Valós idejű beszélőnormalizációs eljárás és alkalmazása a „Beszédmester” beszédjavítás-terápiai rendszerben

KOCSOR ANDRÁS

Szegedi Tudományegyetem, MTA-SZTE Mesterséges Intelligencia Kutatócsoport
kocsor@inf.u-szeged.hu

Kulcsszavak: beszédfelismerés, beszélőnormalizáció, beszédjavítás

A különféle beszélőnormalizálási technikák alkalmazása jelentősen javíthatja a beszédfelismerés pontosságát. Ebbe a módszercsaládba tartoznak azok az eljárások is, amelyek az artikulációs csatorna hossznormalizálására (VTLN) törekednek. A kutatások tanúbizonysága szerint ezek a módszerek jól alkalmazhatók, amikor a beszédfelismerő rendszernek megbízhatóan kell működnie férfi, nő és gyermek beszélők esetén is. Elkészítettünk egy számítógéppel segített beszédjavítás-terápiára és olvasásfejlesztésre alkalmas eszközt; a Beszédmestert. A cikk kettős céllal készült. Egyrészt szeretnénk röviden bemutatni a Beszédmester szoftvert, rámutatni újdonságértékére, illetve betekintést adni a vele elért eredményekbe. Másrészt pedig a Beszédmester hátterében nyugvó beszédtechnológiai módszerek közül szertnénk ismertetni egy újszerű valós idejű VTLN eljárást. Nevezetesen megvizsgáljuk, hogy az irodalomból ismert lineáris diszkrimináns alapú VTLN modell, felépítés után, hogyan közelíthető valós idejű kiértékelést biztosító regressziós neuronhálózattal.

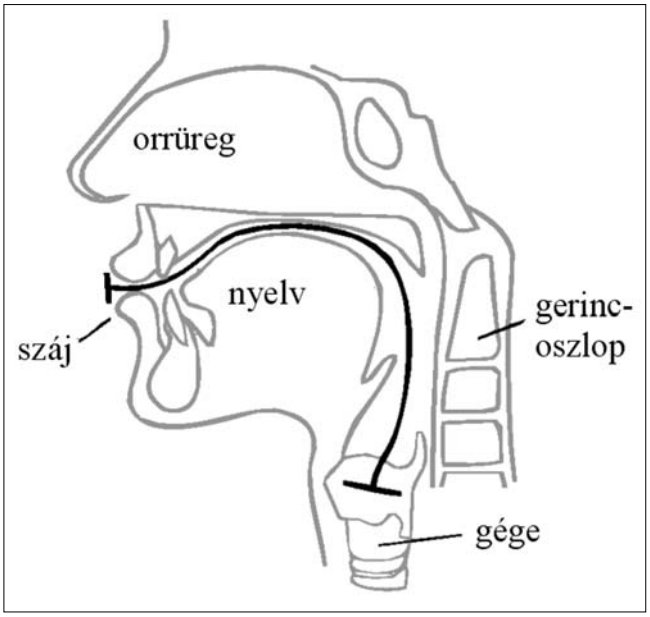
1. Bevezetés

A beszédfelismerési alkalmazások széleskörű elterjedésének előfeltétele a rendszerek beszélőfüggetlensége, azaz a különböző felhasználók esetén is megbízhatóan és pontosan működő felismerési mechanizmus. A rendszerek általánosítási képessége természetes módon fokozható a tanítási folyamat során felhasznált beszédadatbázisok nagy méretével és sokszínűségével, de ez hosszadalmas és költséges munkát követel meg a módszer korlátozott hatékonysága mellett. Egy másik megközelítés ezért az egyes beszélők hangját változtatja meg úgy, hogy az a lehető legjobban hasonlítson az alkalmazott beszédfelismerő rendszer „ideálisan” felismerhető hangjához.

A beszélők hangjai közti eltéréseket a garat-, száj- és orrüreg alkotta artikulációs csatorna *(1. ábra)* egyedekre jellemző különbségei okozzák. Ez a csatorna a gégéből kibocsátott hanghullámot, mint áramló levegőt közvetíti a külvilág felé. Működése során a kiinduló hang bizonyos frekvenciáit felerősíti, másokat elnyom, de a jelenlévő frekvenciák szerkezetét nem módosítja. Emiatt a csatorna jelfeldolgozási szempontból szűrőként viselkedik, azaz matematikailag egy átviteli függvénnyel jellemezhető (megadja az egyes frekvenciák erősítésének mértékét). Az átviteli függvény maximumhelyeit formánsoknak nevezzük. Fonetikai kísérletek alapján kijelenthető, hogy a hang első három formánsa a fonetikai minőséget, míg a többi a beszélő egyedi jellemzőit (például hangszín) határozza meg. Beszédfelismerés szempontjából az itt jelentkező különbségeket célszerű kompenzálni, létrehozva így egy „idealizált” beszédhangot.

A különböző beszélők által ejtett, de azonos fonetikai csoportba tartozó hangok formánsszerkezete hasonló, a frekvenciaértékek egymáshoz képest csak kis mértékben elcsúsztatottak. Minden fonetikai csoport jellemezhető egy átlagos formánsszerkezettel. A legelterjedtebb megközelítés szerint az egyes beszélők formánsainak átlagostól vett eltérése korrelál az artikulációs csatorna hosszával (Vocal Tract Length – VTL), ezért lehetőség nyílik a hangok normalizálására a beszélőre jellemző VTL érték szerint. Ezt a módszert a szakirodalom Vocal Tract Length Normalization-nek (VTLN) nevezi. Ennek során egy transzformációs függvény segítségével próbáljuk a beszélők formánsait az átlagos helyre leképezni [4,13]. A beszélőnormalizálás feladata egy megfelelő típusú transzformációs függvény kiválasztása, és a beszélőre jellemző optimális paraméterek beállítása.

