AKUSZTIKAI HANGOSZTÁLYOK FELISMERÉSÉN ALAPULÓ, NEMLINEÁRIS IDŐVETEMÍTÉS MEGVALÓSÍTÁSA A MONDATHANGLEJTÉS ÉS A SZÓHANGSÚLYOZÁS OKTATÁSÁHOZ

Kiss Gábor – Vicsi Klára

Bevezetés

Kutatási célunk egy olyan általános, több nyelvre működő, nemlineáris időbeli vetemítési eljárás kifejlesztése volt, ami azonos tartalmú beszédminták összehasonlítására alkalmas, főleg a beszédminták prozódiai jellemzőinek összevetésére oktatás-alkalmazásokhoz. Cél volt az is, hogy az eredményt mind vizuálisan, mind számértékkel, illetve egyéb szöveges módon megjelenítsük a tanulónak.

A prozódiának a beszédfolyam automatikus tagolásában, a szintaktikai és szemantikai információ helyes közlésében igen nagy szerepe van (Olaszy 2010). A magyar beszédben a prozódiával fejezzük ki a mondatok modalitását is. A prozódia sok esetben a közlés értelmezését is elősegíti.

Ennek a munkának így egyik lehetséges gyakorlati haszna, hogy az általunk kifejlesztett eljárás az audiovizuális nyelvoktatásban vagy a hallássérültek kiejtéstanításában a prozódia fejlesztéséhez jól felhasználható.

A BME TMIT Beszédakusztikai Laboratóriumában korábban kifejlesztettek egy HMM alapú, automatikus intonációosztályozó rendszert, amelyet hallássérültek beszédterápiájában használtak fel (Sztahó et al. 2009; Szaszák et. al. 2009). Ez a rendszer automatikusan különbséget tudott tenni a helyes és a helytelen mondatintonáció között, de a vizuális visszajelzésnél nehezen volt értelmezhető az etalon és az aktuális minta összehasonlítása. Ennek oka az volt, hogy ez az eljárás csupán mondatok időtartamát egyenlítette ki egyformára, és ez egyfajta lineáris skálázásnak (lineáris vetemítésnek) tekinthető.

Ahhoz, hogy ugyanazon mondat két egyén által kiejtett formáját össze lehessen hasonlítani prozódiai szempontból, azokat hangszinten azonos időbeli hosszra kell skálázni, azaz egymáshoz vetemíteni. A prozódia összehasonlításánál a konkrét probléma az, hogy a beszéd tartalmaz nemlineáris megnyúlásokat és rövidüléseket, és ezek nem feltétlenül számítanak hibás ejtésnek. Amennyiben tehát lineáris időskálázást használnánk a prozódiai jellemzők összehasonlításához, akkor az egyes hangok nagy valószínűséggel nem pontosan ugyanazon két időpont közé esnének. Így a jellemzők összehasonlítása-

nál ott is eltérés mutatkozna, ahol a tényleges mintákban a bemondás jóságában nem lenne lényeges különbség, hiba. Csupán abból fakadna az eltérés, hogy az időbeli eltolódás miatt a beszéd nem megfelelő szakaszából mérnénk az adatokat.

Az ideális időbeli összehasonlítás az, amikor a két mintát az egyes beszédhangok mentén illesztjük egymáshoz. Ezt alkalmazzák gépi beszédfelismerésnél, de ez a megoldás nyelvfüggő (Baker et. al. 2009; Huang–Deng 2010).

Több nyelvre használható prozódiai összehasonlításhoz olyan vetemítő eljárás kell, ahol szigorúan a beszédhangok mentén történik az illesztés, és csak azokon belül lineáris a skálázás. Ezzel nemcsak azt tudná közölni a program a tanulóval, hogy az általa bemondott mondat prozódiai jellemzői nagyban eltérnek a referenciamondattól, hanem képes lehetne a hibát lokalizálni is, és pontosan megfogalmazni, hogy mit ejt rosszul a tanuló. Például, hogy az adott szónak nem az első, hanem az utolsó szótagját ejtette hangsúlyosan. A tanulmányban bemutatunk egy általunk kifejlesztett, olyan több nyelvre működő automatikus szegmentáláson alapuló nemlineáris vetemítő eljárást, amely lehetővé teszi a különböző prozódiai jellemzők meglehetősen pontos összehasonlítását mondat egységű kiejtést tanítási folyamatokban.

Anyag és módszer

A kutatásunk során három, különböző nyelvű beszédadatbázist használtuk, az MRBA magyar (Vicsi et. al. 2004), a Kiel német (Benno 2005) és a TIMIT angol (Garofolo et al. 1993) nyelvű adatbázisokat. Mind a három adatbázist beszédhang egységben szegmentáltunk és címkéztünk. Ezek az adatbázisok olvasott, folyamatos, fonetikailag kiegyensúlyozott szöveget tartalmaznak. A feladat szempontjából fontos jellemzőik leolvashatóak az 1. táblázatról.

