. kötet 1.szám, pp. 67–76., HU ISSN 2062-9737

[S5] Czap László, Pintér Judit: A hangsúly egyik jellemző modalitásának vizsgálata, Alkalmazott Nyelvészeti Közlemények 2014. IX. évfolyam 1. szám, pp. 114–121., ISSN 1788-9979

Konferenciaközlemények

[S6] Dr Czap László, Pintér Judit Mária: Beszédfelismerés hatékonyságának vizsgálata különböző nyelvtanokkal, XVII. Fiatal Műszakiak Tudományos Ülésszaka, Műszaki Tudományos füzetek, 2012, pp. 71–74, ISSN 2067-6 808

[S7] Dr Czap László, Pintér Judit Mária: A szavakon túli kommunikáció az audiovizuális beszédszintézisben cikk, XVII. Fiatal Műszakiak Tudományos Ülésszaka, Műszaki Tudományos füzetek, 2012, pp. 67–70, ISSN 2067-6 808

[S8] Czap, L.; Pinter, J. M.: Improving Performance of Talking Heads by Expressing Emotions, Cognitive Infocommunications (CogInfoCom), 2012 IEEE 3rd International Conference, 2012, pp. 523–526, E-ISBN : 978-1-4673-5188-1; Print ISBN: 978-1-4673-5187-4

[S9] Laszlo Czap, Judit Maria Pinter: Multimodality in a Speech Aid System - International Conference on Human Machine Interaction (ICHMI 2013) VOLUME 01 pp.6–11, ISBN-13: 978-81-925233-1-6, ISBN-10: 81-925233-1-4

[S10] Dr. Czap László, Pintér Judit Mária: A beszédprodukció automatikus minősítése hallássérültek beszélni tanításához, XVIII. Fiatal Műszakiak Tudományos Ülésszaka, Műszaki Tudományos füzetek, 2013, pp. 99–102, ISSN 2067-6 808

[S11] László Czap, Judit Mária Pintér: Relative Intensity for Stress Detection; 8. International Scientific Conference on Mechanical Engineering COMEC 2014, Cuba, 5 p., Paper ISBN: 978-959-250-997-9

[S12] Dr Czap László, Pintér Judit Mária: Beszédfelismerés hatékonyságának vizsgálata különböző nyelvtanokkal, XVII. Fiatal Műszakiak Tudományos Ülésszaka, Műszaki Tudományos füzetek, 2012, pp. 71–74, ISSN 2067-6 808

[S13] Dr. Czap László, Pintér Judit Mária: Gyenge minőségű beszéd szegmentálása, XX. Fiatal Műszakiak Tudományos Ülésszaka, Műszaki Tudományos füzetek, 2015, pp. 119–122, ISSN 2067-6 808

Független hivatkozások

[H1] Illésné Kovács Mária: Pozitív és negatív visszajelzések hallássérültek internetes beszédfejlesztésében, Alkalmazott Nyelvészeti Közlemények IX. évfolyam 1. szám, ISSN 1788-9979, 2014. pp. 135–143.

[S1] Dr. Czap László, Pintér Judit Mária: A beszédasszisztens koncepció, Multidiszciplináris tudományok- A Miskolci Egyetem közleménye, 3. kötet. 1 sz. HU ISSN 2062-9737, 2013. pp. 241–250.

[S2] Bodnár Ildikó, Czap László, Pintér Judit: Kutatási projekt a hallássérültek internetes beszédfejlesztésére, Alkalmazott Nyelvészeti Közlemények VIII. évfolyam 2. szám, ISSN 1788–9979, 2014. pp. 19–32.

[S3] Csetneki Sándorné Dr. Bodnár Ildikó, Czap László, Pintér Judit: Számítógéppel segített beszédfejlesztés, Modern Nyelvoktatás XX. évfolyam, ISSN 1219-628X, 4.szám. 2014.pp. 75–86.

Irodalomjegyzék

[1] Abari K., Olaszy G.: *Magyar hangsúlyadatbázis az interneten kutatáshoz és oktatáshoz*, MSZNY 2014. pp.347–356.

[2] Ali Zilouchian: *Fundamentals of Neural Networks*, CRC Press LLC, 2001.

[3] Anne-Marie Öster: *Computer-Based Speech Therapy Using Visual Feedback with Focus on Children with Profound Hearing Impairments*, [Doctoral Thesis], Stockholm Sweden: KTH Computer Science and Communication, 2006.

[4] Baker, J., Deng, L., Glass, J., Khudanpur, S., Lee, C. H., Morgan, N.: *Updated MINDS Report on Speech Recognition and Understanding*, Part I. IEEE Signal Processing Magazine 26/3. 75–80. Part II. IEEE Signal Processing Magazine 26/4. 2009. pp.76–85.