1. ábra Az emberi artikulációs csatorna



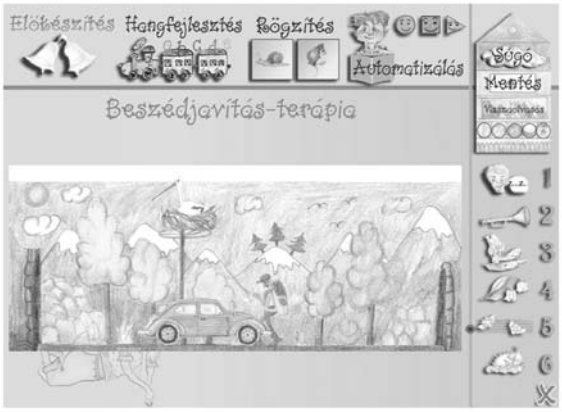
A VTLN során a transzformációs paraméterek beállítása történhet néhány másodperces bemondás alapján, vagy a teljes hanganyag segítségével is. Mindkét esetben számottevően csökkenthető a fonémafelismerés hibája.

Ezeknek a módszereknek az eredeti formában történő alkalmazására azonban nincs lehetőség az általunk kifejlesztett „Beszédmester” szoftver [6-11] működése során, hiszen ebben az esetben valós idejű fonémafelismerésre, és így valós időben elvégezhető normalizációs technikára van szükség. A probléma megoldására egy valós idejű VTLN eljárást ismertetünk, amely a bemondás pillanatában hatékonyan megbecsüli az alkalmazott transzformációs függvény paramétereit [10].

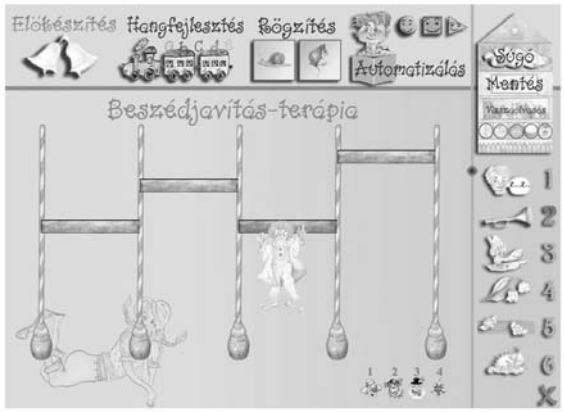
Jelen publikáció felépítése a következő. A második fejezetben, röviden bemutatjuk a „Beszédmester” szoftvert, amely funkcionalitásából adódóan felveti a valós idejű beszálőnormalizáció kérdését. A harmadikban ismertetjük a neuronháló-alapú normalizációs megoldásunkat, amit a módszer teszteléséről és a teszteredmények kiértékeléséről szóló negyedik fejezet követ. Végül levonjuk a cikk tanulságait és felvázoljuk a továbbfejlesztés lehetőségeit.

2. ábra Képek a beszédjavítás-terápia részből

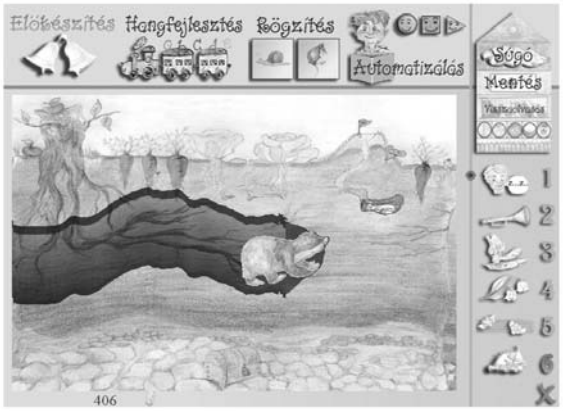
(A) zöngeadás gyakorlása



(B) hangerő változtatás berögzítése

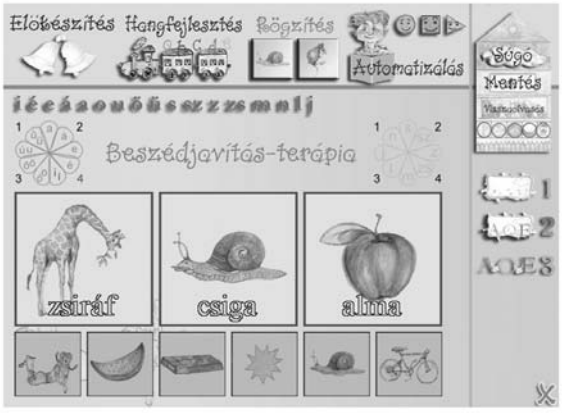


(C) játék a hangmagasság finom változtatásával

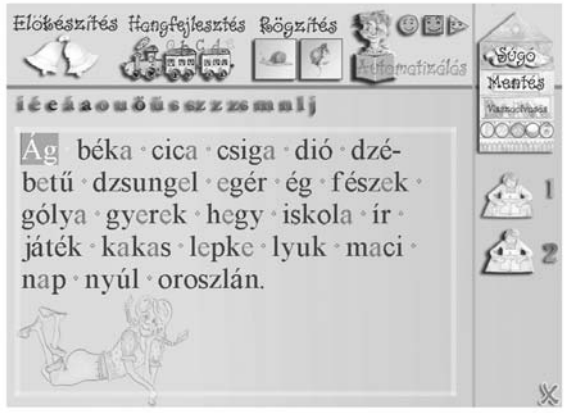


(D) hangfejlesztés valós idejű visszacsatolással

(E) hangok rögzítése szavakban



(F) hangok rögzítése mondatokban



2. A „Beszédmester”

A Beszédmester [6-11], egy komplex beszédjavítás-terápiai és olvasásfejlesztő számítógépes szoftver, amely ingyenesen letölthető a következő weboldalról:

http://www.inf.u-szeged.hu/beszedmester.

2.1. Beszédmester a beszédjavítás-terápiában

Óvodás és kisiskolás életkorban nagyon gyakori jelenség, hogy gyermekeink beszédében zavaróan súlyos hanghibák vannak, sőt előfordul az is, hogy alig, vagy egyáltalán nem értjük őket. Sok gyermeknél műtéti úton állítják elő a szinte teljes értékű, természetes minőségű hallást; ők a cochlea implantáltak. Számukra több éves folyamat a hallás olyan szintű „megtanulása”, ami a hangos beszéd elsajátításához szükséges. Vannak gyermekek, akik ép hallásuk mellett is súlyosan beszédhibásak. Ha szervi okok (szájpadhasadék stb.) akadályozzák a szépen hangzó beszéd kialakulását, akkor a műtét utáni rehabilitációban logopédus szakember segítségével megoldást lehet találni. Ép hallás és beszédszervek esetén is késhet a beszéd megindulása, vagy olyan sok hangzóhibát véthet a gyermek, ami zavaró, alig érthető kiejtéshez vezet. A Beszédmester szoftver *(2. ábra)* az érthető beszédet döntően befolyásoló magánhangzók felismerését végzi el az elhangzás pillanatában (2/D. ábra), ezzel egy olyan új eszközt adva a szakemberek kezébe, amellyel jelentősen lerövidülhet a remélt, kívánatos színvonalú beszédállapot eléréséhez szükséges gyakran több évi munka [3,5].