1. táblázat: Az eljárás kifejlesztése során felhasznált adatbázisok

Az adatbázis neve	A beszélők száma	Bemondási környezet	Az adatbázis hossza percben
MRBA	100	iroda, otthon	330 perc
Kiel	50	iroda, otthon	180 perc
TIMIT	40	iroda, otthon	310 perc

A tárolt hangminták digitalizálása 16 kHz-es mintavételi frekvenciával, 16 bites lineáris kvantálással készült. Az eljárásunkban három alapvető módszert alkalmazunk: gépi szegmentálás, mintaillesztés és vetemítés a referencia és a bemondott mondat összehasonlítására; valamint az előbbiek tesztelése. Az eljárásunk újdonságát az jelenti, hogy a gépi szegmentálás folyamatában általános akusztikai hangosztályokat használunk, ami lehetővé teszi a szegmentálást több nyelv esetén is általános akusztikai környezetben. Ebből következik, hogy a nemlineáris idővetemítést is általánosan több nyelvre alkalmasan oldottuk meg.

A gépi szegmentálásban használt újszerű hangosztályozás alapötletét a Vicsi Klára és Vig Attila által publikált tanulmány (1998) adta. Ennek a lényege, hogy képesek voltak neurális háló segítségével, több nyelven 80-90%-os szegmentálási pontosságot elérni az általuk definiált akusztikai hangosztályokban süketszobai körülmények között rögzített hangfelvételek esetén.

Eljárásunk során mi is ezeket az akusztikai hangosztályokat használtuk a gépi szegmentáláshoz. Az akusztikai hangosztályok és a hozzájuk tartozó beszédhangok, beszédhangrészek a 2. táblázatból olvashatóak le.

2. táblázat: A definiált akusztikai hangosztályok felsorolása és annak jelölése, hogy mely osztályba mely beszédhangok, vagy összetett hangok esetén mely beszédhangrészek kerültek magyar nyelv esetén [+*: a zöngés zárszakasz (zönge), –*: a zöngétlen zárszakasz [csend])

Az akusztikai hangosztály neve	Az akusztikai osztály jele	Osztályba sorolt beszédhangok (SAMPA-jele)
Magánhangzók 1	mv	O, A:, E, o, o:, u, u:, 2, 2:
Magánhangzók 2	hv	e:, i, i:, y, y:
Zöngétlen spiránsok	s-	f, s, S, h,
Zöngés spiránsok	s+	v, z, Z
Nazálisok és likvidák	na	r, l, j, m, n, J
Zöngés zárszakasz	b+	+*
Zöngés felpattanás	vo	b, b:, d, d:, g, g:, dz, dz:, dZ, dZ:, d', d':
Zöngétlen zárszakasz	b-	–*
Zöngétlen felpattanás	uv	p, p:, t, t:, k, k:, ts, ts:, tS, tS:
Csend	si	

Az osztályozás lényege, hogy a nyelvtől független artikulációs konfigurációkat határoz meg, amelyek általánosak. Amennyiben két ugyanazon osztálybeli egység áll egymás mellett, akkor a szegmentáló nem különíti el a két elemet. Ez azonban nem olyan súlyos probléma, ugyanis prozódiai szinten elégséges a szótagszintű pontosság is. Szótaghibát csak az okoz, ha két egy osztálybeli magánhangzó kerül egymás mellé (*boa*), ez azonban ritkán fordul elő, és egyszerű szabályrendszerrel nagyrészt a hiba kiküszöbölhető (elemhosszfigyelés).

A gépi szegmentáláshoz Hidden Markov Model (HMM) elvű felismerő motort használtunk (Mihajlik 2010), amellyel a 10 fonetikai osztály akusztikus modelljét tanítással hozzuk létre. Az összehasonlításhoz nemlineáris vetemítéssel egyforma időskálára hozzuk a mondatok hangelemeit. Adott a referenciamondat, ahol ismert az időszerkezet, és a felhasználó ugyanezt a mondatot mondja be, ahol viszont nem ismerjük a beszélő ejtési sajátosságait (hadar, lassan beszél stb.). Eljárásunk feltételezi, hogy címkézési szinten ismerjük a referenciamondat időszerkezetét, illetve hogy a felhasználó kooperál.

A megfelelő vetemítéshez szükséges, hogy ismerjük a bemondott mondat beszédhang- vagy szótagszintű időszerkezetét is. Ehhez a bemondást automatikusan szegmentálni kell. A szegmentálás után összehasonlítjuk a bemondás és a referencia magánhangzóit, és a megfelelőket összekapcsoljuk. Ezután a referencia- és a bemondott mondatot vetemítjük egymáshoz, amit a magánhangzók időkoordinátái alapján végezzük el. Ezután teszteljük a vetemítés jóságát a mondatintonáció és a mondatintenzitás dinamikája alapján. Az eredmény összehasonlíthatóságához a tesztelést elvégezzük lineáris skálázás esetén is. Ehhez kirajzoljuk a mondatintonáció és mondatintenzitás dinamikáját, illetve kiszámoltatjuk ezeknek a referenciához képesti négyzetes távolságát. (Vagyis vettük mintapontonként a négyzetes eltérést, ezeknek az átlagát, majd gyököt vontunk az eredményből.)