[5] Ben. K., Patrick, S.: *An Introduction to Neural Networks*, University of Amsterdam, 1996.

[6] Bolla Kálmán: *A Phonetic Conspectus of Hungarian*, Tankönyvkiadó, Budapest. 1995.

[7] C.Jeyalakshmi, Dr.V.Krishnamurthi , Dr.A.Revathy: *Deaf Speech Assessment Using Digital Processing Techniques*, Signal & Image Processing : An International Journal(SIPIJ) Vol.1, No.1, 2010. pp.14–25.

[8] Crystal David: *A nyelv enciklopédiája,* Budapest: Osiris. 1998.

[9] Czap L.: *Audio-Visual Speech Recognition And Synthesis*, Phd Thesis, Budapest University Of Technology And Economics, 2004.

[10] Czap, L., Kovács, Zs., Tóth, Á., Váry, Á.: *A Beszédasszisztens használata hallássérültek egyéni beszédfejlesztésében* (Közlés alatt)

[11] Csányi Yvonne: *Bevezetés a hallássérültek pedagógiájába*, B. G. Gyógypedagógiai Tanárképző Főiskola Budapest, 1998.

[12] Csányi Yvonne: *Tanulmányok a hallássérültek beszéd-érthetőségének fejlesztéséről*, Bárczi Gusztáv Gyógypedagógiai Tanárképző Főiskola Budapest, 1995.

[13] Csányi, Y., Zsoldos, M., Perlusz, A.: *Hallássérült (hallásfogyatékos) gyermekek, tanulók komplex vizsgálatának diagnosztikus protokollja,* Budapest: Educatio Társadalmi Szolgáltató Nonprofit Kft., 2012.

[14] Denkinger Géza: *Valószínűségszámítás*, Nemzeti Tankönyvkiadó, 2001.

[15] Deza, E, Deza, M.: *Dictionary of Distances*, Elsevier, ISBN 0444520872, 2006.

[16] Faragó, A., Fülöp, T., Gordos, G., Magyar, G., Osváth, L, Takács, Gy.: *Egyszerű izolált szavas beszédfelismerő,* Kutatási anyag, 1985.

[17] Farkas Miklós: *A hallássérültek kiejtés- és beszédfejlesztésének elmélete és gyakorlata*, B. G. Gyógypedagógiai Tanárképző Főiskola Budapest, 1996.

[18] Farkas, M., Perlusz, A.: *A hallássérült gyermekek óvodai és iskolai nevelése és oktatása*, In: Illyés Sándor (szerk.) Gyógypedagógiai alapismeretek. Budapest: ELTE Bárczi Gusztáv Gyógypedagógiai Főiskola, 2000.

[19] Fujisaki, H., Ohno, S.: *The Use of a Generative Model of F0 Contours for Multilingual Speech Synthesis,* Fourth International Conference on Signal Processing, Vol. 1, 1998. pp. 714–717.

[20] Futó Iván: *Mesterséges intelligencia*, Aula Kiadó, 1999.

[21] Gordos, G., Takács, Gy.: *Digitális beszédfeldolgozás*, Műszaki Könyvkiadó, Budapest, 1983.

[22] Gósy Mária: *Fonetika, a beszéd tudománya,* Osiris, Budapest, 2004. pp.182–243.

[23] Győrffy, P., Bődör, J.: *1978. Gyógypedagógiai ismeretek a hallási fogyatékosság köréből*, Budapest: Tankönyvkiadó, 1978

[24] Hermansky Hynek: *Perceptual linear predictive (PLP) analysis for speech,* Journal of the Acoustical Society of America, 87(4), 1990. pp.1738–1752.

[25] Huang, X., Deng, L.: *Overview of modern speech recognition*. In Indurkhia, Nitin – Damerau. Fred (eds.): Handbook of natural language processing. CRC Press Boca Raton, London–New York. 2010.

[26] Illyés Sándor: *Gyógypedagógiai alapismeretek*, ELTE Bárczi Gusztáv Gyógypedagógiai Főiskolai Kar Budapest, 2000.

[27] Imai, S.: *Cepstral analysis synthesis on the mel frequency scale,* IEEE International Conference on ICASSP '83. (Volume:8), Yokohama, Japan , 1983. pp. 93–96.

[28] J.D. Subtelny: *Speech assessment of the deaf adult,* J. Acad. Rehab. Audiol., 8 (1&2), 1975. pp. 110–116.

[29] Kálmán, Zs., Könczei Gy.: *A Taigetosztól az esélyegyenlőségig*. Budapest: Osiris Kiadó. 2002.

[30] Karl Pearson: *Notes on regression and inheritance in the case of two parents,* Proceedings of the Royal Society of London (58), 1895. pp. 240–242.