A Beszédmester szurdopedagógiai felhasználását, kipróbálását a program fejlesztésének kezdeti szakaszától a Kaposvári Siketek Iskolájában végezték. A kipróbálás eredményeit egyrészt az artikulációfejlesztő terápiát irányító szakemberek (szurdopedagógusok, szurdologopédusok, logopédusok), másrészt a terápiában részt vevő gyermekek megjegyzéseiből, kérdéseikből szűrték le, illetve hasznos kiegészítésül szolgáltak a tesztelés külső megfigyelőinek feljegyzései is. Ezek a megfigyelések adták a programot használó gyermekek metakommunikativ jelzéseinek legátfogóbb képet.

A foglalkozást követő elemzések során nagyon fontosak voltak a gyermekek reakcióiról összegyűjtött feljegyzések. Az olyan apróságnak tűnő megjegyzések, mint: *„...felcsillanó szemmel, mosolyogva nézett a monitorra...”, „...tapsikolt örömében, amikor meglátta a ...”*, pozitív minősítését adták az adott programrész felépítésének, működésének. A Beszédmester szoftver kipróbálását súlyos fokban hallássérült (siket) gyermekek kiejtésjavítására, nagyothalló és cochlea implantált gyermekek beszédérthetőségének fejlesztésére, és beszédfogyatékos gyermekek artikuláció fejlesztésére terápiás foglalkozásokon végezték. A nem beszelő siket gyermekek kiejtésjavítását öt hónapon keresztül végezték. A gyermekek magánhangzóinak ejtését a terápia előtt és után vizsgálták. Az eredmények azt mutatták, hogy a terápia előtt 1-3 magánhangzót a fejlesztés hatására pedig már 4-7 magánhangzót ejtettek helyesen a gyermekek.

A látványos növekedésben a Beszédmester szerepét bizonyítja, hogy valamennyi gyermek 1-2 éves szurdopedagógiai fejlesztésben vett részt, tehát a terápia előtti számadat 1-2 év kiejtésnevelésének a hozadéka. Már beszélő siket gyermekek és a logopédiai fejlesztés során az ajak- és szájpadhasadékos gyermekek „gépi” beszédjavító terápiája is látványos eredményt hozott a magánhangzók tisztaságában. A Beszédmesterrel „tanuló” gyermekek magánhangzói teljes mértékben orrhangzósságtól színezetmentessé váltak, és sokkal tisztábbak lettek.

A Beszédmester hatásfoknövelő szerepét a nagyothalló és a cochlea implantált gyermekek beszédérhetőségének fejlesztésében is vizsgálták. Mindkettő esetében a Beszédmester alkalmazásának legnagyobb hozadéka, hogy az állandó értékelés, visszajelzés (a számítógép automatikusan végzi el!) mellett tudják saját kiejtésüket javítani. A színes, motiváló, mozgó grafika azokból a gyermekekből váltott ki újra és újra hangos megnyilatkozást, akik eddig ritkán hallatták hangjukat. Az alacsony életkorú, figyelemzavaros gyermekeket tartós figyelmi helyzetbe hozta a változatos gyakorlási módok sokasága.

2.2. Beszédmester az olvasásfejlesztésben

A Beszédmester szoftver az olvasástanítás segítését, az olvasás terápiáját, fejlesztését is célul tűzte ki. A szoftver segítségével játékos úton, szinte észrevétlenül lehet gyakoroltatni az olvasást *(3. ábra)*. Használható az iskolai olvasástanítás során és egyéni gyakorlásra. A részképességükben sérült gyermekek fejlesztő terápiájában komplex készségfejlesztést biztosít: memória- és figyelemfejlesztés, irányfelismerés, iránytartás kialakítása, finommotorika fejlesztése, hallási diszkriminációs készség, hallási figyelem, vizuális differenciáló képesség fejlesztése [15]. Segítheti a diszlexiaterápiát, hiszen a betűk újratanításának feltételei biztosítottak, a fonémapárok feladatai pedig célzottan a kritikus párok gyakorlására, differenciálására készültek (3/F. ábra). A teszteredmények azt mutatják, hogy nemcsak az első osztályosok fejlesztésére, tanítására alkalmas a szoftver, hanem a 8-10 évesek terápiájában is kiválóan használható.

Az olvasásfejlesztés modul célja eredetileg egy programfüggetlen olvasástanítási szoftver előállítása volt, azonban az előzetes tesztek azt bizonyították, hogy az olvasási nehézségekkel küzdő gyermekek munkáját is hatékonyan segíti a Beszédmester. A munka eredményességét mutatja, hogy a gyermekek szívesen és hosszan dolgoznak a szoftverrel, s a feladatok megoldása segíti a fonológiai tudatosság kialakulását. A részképességükben sérült gyermekek, de még egészséges társaik is sokszor nehezen tájékozódnak a tankönyv feladatai között, figyelmüket könnyen elterelik a színes ábrák. A Beszédmester maximálisan biztosítja az egyéni tempóban való haladást. Az egyéni irányított munkaformában a gyermekek szívesen fogadták a mikrofonnal kiegészített szoftvert. Az egészséges óvodás és iskoláskorú gyermekek is könnyen dolgoztak vele. A magánhangzó-felismerése egyszerűbb feladatnak bi-

zonyult. A szófelismerés során azt tapasztaltuk, hogy a gyerekek egészen addig próbálkoznak az adott szó helyes kiejtésével, amíg fel nem villan a hívókép írásos változata, a szókép. Csoportos, irányított munkaformában a program hang- és szófelismerő része alkalmat ad a tehetséggondozásra és a felzárkóztatásra.