Az eljárás megvalósítása

A szegmentáláshoz szükséges HMM modellek tanításához az MRBA adatbázis 30%-át használtuk fel. A program bemenetként a tanításra felhasznált beszédmintákat, illetve a hozzájuk tartozó annotációs fájlokat kapja meg. Első lépésben előfeldolgozást végez, majd a hangfájlokat feldarabolja az annotációs fájl ismeretében. Az előfeldolgozás során a periférikus hallásnak megfelelő felbontásban (kritikus sávokban) színképelemzést végez. Kiszámítja a hangminta kritikus sávokban lévő energiaértékeit, illetve azok deriváltjait. 20 Hztől 7,7 kHz-ig huszonegy sávra bontja a hangmintát. Ezek az értékek alkotják a jellemzővektorokat a modelleket létrehozó szoftver bemeneteként. A program hangosztályonként háromállapotú Markov-lánccal dolgozik. Az előfeldolgozással nyert jellemzők alapján és az annotációs fájl segítségével létrehozza az egyes akusztikai hangosztályok modelljeit. Így minden fonetikai hangosztályhoz létrehoz a tanító eljárás egy külön akusztikai modellt. Ennek a módszernek az egyik előnye például, hogy bármikor bármelyik modell könnyen cserélhető, illetve bővíthető vagy csökkenthető a modellek száma.

Ezután a felismerés, illetve a szegmentálás a következőképpen történik. Adott a bemondás hipotetikus leirata (feltesszük, hogy kooperatív a bemondó, és törekszik a referenciamondat bemondására, legfeljebb kisebb hibákat ejt). A leiratot egy fonetikai átíró átírja a megfelelő akusztikai hangosztályok szerint. Ezután a HMM elvű gépi felismerő megkapja a bemondást, az akusztikai hangosztályokká átírt leiratot és az akusztikai hangosztályok akusztikai modelljeit. Ezeknek a segítségével elvégzi a felismerést, szegmentálást a bemondáson, és azt a PRAAT program (Boersma–Weenink 2001) annotációs formátumában visszaadja.

Ezután valósítottuk meg a nemlineáris idővetemítést. Ideális esetben a vetemítési feladat egyszerű. Tegyük fel, hogy a két mondat beszédhangjainak sorozata megegyezik, ebben az esetben a vetemítés csupán annyi lenne, hogy a bemondott mondat beszédhangjainak hosszát lineárisan úgy módosítanánk egyesével, hogy azok a referenciával azonos hosszúak legyenek. Sajnos ezzel a feltételezéssel nem élhetünk, mivel a beszéd rendkívül változatos (beszéd-

hangbeszúrások, kihagyások, elharapások, másképp ejtések). Még akkor sem egyezik meg feltétlenül a két mondat, ha a bemondó törekszik a pontosságra.

A szegmentáló a legjobban a magánhangzókra működött 90%-os körüli pontossággal (magyar nyelv esetén). Másrészt a beszéd prozódiai elemzéséhez elég a magánhangzónkénti pontosság, vagyis elégséges, ha magánhangzónként végzünk időbeli vetemítést nemlineárisan, és két magánhangzó között pedig lineárisan. Így kézenfekvőnek látszott úgy dönteni, hogy a két mondatban a magánhangzók helyzetét illesztjük dinamikusan, és a köztük lévő részt pedig lineárisan.

Kulcspontoknak neveztük el azokat a pontokat a referencia- és a vetemítendő mondatban, amelyek illesztése dinamikus lesz. Ilyen kulcspontok a magánhangzók, illetve a beszéd kezdete és vége. Mivel a kulcspontok sorozata nem feltétlenül egyezik meg, szükséges egy mintaillesztési eljárás. Amennyiben a két mintasorozat ugyanaz, akkor az egymás alatti kulcspontokat illeszti. Probléma akkor adódik, ha a két mintasorozat eltér egymástól. Feltehetjük, hogy az eltérés nem lesz jelentős (a bemondó kooperatív), emiatt háromféle eltéréssel számolhatunk:

– beszúrás: a bemondó beszúrt egy, esetleg több hangot, amely eredetileg nincs jelen;

– kihagyás: a bemondó kihagyott egy, esetleg több hangot (elharapta), amely a referenciamondatban megtalálható;

– másképp ejtés: a bemondó más hangot ejt ki, mint a referenciában eredetileg volt.

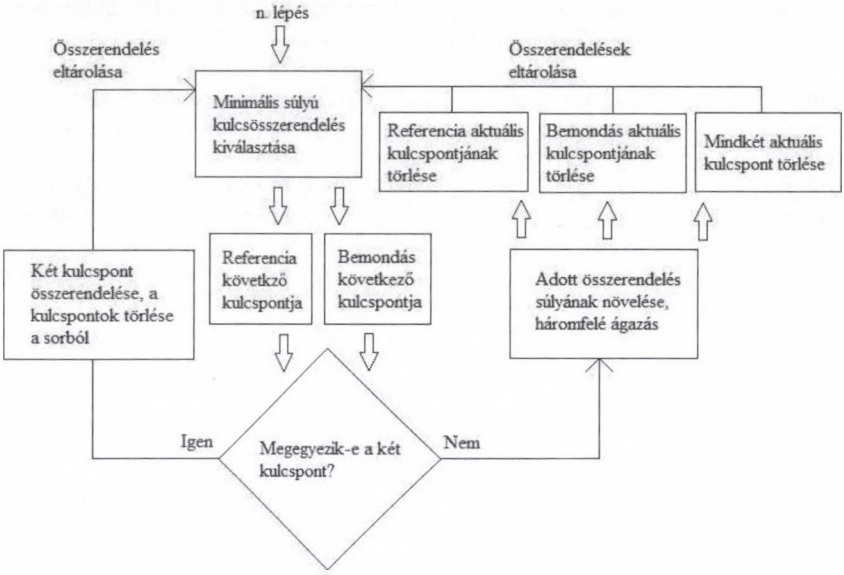
Ezek alapján az illesztést a következő gépi eljárás szerint végeztük. Felvettük kulcspontoknak a magánhangzókat, illetve a mondat elejét és végét. Az algoritmus végigment a referencia- és a vetemítendő mondat kulcspontjain egyesével (mondat eleji kulcsponttal kezdve), ha a két elem megegyezett, mindkettőn lépett egyet előre, és összekapcsolta őket, ha nem egyeztek meg, az adott út súlyát növelte eggyel, és háromfelé ágazott az algoritmus:

– a referenciamondatban lép egyet előre az algoritmus, míg a vetemítendőben nem, és ehhez a referenciaponthoz semmit sem kapcsol (törli a kulcspontot);

– a vetemítendő mondatban lép egyet előre az algoritmus, míg a referenciában nem, és ehhez a vetemítendő ponthoz semmit sem kapcsol (törli a kulcspontot);

– mindkét sorban lép egyet az algoritmus, és ezeket a pontokat senkivel se kapcsolta össze (mindkét kulcspont törlésére került).

Ezután iteratívan folytatódott az algoritmus futása. Ha megint eltért egymástól a két mintasor, akkor megint háromfelé ágazott, míg az adott útvonal súlyát növelte eggyel. Mindig az aktuálisan minimális súlyú úton végezte a számítást. Az algoritmus akkor fejeződött be, ha mindkét sor esetén elérte a vége kulcspontokat (ezek nem törölhetőek). A mintaillesztő algoritmus egy lépése látható az 1. ábrán.



1. ábra
A mintaillesztő algoritmus egy lépése

Ha a két mintasorozat megegyezik, akkor valóban az ideális működést biztosítja. Egyébként a vetemítés működésének helyességét a végeredményekből adódóan láthatjuk, ahol kifejezetten szélsőséges példán is kipróbáltuk az eljárást. Ezután a következő a vetemítési algoritmust, programot implementáltuk az időben nemlineáris vetemítéshez:

– beolvassa a referencia- és a vetemítendő mondat szegmentálását;

– bejelöli a referencia- és a vetemítendő mondatban a beszéd elejét és a végét, ami nem feltétlenül egyezik meg a hangminta elejével, illetve végével;

– levágja a két mondat felesleges részét (vagyis ami a beszéd kezdete előtt, illetve után van), és ezután a maradékot 0–1 időintervallumra skálázza;

– a magánhangzók közepei kulcspontok lesznek, az „mv"-beliek A típusú, a „hv"-beliek B típusú kulcspontok;

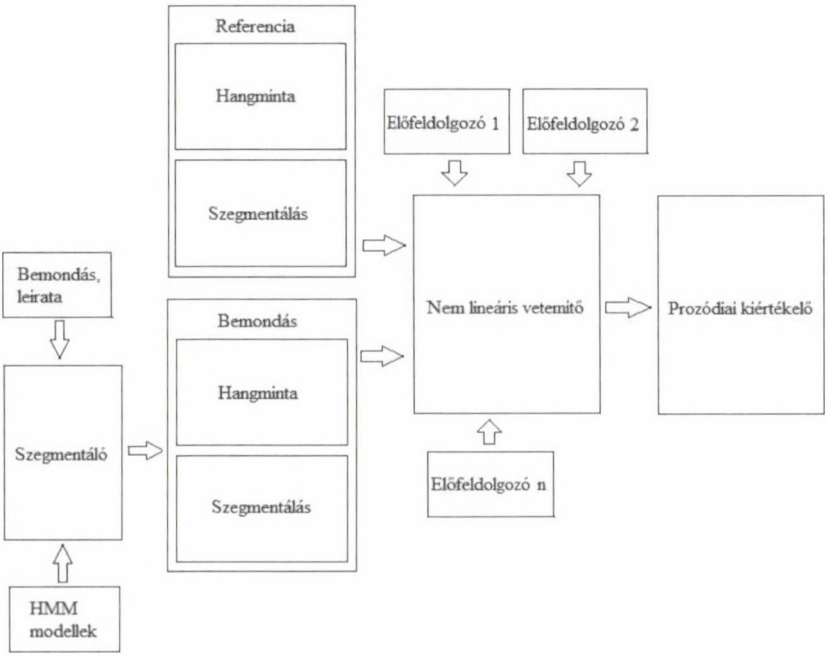
– ezután elvégzi a korábban tárgyalt mintaillesztési eljárást, ahol véglegesíti, illetve összerendeli a kulcspontokat;

– a vetemítendő mondatban a kulcspontokat a referenciamondatban lévő időszerkezetre helyezi át, és a köztük lévő részt lineárisan skálázza.

Ezután pedig egy megfelelő távolságszámító algoritmussal a két mondat prozódiai távolsága meghatározható, így eldönthető, hogy helyes-e a bemondott mondat prozódiai szempontból vagy sem. Egy ennél még „okosabb" ki-

értékelő program akár azt is közölheti a felhasználóval, hogy milyen tévedéseket ejtett, és azokon hogyan képes változtatni.

Az időben nemlineáris vetemítő eljárást az eddigiek alapján implementáltuk, illetve megalkottunk egy rendszert, ami képes két mondat prozódiai jellemzőinek összehasonlítására. A teljes eljárás folyamatábrája 2. ábrán látható.