[31] Kárpáti Árpád Zoltán *Gondolat – ébresztő! Egy apa feljegyzései a siket gyermekek oktatásáról,* Budapest: Underground Kiadó. 2011.

[32] Kassai Ilona: *Fonetika,* Nemzeti Tankönyvkiadó, Budapest, 1998.

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

[34] Kun L., Xiaojun Q., Shiyin K., Helen M.: *Lexical Stress Detection for L2 English Speech Using Deep Belief Networks*, INTERSPEECH, 2013. pp. 1811–1815.

[35] Mády Katalin: *Beszédpercepció és pszicholingvisztika*, Pszicholingvisztikai kézikönyv, 2008.

[36] Maier A., Haderlein T., Eysholdt U., Rosanowski F., Batliner A., Schuster M., Nöth E.: *Peaks – A System for the automatic evaluation of voice and speech disorders,* Speech Communication, 2009.

[37] Maier A., Hönig F., Hacker C., Schuster M., Nöth E.,: *Automatic evaluation of characteristic speech disorders in children with cleft lip and palate*, Proc. of 11th Int. Conf. on Spoken Language Processing, Brisbane, Australia, pp. 1757–1760.

[38] Markó Alexandra: *A spontán beszéd néhány szupraszegmentális jellegzetessége*, PhD értekezés, Eötvös Lóránd Tudományegyetem, Budapest, 2005.

[39] Martin Kroul: *Automatic Detection of Emphasized Words for Performance Enhancement of a Czech ASR System*, SPECOM'2009, St. Petersburg, 21-25 June 2009. pp.470–473.

[40] Molnár József: *The Map of Hungarian Sounds*, Tankönyvkiadó, Budapest, 1986.

[41] Neha P. Dhole, Dr. Ajay A.Gurjar: *Detection of Speech under Stress*, A Review , International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 10, 2013. pp.36–38.

[42] Németh, G., Olaszy, G.: *A Magyar Beszéd*, Akadémiai Kiadó, Budapest, 2010.

[43] Olaszy Gábor: *Az alapfrekvencia és a hangsúlyozás kapcsolata a magyarban*, In: Kísérleti fonetika - Laboratóriumi fonológia 2002. (szerk.: Hunyadi László) Kossuth Egyetemi Kiadó, Debrecen, 2002.

[44] Olaszy, G., Abari, K., Bartalis, M.: *Magyar hangsúlyjelölési szöveges adatbázis fejlesztése és referenciavizsgálata*, Beszédkutatás 2014. pp. 205–219.

[45] Pytel József: *Audiológia*, Budapest: Victoria Kft., 1996.

[46] Rabiner, L. R.: *A tutorial on hidden Markov models and selected applications in speech recognition*, Proceedings of the IEEE, 1989. 77(2):257–286.

[47] Rabiner, L., R.: *On the use of autocorrelation analysis for pitch detection*. IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume 25, No. 1. 1977. pp. 24–33.

[48] Ross, M. et al.: *Average magnitude difference function pitch extractor*, IEEE Transactions on Acoustics, Speech, and Signal Processing, Volume 22, No. 5. 1974. pp. 353–62.

[49] Szaszák György: *A szupraszegmentális jellemzők szerepe és felhasználása a beszédfelismerésben,* PhD értekezés, Budapesti Műszaki és Gazdaságtudományi Egyetem, Budapest, 2008.

[50] Tarnóczy Tamás: *Zenei akusztika,* Zeneműkiadó, Budapest, 1982. pp. 151–82.

[51] Van Kuijk, D., Boves, L.: *Acoustic characteristics of lexical stress in continuous telephone speech,* Speech Communication, 27(2), 1999. pp. 95–112.

[52] Ververidis, C. Kotropoulos: *Emotional speech recognition: resources, features and methods,* Speech Commun., 48 (9), 2006. pp. 1162–1181.

[53] Vicsi, K., Szaszák, Gy.: *Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features,* International Journal of Speech Technology, Vol. 8, Num. 4, 2005. pp. 363–70.

[54] Waibel Alex: *Prosody and Speech Recognition*, Pitman, London, UK. 1988.

[55] William M. Hartmann: *Signals, Sound, and Sensation*, American Institute Of Physics, 2004. ISBN 1-56396-283-7.

[56] Xie, H., Andreae, P., Zhang, M., Warren, P.: *Detecting stress in spoken English using decision trees and support vector machines*, In: Proceedings of the second workshop on Australasian information security, Data Mining and Web Intelligence, and Software Internationalisation, Australian Computer Society, Inc. 2004. 145–150.

[57] Young, S. et al.: *The HTK Book (For Version 3.3)*, Cambridge University, 2005.

[58] Zwicker, E.: *Subdivision of the audible frequency range into critical bands*, The Journal of the Acoustical Society of America, 33, Feb., 1961.