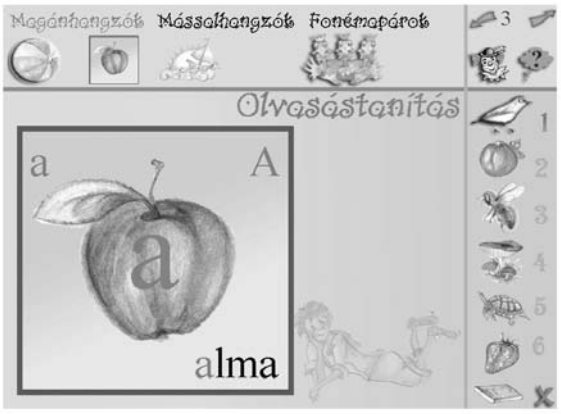
2.3. Beszédtechnológia a Beszédmesterben

A Beszédmester program, különösen a beszédorientált részeket tekintve, a számítógéppel segített oktatás területén innovatív jelentőségű, hiszen az interakció a beszédinterfész által a számítógép és a felhasználó között még emberibbé válik. A tanulás/terápia a tanuló/sérült gyermek és a számítógép manipulatív, „barátságos” interakciója alapján valósulhat meg.

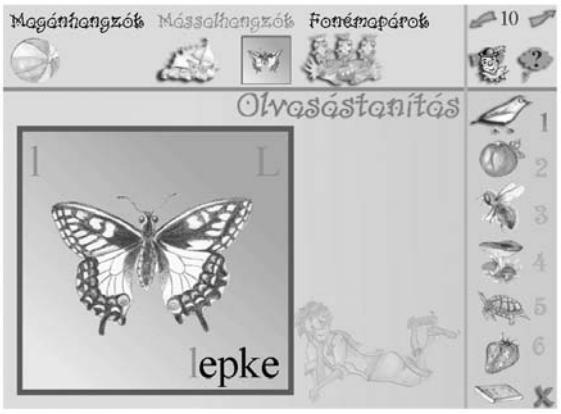
A szoftver kulcseleme a beszédtechnológiai modul, amely lehetővé teszi, hogy a rendszer a mikrofonba bemondott beszédhangokat, illetve szavakat valós időben visszajelezze. A Beszédmesterben alkalmazott technológia segítségével nem a hang oszcillogramja, vagy spektruma jelenik meg a gyakorlás során, hanem maga a hozzá kapcsolt betű jele tűnik fel.

3. ábra Képek az olvasásterápia részből

(A) magánhangzó tanítás valós idejű fonémafelismeréssel

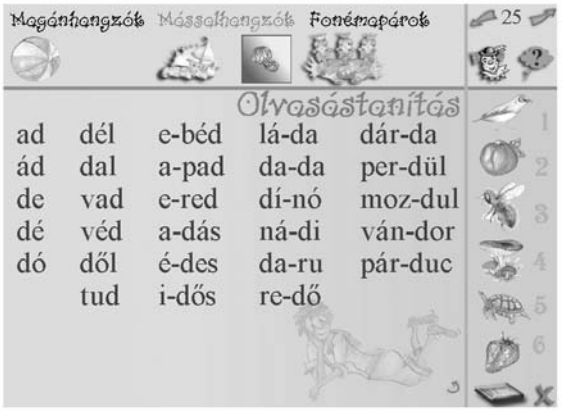


(B) mássalhangzó tanítás gépi szófelismeréssel



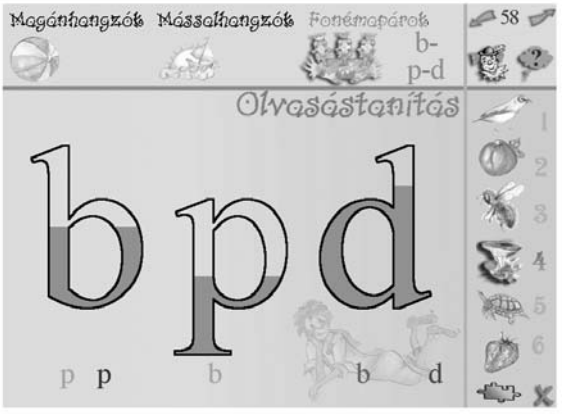
(C) fonémák helyének keresése

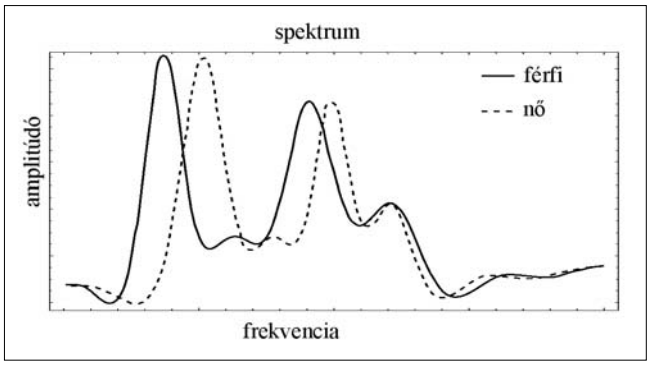
(D) olvasási gyakorlat



(E) kritikus párok újratanítása gépi szófelismeréssel

(F) betűfeltöltés





4. ábra
Az artikulációs csatorna hosszának frekvenciaeltoló hatása. Az ábrán egy férfi és egy nő által bemondott magánhangzó spektruma látható egy adott pillanatban.