2. ábra
A prozódiai kiértékelő folyamatábrája

Az eljárás bemenete a referenciamondat hangmintája, illetve a hozzátartozó annotáció. Ez látható az ábrán a referenciarésznél (középen fent). Az eljárás további bemenete a bemondott mondat hangmintája, illetve a hozzátartozó leirat. Ezután az eljárás a leirat alapján elkészíti a bemondott mondat hangosztályszintű szegmentálását. A szegmentálás eredményét és a bemondott hangmintát a referenciamondat hangmintájával és szegmentálásával együtt megkapja a nemlineáris vetemítő modul, amely a korábban tárgyaltak szerint elvégzi az időben nemlineáris vetemítést az egyes előfeldolgozók által mért jellemzők mentén. Ezután a prozódiai kiértékelő elvégzi a görbék elemzését, és a felhasználó számára valamilyen visszajelzést nyújt.

Eredmények

Az eredményeket a feldolgozási sorrend szerint mutatjuk be. Először a gépi szegmentálás, majd a vetemítési eljárás eredményeit közöljük. A gépi szegmentáló kiértékelése a következő automatikus eljárással készült:

– először végignézte az eljárás a szegmentálást a referenciához képest. Ha bárhol is két szegmenshatár csupán 25 ms-on belül tért el, azt nem tekintettük hibának. Ilyenkor a felismert mondatban a szegmenshatárt áthúzta a referencia-szegmenshatár időpontjába. Erre azért volt szükség, mert a kézi szegmentálás is legalább ekkora hibahatárral rendelkezett.

– A javítás után 10 ms-os lépésközzel összehasonlította a referenciamintát és a felismert eredményt. Ha az adott mintavételi pontban megegyezett a két annotáció, akkor az abban az időpontban történő felismerést sikeresnek vette, ha eltért, akkor sikertelennek, és eltérés esetén megjegyezte a tévesztést is, így képes volt tévesztési mátrixot építeni.

A fent definiált kiértékeléssel az eljárást három nyelvre teszteltük: magyar, német és angol. Bár a tanítás során csak magyar beszédmintákat használtuk fel az akusztikai modellek létrehozásához, így várható, hogy az angol és német nyelv esetén valamivel alacsonyabb lesz a gépi szegmentálás pontossága. Magyar nyelv esetén az MRBA adatbázis azon beszédmintáin teszteltük, amelyek nem szerepeltek a tanítás során. Vagyis a teljes adatbázis 70%-án, ez körülbelül 230 perc. A német nyelven való teszteléshez a Kiel Korpuszt használtuk, ez körülbelül 180 perc. Az angol nyelven való teszteléshez a TIMIT adatbázist használtuk, ez körülbelül 310 perc. A tesztelési eredményeket a 3., 4. illetve az 5. táblázat tévesztési mátrixaiból lehet leolvasni.

A magyar nyelv esetén az eredmény valamivel jobb, mint a már korábban megvalósított neurális hálókkal való felismeréssel kapott eredmények (85%) (Vicsi–Vig 1998), annak ellenére, hogy ott süketszobai, zaj nélküli felvételeket szegmentáltak. A mi esetünkben pedig természetes irodai körülményeknek megfelelő zajos környezetben rögzített hangfelvételeken történik a szegmentálás. A többi nyelv esetén viszont alacsonyabb a szegmentálási pontosság, bár az összehasonlítást nem lehet tökéletesen elvégezni, mivel Vicsi Klára és Vig Attila tanulmányában a többi nyelvből is tanítottak a felismerés esetén. Feltehetőleg ha nyelvenként külön modelleket használnánk, akkor a magyarhoz hasonló szegmentálási pontosságot kapnánk. A táblázatból jól látható, hogy a magánhangzók (mv és hv) felismerési eredménye a legjobb (80–93%). Így ez megfelelő lehet a későbbi időben nemlineáris vetemítéshez. Az összes európai nyelv esetén az akusztikai hangosztályok használata elegendő az adott nyelv összes hangjának a besorolásához. A prozódiai összehasonlításhoz a tíz osztály használata elégséges, ugyanis a célunk nem felismerni a bemondott hangsorozatot, csupán megtalálni a beszédhangok, kiváltképp a magánhangzók határait. A vetemítő eljárásunk tesztelésénél a következő szempontokat vettük figyelembe:

– az eljárásban az újdonság a „nemlineáris idővetemítés";

– az eljárásunk helyességének a tesztelése a cél, ezért törekedtünk, hogy a bemondás dallammenete és intenzitásgörbéje hasonló legyen, miközben tartalmazzon nemlineáris megnyúlásokat, rövidüléseket.

3. táblázat: Magyar nyelven történt tesztelés tévesztési mátrixa

	mv	hv	s+	s-	na	vo és b+	uv és b-	si
mv	**93%**	2%	0%	0%	2%	1%	1%	1%
hv	4%	**90%**	0%	0%	4%	1%	1%	0%
s+	9%	5%	**74%**	2%	5%	3%	2%	0%
s-	5%	1%	1%	**86%**	2%	1%	4%	0%
na	17%	9%	0%	0%	**68%**	3%	2%	1%
vo és b+	5%	4%	1%	0%	6%	**82%**	1%	1%
uv és b-	4%	1%	0%	1%	3%	1%	**88%**	2%
si	4%	1%	0%	4%	2%	3%	2%	**84%**

Összesített felismerés: 86% (jó minták/összes minta).