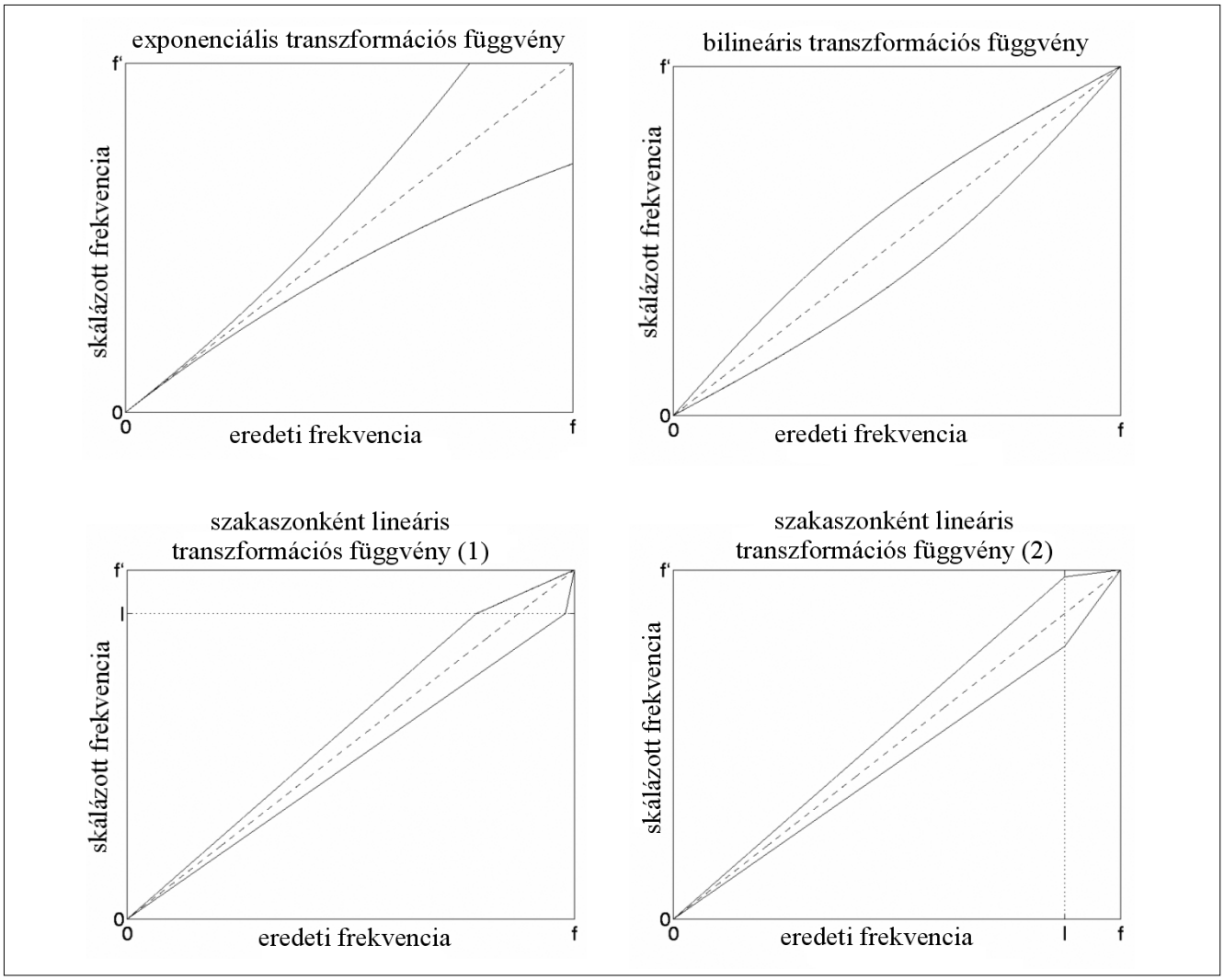
A vizuális kijelzés az elhangzással azonos időben történik. Mind a beszéd analízise, mind a beszédhangok azonos idejű és fonémaszintű feldolgozása önműködően megy végbe. Tehát ez a megoldás már nem a mikrofon előtt közvetlenül elhangzó egyedi beszédhangot jeleníti meg, hanem azt, amit „megért" a rendszer, azaz a hanghoz társított betűképet.

3. Beszélőnormalizáció az artikulációs csatorna modellezésével

Az emberi populáció egyedeinek beszédkarakterisztikája nagy különbözőséget mutat. Ennek egyik fő fiziológiai forrása a beszélők artikulációs csatornájának a hossza. A férfiak esetében az átlagos hossz 17, a nőknél 15, míg a gyermekeknél 14 cm [2]. Az artikulációs csatorna voltaképpen a beszélő formánsait eltolja a frekvencia tengelyen az adott fonetikai csoportra jellemző átlagtól [13]. Az eltolás azonban nem lineáris, az alsó és a felső frekvencia tartományokban különböző mértékű lehet. Az eltolási jelenség első fizikai leírása a Helmholtz rezonátor volt, amely az artikulációs csatornát egy csővel modellezte, így az eltolási mértékre exponenciális összefüggés adódott. A későbbi, finomabb modellek már figyelembe vették azt a tényt, hogy a magasabb frekvencia tartományokban csökken az eltolás mértéke, így keletkezett a bilineáris és a szakaszonként lineáris transzformációs függvény [12].

Az alkalmazott transzformációs függvények *(5. ábra)* általában kevés paraméterrel definiálhatóak, ez megkönnyíti mind az egyedre jellemző optimális paraméter meg-

5. ábra Lehetséges transzformációs függvények.
Az ábrák az eredeti f és a transzformált f' frekvenciaértékek közötti kapcsolatot mutatják.



határozását, mind a paraméteres modell kiértékelését: az exponenciális, a bilineáris és a szakaszonként lineáris függvények mindegyike egyetlen paraméterrel írható le.

Miután kiválasztottunk egy transzformációs függvénytípust, a beszédadatbázisban szereplő beszélőkre egyénenként meghatározzuk az optimális paraméter értéket egy iteratív eljárás segítségével. Bár a transzformációs függvény paraméterértéke folytonos, mégis célszerű a lehetőségeket egy kis elemszámú, véges halmazra korlátozni (általában 10-20 érték). Az egyedre jellemző optimális paraméterérték kiválasztásának stratégiája és az optimalizálás kritériuma módszerről módszerre változik [4,16-18]). Jelen dolgozat nem a létező módszerek összehasonlítására összpontosít, helyette vizsgálódásainkat csak a [17]-ben említett lineáris diszkrimináns (LD) alapú módszerre szűkítjük.