4. táblázat: Német nyelven történt tesztelés tévesztési mátrixa

	mv	hv	s+	s-	Na	b+	b-	vo	uv	si
mv	**83%**	2%	2%	3%	4%	2%	0%	1%	2%	1%
hv	6%	**80%**	1%	1%	6%	3%	1%	1%	1%	0%
s+	8%	13%	**74%**	0%	2%	2%	1%	0%	0%	0%
s-	6%	4%	0%	**81%**	4%	1%	2%	1%	1%	0%
na	12%	7%	1%	3%	**69%**	5%	1%	1%	1%%	0%
b+	9%	7%	1%	2%	2%	**73%**	0%	2%	0%	4%
b-	2%	2%	1%	6%	6%	3%	**76%**	0%	4%	0%
vo	4%	3%	3%	0%	2%	10%	4%	**74%**	0%	0%
uv	7%	6%	0%	6%	6%	0%	20%	1%	**51%**	3%
si	5%	0%	1%	0%	3%	4%	2%	0%	0%	**86%**

Összesített felismerés: 78% (jó minták/összes minta).

5. táblázat: Angol nyelven történt tesztelés tévesztési mátrixa

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**93%**	3%	0%	1%	1%	0%	1%	0%	0%	0%
hv	6%	**86%**	1%	1%	3%	1%	1%	0%	0%	0%
s+	11%	4%	**76%**	1%	2%	1%	1%	1%	1%	2%
s-	5%	3%	1%	**87%**	1%	0%	1%	0%	1%	0%
na	26%	7%	1%	2%	**59%**	2%	1%	1%	1%	0%
b+	5%	2%	1%	3%	5%	**74%**	2%	7%	1%	1%
b-	5%	2%	1%	5%	2%	1%	**79%**	1%	4%	1%
vo	7%	4%	2%	3%	7%	2%	2%	**70%**	2%	1%
uv	8%	5%	2%	1%	5%	1%	4%	1%	**71%**	1%
si	2%	2%	1%	3%	2%	1%	6%	1%	1%	**80%**

Összesített felismerés: 80% (jó minták/összes minta).

Tehát kettős célunk volt, egyrészt megmutatni, hogy az eljárás nem rosszabb, mint a lineáris skálázás, másrészről korrektebb az általunk vizsgált prozódiai jellemzők távolságának meghatározása esetén. Háromféle bemondásra teszteltük le az eljárást, majd végeztük el a prozódiai kiértékelést. A bemondott mondat így hangzott: *Az északi szél nagy vitában volt a nappal, hogy kettejük közül melyikük az erősebb.* A bemondás módjai:

– normális tempóban: referenciának (6320 ms),

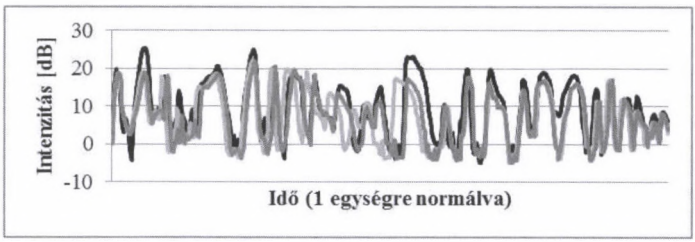
– normális tempóban (törekedve, hogy a bemondás időszerkezete megegyezzen a referenciával): bemondottnak (6213 ms),

– más tempóban (szüneteket beleütemezve) bemondva, ezáltal erős, nemlineáris megnyúlásokat garantálva a mondatban (8101 ms).

A két mondat prozódiai távolságának kiszámításához, szemléltetéshez az alaphang és az intenzitás dinamikáját használtuk fel. A távolságszámítás során négyzetes távolságszámítást választottuk, és ennek átlagát számoltuk. A mondatintonáció (dallammenet), illetve az intenzitás kiszámításához PRAAT programot használtuk. Az alaphang kiszámítása során az ablak mérete 50 ms, a lépésköz 10 ms volt. Az intenzitás kiszámítása során az ablak mérete 100 ms, a lépésköz 10 ms volt.

Az adott mondat 26 magánhangzót tartalmaz, így 28 pont dinamikus vetemítése történik meg, és a pontok között a lineáris skálázás lesz. A kiértékelő eljárással kirajzoltattuk az alaphangot, illetve az intenzitást lineáris skálázással és nemlineáris vetemítéssel is, összehasonlítva a referenciával mindkét bemondás esetén. Az alaphangot, illetve az intenzitást normáltuk az első valódi értékre (az első mért értéket kivontuk a többiből), illetve az alaphangot interpoláltuk azokon a szakaszokon, ahol nem volt zönge. (A zöngétlen szakasz két végpontját összekötöttük egy egyenes szakasszal.)