Ez az eljárás ugyanolyan hatékonynak, de stabilabbnak bizonyult, mint a legelterjettebb Maximum Likelihood módszer [4]. A Maximum Likelihood (ML) alapú paraméteroptimalizáció kezdetben minden beszélőhöz azt a paraméterértéket rendeli, amely nem változtatja meg a formánsait. Az így betanított beszédfelismerőt felhasználva, a lehetséges paraméterértékek egyénenkénti kipróbálásával minden beszélőre meghatározható az optimális érték, amely mellett a legnagyobb valószínűséggel ismerhetők fel a mintái. Ezután, az így meghatározott egyedi paraméterek felhasználásával transzformált beszédadatbázis mintáin újra betanítjuk a beszédfelismerő algoritmust. Az iterációt addig folytatjuk, amíg a változás meghalad egy bizonyos minimális mértéket.

A Lineáris Diszkrimináns Analízis (LDA) alapú LD-VTLN módszer [17] szintén iteratív eljárás, amely az iteráció során az előző módszertől eltérően nem használ beszédfelismerő eljárást. A módszer az egyénenként megadott paraméterhalmaz kiértékelését a lineáris diszkrimináns (LD) kiszámításával végzi. Az LD érték meghatározása az egyénenként különböző módon transzformált mintákat tartalmazó beszédadatbázis alapján történik. A beszédadatbázisban a minták címkézettek, egy-egy fonetikai szimbólum és az adatközlő sorszáma, illetve adatai vannak hozzárendelve. A fonetikai szimbólumok alkotják az osztályokat, amelyek minnél pontosabb megkülönböztetése a cél ismeretlen mintára.

Először meghatározunk egy B mátrixot, amely az egyes osztályok mintái átlagának szórását reprezentálja. Majd előállítjuk a W mátrixot, ami az osztályokon belüli minták szórásának átlagára jellemző. Az LD értéke a két mátrix determinánsának hányadosa: LD=|B|/|W|. Ez az érték nagy, ha az egyes osztályok elemei kis szórást mutatnak, miközben a különböző osztályok átlagai távol esnek egymástól. Az LD érték növekedésével voltaképpen egyre nagyobb osztály-szeparabilitást (fonémaszeparációt) érhetünk el.

Az iteratív LD-VTLN algoritmus pszeudókódja a következő:

1. Válasszunk minden beszélőhöz egy iniciális paraméterértéket és végezzük el az általa bemondott minták frekvencia-transzformációját.
2. Minden beszélőre számoljuk ki az LD értéket úgy, hogy a hozzá társított egyedi paraméterértéket kicsit megnöveljük, illetve csökkentjük. Ezután válasszuk ki azt a paramétert, melyre a legjobb LD értéket kaptuk.
3. Transzformáljuk az adatbázist a kapott paraméterhalmaz figyelembevételével.
4. Ugorjunk a 2-es lépésre mindaddig, amíg az átlagos paraméterváltozás kisebb egy előre beállított küszöbszámnál. Különben az algoritmus végrehajtását megszakítjuk, és a kapott paraméterekkel transzformált adatbázison betanítjuk a beszédfelismerő módszert.

Ismeretlen mintára a spektrumot transzformáljuk a lehetséges transzformációs paraméterek mindegyikével, majd meghatározzuk a hozzájuk tartozó legvalószínűbb szimbólumsorozatot az LD-VTLN során betanított beszédfelismerő segítségével. A felismerés eredménye ezek közül a legkisebb költségű lesz. Sajnos ez az eljárás azonban nagyon költséges, valós időben nem kivitelezhető.

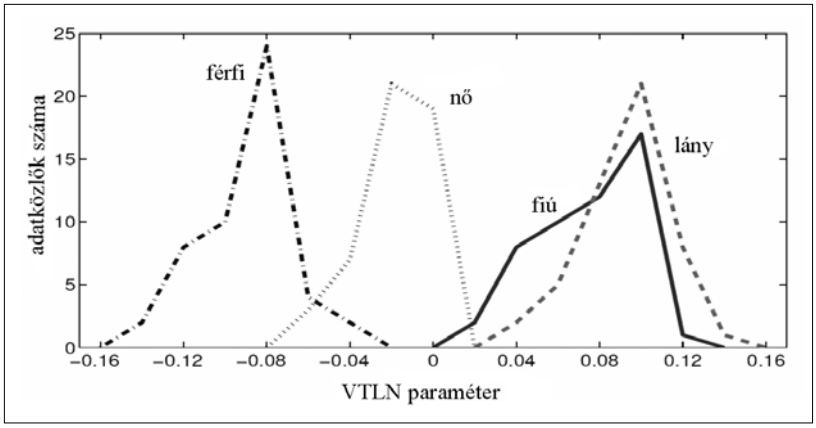
3.1. Valós idejű beszélőnormalizáció a Beszédmesterben

A BeszédMester beszédjavítás-terápia részében valós idejű fonéma-felismerési feladatot kell megoldani. A feladat során a tanár, illetve a diák felváltva használja a rendszert, ami tovább nehezíti a beszédfelismerőrendszer feladatát.

A rendszer tanításához bemondásokat rögzítettünk a célcsoportbeli felnőttek, illetve gyerekek segítségével. A gyerekek hangjához nehezebb megbízható felismerőt készíteni, mivel ebben a korban még sokat változik az artikulációs csatorna hossza és formája, ezért az elkészült beszédadatbázis sok különböző korú gyermek hangját tartalmazza. Első lépésben a tanítás során LD-VTLN alapú beszélő-normalizációt alkalmaztunk a fonémákból álló beszédadatbázison. A paraméterek eloszlása pontosan visszaadta a beszélők kor és nem szerinti eloszlását, amelyet megfigyelhetünk az *6. ábrán*.

Ezután a rendelkezésre álló optimális transzformációs paraméterek közelítésére egy regressziós algoritmust taníthatunk be az adatbázisban lévő bemondá-

6. ábra
A VTLN paraméterek eloszlása kor és nem szerint



sok alapján. A modell feladata, hogy a beszélő kilétének ismerete nélkül, mindössze a bemondásból kinyerhető akusztikai információk alapján becsülje meg az LD-VTLN algoritmus által kihozott optimális normalizációs paraméter értéket. Így a fonéma-felismerés során első lépésben a transzformációs paraméter becslése történik meg, majd ezután a transzformált hangon kell lefuttatni a beszédfelismerő rendszert. A felismerés így valós idejű marad, és teljesítménye alig marad el a globális keresés eredményétől. A tesztelés menetét és technikai részleteit a következő fejezetben írjuk le.