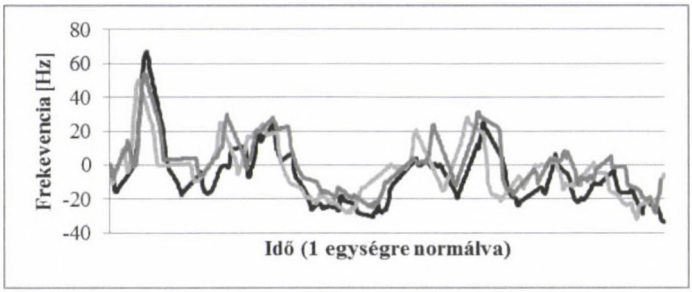
A 3., 4., 5. és 6. ábrákon a vízszintes tengelyen a referenciamondat szerint (6320 ms) 1 egységre lett skálázva, és a tengely 0–1 intervallumban lineáris.



3. ábra

A referenciamondat és a „jól" bemondott mondat alaphang-dinamikájának összehasonlítása, különböző vetemítések esetén (fekete: referencia, világosszürke: bemondott mondat lineáris vetemítéssel, sötétszürke: bemondott mondat dinamikus vetemítéssel)

Az átlagos négyzetes távolság lineáris vetemítés esetén 13,5 Hz lett, ezzel szemben dinamikus vetemítés esetén 10,1 Hz.

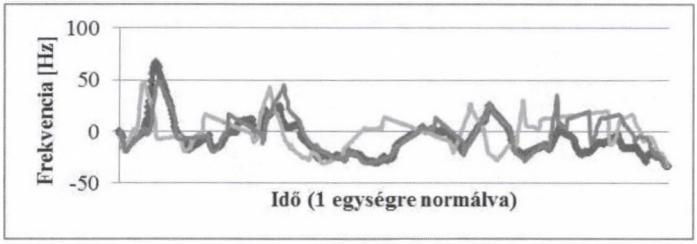


4. ábra
A referenciamondat és a „jól" bemondott mondat intenzitás dinamikájának összehasonlítása, különböző vetemítések esetén (fekete: referencia, világosszürke: bemondott mondat lineáris vetemítéssel, sötétszürke: bemondott mondat dinamikus vetemítéssel)

Az átlagos négyzetes távolság lineáris vetemítés esetén 6,9 dB, ezzel szemben dinamikus vetemítés esetén 5,5 dB lett. A 3. ábrán látható a két „ugyanúgy" bemondott mondat esetén a mondatintonáció (dallammenet), illetve a 4. ábrán az intenzitás dinamikája az egyes vetemítések esetén. Az ábrák alapján levonhatóak az első teszt tanulságai az időben nemlineáris vetemítésről. Jól látható, hogy az időben nemlineárisan vetemített görbe (sötétszürke) lényegében megegyezik a referenciagörbe dinamikájával (fekete). A lineárisan vetemített görbe (világosszürke) is nagyjából követi a referenciagörbe dinamikáját, bár nem teljesen, az eltérés láthatóan nem jelentős, attól még az időben nemlineárisan vetemített görbe láthatóan pontosabban követi a referenciát. Ez abból adódik, hogy ebben az esetben a referenciabemondás és a vetemítendő mondat időszerkezete lényegében megegyezik, és így láthatjuk, az időben nem lineáris vetemítés nem ad rosszabb eredményt, mint a lineáris.

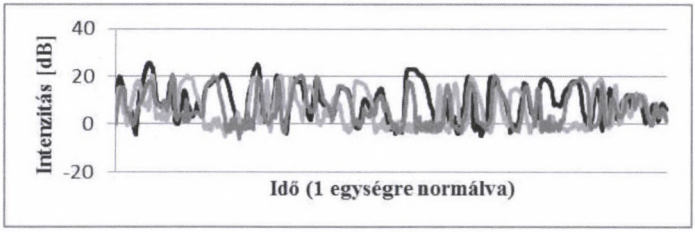
Az előzővel szemben, a második teszt esetén szembetűnő különbség mutatkozik a két vetemítési eljárás között (5., 6. ábra). Az ábrákon kifejezetten jól látszik az időben nem lineáris vetemítés előnye. A sötétszürke görbe (időben nemlineárisan vetemített) lényegében együtt mozog a referenciagörbével (fekete). Ezzel szemben a lineárisan vetemített görbe (világosszürke) a referenciától lényegesen eltérően mozog helyenként ellenkezően, mint a referenciagörbe. Ez a jelenség a nemlineáris megnyúlásokból és rövidülésekből adódik. Tehát ilyen esetben, ha lineáris vetemítéssel hasonlítanánk össze a két mondatot, akkor a prozódiai kiértékelő program több helyen is hibát jelezne a

lineáris skálázás esetén, míg az időben nemlineáris vetemítés esetén feltehetőleg azt közölné, hogy a bemondás prozódiai szempontból helyes.



5. ábra
A referenciamondat és az időben „rosszul" bemondott mondat alaphang-dinamikájának összehasonlítása, különböző vetemítések esetén (fekete: referencia, világosszürke: bemondott mondat lineáris vetemítéssel, sötétszürke: bemondott mondat dinamikus vetemítéssel)

Az átlagos négyzetes távolság lineáris vetemítés esetén 22,1 Hz, ezzel szemben dinamikus vetemítés esetén 11,5 Hz lett.



6. ábra
A referenciamondat és az időben „rosszul" bemondott mondat intenzitás dinamikájának összehasonlítása, különböző vetemítések esetén (fekete: referencia, világosszürke: bemondott mondat lineáris vetemítéssel, sötétszürke: bemondott mondat dinamikus vetemítéssel)

Az átlagos négyzetes távolság lineáris vetemítés esetén 11,4 dB, ezzel szemben dinamikus vetemítés esetén 7,4 dB lett.

Összegzés

Az eddig tárgyalt eredmények alapján kijelenthetjük, hogy az általunk tervezett eljárás működik, és az alkalmas lehet két mondat prozódiai távolságának kiértékelésére, illetve azok korrekt módon való vizuális ábrázolására több nyelv esetén is. Ez az eljárás segítheti a beszédoktató programok hatéko-

nyabb használatát, pontosabb, illetve összetettebb visszajelzést adhat a tanulóknak.

A 6. táblázatban látható, hogy az időben lineáris, illetve az időben nemlineáris vetemítés esetén, egy referenciamondattól mekkora a prozódiai távolság az egyes jellemzők alapján (alaphang, intenzitás), ha a bemondó törekszik arra, hogy „jól" mondja vissza a mondatokat úgy, hogy az első esetben időszerkezetileg is hasonlóan mondja be, míg a második esetben erre nem törekszik.

6. táblázat: Az alaphang, illetve az intenzitás átlagos négyzetes távolsága a referenciától lineáris és dinamikus vetemítés esetén

	Azonos tempóban bemondott	**Más tempóban bemondott**
Alaphang:		
Lineáris vetemítés esetén	13,5 Hz	**22,1 Hz**
Dinamikus vetemítés esetén	10,1 Hz	**11,5 Hz**
Intenzitás:		
Lineáris vetemítés esetén	7,2 dB	**11,4 dB**
Dinamikus vetemítés esetén	5,5 dB	**6,8 dB**

A 6. táblázatból leolvasható, hogy a nem megfelelő vetemítés nagyban megnöveli az átlagos négyzetes távolságot. A 6. táblázat, a 6. és a 7. ábra alapján jól látható, hogy az időben nemlineáris vetemítés jobb eredményt ad vissza, mint a lineáris. Az is látható, hogy az időben nemlineáris vetemítés szükséges a helyes összehasonlításhoz, elengedhetetlen a korrekt vizuális megjelenítéshez. Ezek alapján kijelentjük, hogy egy időben nemlineáris vetemítő eljárást használó tanító program interaktívabb, magasabb szintű tanulási élményt nyújthat a tanuló számára. Részletesebb, pontosabb visszajelzéseket adhat, ami a lineáris vetemítés esetén nem volna lehetséges. És mindez több nyelv esetén is megvalósítható ugyanazzal az eljárással.

Irodalom

Baker, Janet – Deng, Li – Glass, James – Khudanpur, Sanjeev – Lee, Chin Hui – Morgan, Nelson 2009a. 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. 76–85.

Benno, Peters 2005. *The Kiel Corpus of spontaneous speech.* http://www.ipds.uni-kiel.de/kjk/pub_exx/aipuk35a/aipuk35a_1.pdf. (A letöltés ideje: 2013.január.12.)

Boersma, Paul – Weenink, David 2001. Praat, a system for doing phonetics by computer. *Glot International.* 341–345.

Garofolo, John S. – Lamel, Lori – Fisher, William – Fiscus, Jonathan – Pallet, David – Dahlgren, Nancy – Zue, Victor 1993. *TIMIT Acoustic-Phonetic Continuous Speech Corpus.* Linguistic Data Consortium, Philadelphia.

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

Mihajlik Péter 2010. Rejtett Markov-modellek. In Németh Géza – Olaszy Gábor (szerk.): *A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.* Budapest, Akadémiai Kiadó. 242–243.

Olaszy Gábor 2010. A beszéd szupraszegmentális szerkezete. In Németh Géza – Olaszy Gábor (szerk.): *A magyar beszéd. Beszédkutatás, beszédtechnológia, beszédinformációs rendszerek.* Akadémiai Kiadó, Budapest. 171–204.

Szaszák György – Nagy Katalin – Sztahó Dávid – Vicsi Klára 2009. Automatikus intonációs osztályozó felhasználása hallássérültek beszédterápiájában. In Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): *VI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2009).* Szeged. 381–385.

Sztahó Dávid – Nagy Katalin – Vicsi Klára 2009. Subjective tests and automatic sentence modality recognition with recordings of speech impaired children. In Esposito, Anna – Campbell, Nick – Vogel, Carl – Hussein, Amir – Nijholt, Anton (eds.): *Developement of multimodal interfaces: Active listening and synchrony: Second COST 2102 International School Dublin.* Dublin, Ireland. 397–405.

Vicsi Klára – Vig Attila 1998. LIAS: Language Independent Automatic Segmentation technique using SAMPA labeling of phonemes. In: *Proceedings on First International Conference on Language Resources & Education.* Granada. 1317–1323.

Vicsi Klára – Kocsor András – Teleki Csaba – Tóth László 2004. Beszédadatbázis irodai számítógép-felhasználói környezetben. In Alexin Zoltán – Csendes Dóra (szerk.): *Second Conference on Hungarian Computational Linguistics (MSZNY 2004).* Szeged. 348–359.

Ez a kutatás az Alap- és alkalmazott kutatások hallássérültek Internetes beszédfejlesztésére és az előrehaladás objektív mérésére TÁMOP-4.2.2.C-11/1/KONV projekt és a „FuturICT.hu" TÁMOP-4.2.2.C-11/1/KONV projekt keretein belül készült.