4. A módszer tesztelése

Ebben a fejezetben megvizsgáljuk, hogy az eredeti LD-VTLN algoritmus, illetve a valós idejű kiterjesztése milyen hatékonyságnövelésre képes a fonémaklasszifikáció feladatán. Először leírást adunk a teszteléshez használt beszédkorpuszról, majd számba veszzük az alkalmazott tulajdonságkinyerő módszert, illetve az osztályozó eljárást, végül bemutatjuk az eredeti LD-VTLN eljárás alkalmazásának részleteit és a teszt során használt regressziós algoritmust.

Beszédkorpusz. Tanítási és tesztelési céllal 200 beszélőtől rögzítettünk hangmintákat. A nemek közötti eloszlás a következő volt: 50 nő, 50 férfi, 50 lány és 50 fiú. A gyerekek a 6-9 éves korosztályból kerültek ki. A beszédjeleket 22050 Hz-es mintavételezéssel, 16-bit-es minőségben rögzítettük és tároltuk el. Minden beszélő rövid szünetekkel elválasztva az összes magyar magánhangzót kiejtette. Mivel úgy határoztunk, hogy nem teszünk különbséget a hosszú és rövid magánhangzók között, ezért összesen 9 magánhangzóval dolgoztunk.

Tulajdonságkinyerés. A jeleket 10 ms-os keretekben dolgoztuk fel, ezután 24 kritikus sáv energiáját számoltuk ki a logaritmikus skálán, az FFT és háromszög súlyozás felhasználásával [12]. Minden egyes keret energiáját külön-külön normalizáltuk, ami azt eredményezte, hogy csak a spektrális eloszlás alakját használtuk fel a fonémaosztályozáskor.

Osztályozók. A kísérletek során mesterséges neuronhálókat [1] használtunk fel a fonémaklasszifikációs feladat megoldására. A szokásosan alkalmazott háromrétegű, előrecsatolt többszintű neuronhálós modellt használtuk (feed-forward MLP), amelyet a backpropagation tanulóalgoritmussal tanítottunk. A rejtett neuronok száma minden esetben 16-ra lett állítva.

LD-VTLN. Az LD-VTLN algoritmust két különböző transzformációs függvény felhasználásával alkalmaztuk. A bilineáris és a szakaszonként lineáris transzformációs függvények, zárt alakban, rendre a következő formulákkal definiálhatók (1)(2):

$$f' = \arctan\frac{(1-\alpha^2)\sin f}{(1+\alpha^2)\sin f - 2\alpha} \qquad \alpha \in [-0.18, 0.18],$$

$$f' = \begin{cases} \alpha f & \text{if } 0 < f < 0.7/\alpha \\ \gamma f + (1-\gamma) & \text{egyébként.} \end{cases} \qquad \alpha \in [0.825, 1.175],$$

$$\gamma = \frac{0.3\alpha}{\alpha - 0.7}.$$

Az α transzformációs paraméter kezdőértékét 0-ra állítottuk az (1)-es és 1-re a (2)-es transzformációs függvény esetében. Az α lehetséges értékei rendre a formulák után láthatóak. Az optimalizáció elvégzéséhez α lehetséges értékeit kvantáltuk, az adott intervallumot ekvidisztáns módon 15 részre vágtuk. Az iterációt addig folytattuk, amíg a transzformációs paraméter átlagos változása 10^{-2} alá nem esett.

Regresszió. Az LD-VTLN módszer által kinyert beszélőkre jellemző paraméterek becslésére egy speciális MLP hálózatot hoztunk létre egy kimeneti neuronnal és két rejtett réteggel, amely 24-24 neuronból épült fel. A tanítást az átlagos négyzetes hiba minimalizálásával végeztük. A neuronháló inputját a 24 kritikus sáv energiája (ld. tulajdonságkinyerés) alkotta.

Tesztek. A kísérletek a következőképpen lettek elvégezve. Az összességében 200 ember bemondásaiból álló beszédkorpusz 3 db 50, 100 és 200 beszélőből álló részre lett felosztva, minden esetben egyenletes volt a fiúk, lányok, férfiak és nők aránya. Az így kapott beszédadatbázist 80/20% arányban bontottuk fel tanuló és tesztelő mintákra. A tanuló minták esetében mindkét transzformációs függvényre (bilineáris, szakaszonként lineáris) és minden lehetséges α értékre előállítottuk a transzformált mintát, majd kinyertük a 24 sáv energiaértékeit. Ezután végrehajtottuk az LD-VTLN algoritmust, amely beszélőnként kiválasztotta az optimális paraméter értéket. Ezek alkották a regressziós neuronháló kívánt outputját, míg az input adatokat az eredeti minta sávenergiái jelentették. A tesztek során neuronháló alkalmazásával a következő adatokon végeztünk osztályozást: transzformáció nélküli adatok (NO VTLN), LD-VTLN transzformálás utáni adatok (LD-VTLN), illetve a valós idejű transzformálással kapott minták (Realtime(RT)-VTLN).

Az eredményeket az alábbi *táblázat* foglalja össze:

Klasszifikációs hiba a magánhangzófelismerés feladatán. A sorok az adatközlők száma szerint, az oszlopok pedig a transzformációs függvény és a normalizációs algoritmusok szerint rendezettek.

Adatközlők száma	Bilineáris transzformáció			Szakaszonként lineáris transzformáció		
	NO-VTLN	LD-VTLN	RT-VTLN	NO-VTLN	LD-VTLN	RT-VTLN
50 adatközlő	18.52%	13.07%	14.09%	18.52%	13.86%	14.35%
100 adatközlő	15.36%	12.02%	13.02%	15.36%	12.17%	13.19%
200 adatközlő	14.33%	11.02%	11.52%	14.33%	10.87%	11.04%

Kiértékelés. Az eredményekből láthatjuk, hogy az LD-VTLN módszer a klasszifikációs hibát akár 21-29%-kal is csökkentheti. A regressziós becslést alkalmazva közelítőleg 14-24%-os hibacsökkenést kaptunk, amely megközelíti az LD-VTLN eredményeket. Az adatbázis méretének növelésével (50, 100, 200 adatközlő), a tényleges eltérés a két módszer között csökken. A két transzformációs függvény között nem tapasztaltunk jelentős különbséget.

5. Összefoglalás

A kifejlesztett Beszédmester szoftver egyrészt azzal a céllal készült, hogy segítse az iskolások olvasásfejlesztését, másrészt, hogy a hallássérült, siket és logopédiai kezelésben részesülő gyermekek számára kínáljon gyorsabb fejlődési lehetőséget. Előnye, hogy játékosan, színes képekkel, a számítógép motivációs erejét felhasználva próbálja meg a kisiskolásokat az olvasás rejtelmeire megtanítani, és a gyermekeket a tiszta, hangos beszéd birtokosává tenni. A szoftver célirányos, tudatos, oktató-fejlesztő program, mely figyelembe veszi az életkori sajátosságokat, a komplex készségfejlesztést és nagyfokú önállóságot biztosít.

A Beszédmester szoftver használati értékét a beszélőfüggetlen automatikus beszédfelismerő technológia nagymértékben megnöveli. Ennek a technológiának az egyik kulcseleme a beszélőnormalizáció. A beszédterápia során a tanár és a diák felváltva beszél, ezért gyorsan változhat az optimális normalizációs függvény, vagy ennak paramétere. A probléma megoldására egy valós idejű beszédnormalizációt lehetővé tévő VTLN eljárást ismertettünk, amely a bemondás pillanatában hatékonyan megbecsüli az alkalmazott transzformációs függvény paramétereit. Összegzésül, kijelenthetjük, hogy érdemes a VTLN eljárások – mint például az LD-VTLN – regresszió-alapú becslésével foglalkozni, mert az így kapott modell már valós időben kiértékelhető és majdnem olyan hatékony, mint a közelített eredeti eljárás.

Végezetül, meg kell említenünk, hogy fonológiai tudatosságra nevelő rendszerekben – mint a Beszédmester – nem csak a fonémák osztályozásának hatékonysága, hanem az egyes fonémaosztályok elkülönülésének milyensége is nagy jelentőséggel bír. Bizonyos beszélőnormalizációs módszerek, köztük az LD-VTLN és a bemutatott RT-VTLN, az egyes osztályok minél élesebb elválasztására törekszenek. Az e módszerek által létrehozott szeparációs modell vizsgálata további kihívást jelent és lehetőséget biztosít a kutatások folytatására.

Irodalom

[1] Bishop, C. M., (1995)
Neural Networks for Pattern Recognition, Oxford Uni. P.

[2] Claes, T., Dologlou, I., Bosch, L., Compernolle, D.(1998)
A Novel Feature Transformation for Vocal Tract Length Normalization in Automatic Speech Recognition,
IEEE Trans. on Speech and Audio Processing,
Vol. 6., pp.549–557.

[3] Csányi, Y. (1990) Hallás-beszédnevelés,
Tankönyvkiadó Budapest.

[4] Eide, E., Gish, H., (1997) A Parametric Approach to Vocal Tract Length Normalization, Proc. ICASSP'97,
Munich, Germany, pp.1039–1042.

[5] Farkas, M. (1996)
A hallássérültek kiejtés- és beszédfejlesztésének elmélete és gyakorlata, BGGYPF, Budapest.

[6] Kocsor, A., Tóth, L., Paczolay, D. (2001)
A Nonlinearized Discriminant Analysis and its Application to Speech Impediment Therapy, In:
V. Matousek, P. Mautner, R. Moucek, K. Tauser (eds):
Proc. of the 4th Int. Conf., Speech and Dialogue,
LNAI 2166, Springer Verlag, pp.249–257.

[7] Kocsor, A., Kovács, K. (2002)
Kernel Springy Discriminant Analysis and Its Applic. to a Phonological Awareness teaching System, In:
P. Sojka, I. Kopecek, K. Pala (eds.):
TSD 2002, LNAI 2448, Springer Verlag, pp.325–328.

[8] Kocsor, A., Tóth, L. (2004)
Kernel-Based Feature Extraction with a Speech Techn. Applic., IEEE Trans. on Signal Processing,
Vol. 52., No.8., pp.2250–2263.

[9] Paczolay, D., Kocsor, A., Sejtes, Gy., Hégely, G.(2004)
A „Beszédmester" csomag bemutatása: informatikai és nyelvi aspektusok. Alkalmazott Nyelvtudomány,
Veszprém, 4/1.szám, pp.57–79.

[10] Paczolay, D., Kocsor, A., Tóth, L. (2003)
Real-Time Vocal Tract Length Normalization in a Phonological Awareness Teaching System,
Matousek, V., Mautner, P. (eds.): TSD 2003,
LNCS 2807, Springer Verlag, pp.309–314.

[11] Paczolay, D., Tóth, L., Kocsor, A., Kerekes J. (2002)
Gépi tanulás alkalmazása egy fonológiai tudatosságfejlesztő rendszerben, Alkalmazott Nyelvtudomány,
2/2.szám, pp.55–67.

[12] Pitz, P., Molau, S., Schlter, R., Ney, H. (2001)
Vocal Tract Normalization Equals Linear Transform. in Cepstral Space, Proc. EUROSPEECH 2001,
Vol. 4., pp.2653–2656.

[13] Rabiner, L. R., Juang, B. H. (1993)
Fundamentals of Speech Recognition,
Englewood Cliffs, NJ, Prentice Hall.

[14] Uebel, L. F., Woodland, P. C. (1999)
Investigation into Vocal Tract Lenght Normalisation,
Proc. EUROSPEECH 99, Hungary,
Vol. 6., pp.2527–2530.

[15] Ványi, Á. (1998) Olvasástanítás a diszlexia-prevenciós módszerrel. Project-X. Budapest, pp.4–7.

[16] Wegmann, S., McAllaster, D., Orloff, J.,
Peskin, B. Speaker (1996)
Normalization on Conversational Telephone Speech,
Proc. ICASSP'96, Atlanta, Vol. 1., pp.339–341.

[17] Westphal, M., Schultz, T.,Waibel, A. (1998)
Linear Discriminant – A New Criterion for Speaker Normalization, Proc. ICSLP'98, pap.no.755, Sydney.

[18] Zhan, P., Westphal, M. (1997)
Speaker Normalization based on Frequency Warping,
Proc. ICASSP-97, Munich, Vol. 1, pp.1039–1042.