Automatikus intonációs osztályozó felhasználása hallássérültek beszédterápiájában

Szaszák György, Nagy Katalin, Sztahó Dávid, Vicsi Klára

Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék, e-mail:{szaszak, sztaho, vicsi}@tmit.bme.hu

Kivonat A BME-TMIT egy prozódiai rejtett Markov-modell alapú modalitásfelismerőt fejlesztett ki, amely szupraszegmentális akusztikai előfeldolgozás után tagmondatok és mondatok határait és a mondat modalitását ismeri fel. Cikkünkben bemutatjuk a modalitásfelismerő automatikus intonációs osztályozásra való felhasználását hallássérültek vagy idegen nyelvet tanulók beszédterápiájában. A rendszer teljesítményét ép hallású bemondóktól származó anyagon optimalizáljuk, majd vizsgáljuk a hallássérült bemondók által bemondott mondatok automatikus osztályozásában. A jobb összehasonlíthatóság érdekében az eredményeket szubjektív lehallgatási tesztek eredményeivel is összevetjük.

1. Bevezetés

A beszéd szupraszegmentális szintje - a prozódia - igen fontos az emberi beszédpercepcióban, és hatékonyan felhasználható a gépi beszédtechnológiában is [1]. A jó minőségű beszédszintézis például elképzelhetetlen a prozódia megfelelő modellezése nélkül [2]. A prozódia beszédfelismerésbeli felhasználása kevésbé elterjedt, mindazonáltal számos kutatás igazolja, hogy a beszédfolyam automatikus tagolásában, a beszédfelismerés eredményességének növelésében, a szintaktikai és szemantikai szintű információ kinyerésében fontos szerepe van (Vö.: [4], [5], [6]).

Az emberi beszédben gyakorlatilag a prozódia az egyetlen akusztikai jellemző, amely a modalitásra utal, néhányan ezt a lehetőséget is vizsgálták már [7], [8]. Az utóbb hivatkozott műben a szerzők olyan rejtett Markov-modell (HMM) alapú rendszert muattak be, amely az F0 és az energia menete alapján végez modalitásfelismerést. Jelen cikkükben a szerzők ezt modalitásfelismerőt vizsgálják beszédterápiás rendszerbe ágyazottan.

A számítógépes beszédterápiás rendszerek interaktív felületet biztosítanak a nyelvtanulónak, amelyet a hallássérültek hatékonyan használhatnak helyes beszéd - a helyes artikuláció vagy a helyes hangsúlyozás és intonáció - elsajátításához. A vizuális visszacsatolás révén ugyanis értékelhetik saját kiejtésüket, "produktumukat", ily módon kiváltva a hiányzó auditív visszacsatolást [9]. A módszert a prozódia elsajátítására használva bizonyított, hogy a vizuális visszacsatolás hatékonyabb, mint a puszta auditív [10], különösen, ha a tanuló referenciamintát is lát - például a kívánatos F0-kontúrét.

A legtöbb napjainkban elérhető beszédterápiás rendszer a helyes artikuláció tanítására koncentrál, emellett a prozódia gyakran elhanyagolt szerepbe szorul. A létező alkalmazások egy csoportja távolságszámítás alapján automatikusan értékeli a tanuló kiejtését (vö. SPECO, [11]), míg más rendszerekben HMM fonéma modelleket használnak a kiértékeléshez [12].

Célunk a prozódia oktatása és automatikus kiértékelésének megvalósítása magyar nyelven. Az így előálló rendszert hallássérült gyerekek használhatják a helyes hangsúlyozás és a modalitásnak megfelelő intonáció elsajátítására. Az automatikus kiértékelés elvégzésére a már említett modalitásfelismerőt adaptáljuk [8], ennek során egy speciálisan erre a célra kialakított ún. intonációs beszédadatbázist is felhasználunk.

2. A modalitásfelismerő

Jelen cikk alapja a korábban már részletesen bemutatott [8] HMM alapon intonáció osztályozását végző modalitásfelismerő. Ez az osztályozó magyar nyelvre 7 különböző modalitás elkülönítésére alkalmas, pontosabbana szükséges csönd és nem mondatzáró modelleket leszámítva a véglegesen elkülönítendő modalitások száma 5, mégpedig: kijelentő, kiegészítendő kérdő, igen-nem kérdő, felkiáltó vagy felszólító, választó.

3. Az intonációs adatbázis

A modalitásfelismerő betanításához külön intonációs adatbázis készült a budapesti Dr. Török Béla - hallássérültekre specializált - Általános Iskolában. Az adatbázis anyagát a tervezett feladatoknak megfelelően állítottuk össze: abban minden modalitású mondat előfordul, mind hosszabb, mind rövidebb, akár egyetlen szóból álló mondat formájában. A felvételeket 60 ép hallású és 19 hallássérült gyermekkel készítettük el. Az előbbi csoport a betanításhoz, míg utóbbi a végső rendszer teszteléshez szükséges.

Az adatbázisban az egyes modalitásoknak megfelelően címkéztük az intonációs kontúrokat. A címkézés kritériuma a megvalósult intonáció, amelyet szakértő ítélt meg. A nem pontos vagy nem helyesen intonált bemondásokat nem használtuk fel. Az osztályozás során használt osztályokat és megfelelő címkéiket az 1. táblázat tartalmazza. Ne feledjük, hogy az 1. táblázatban szereplő 6 osztályon kívül a csönd is modellezendő.

4. Az intonációs sémák betanítása

Az intonációs sémák HMM-jeit az 1. táblázatban szereplő osztályokra az intonációs adatbázis ép hallású beszélőkkel készített részének 2/3-án tanítottuk be. A fennmaradó 1/3 validálási célokat szolgál. A tanított HMM-ek 7 állapotú, balról jobbra felépítésű, a kibocsátási valószínűséget 1 vagy 2 Gauss komponenssel leíró modellek. A használt prozódiai-akusztikai jellemzők az F0 és az energia.

1. táblázat. *A címkézéshez használt intonációs osztályok.*

Intonáció	Címke	Példa	Kontúr
Ereszkedő	DE	Anna áll.	
Eső	FA	Miért áll ott?	
Emelkedő-eső	AF	Anna áll ott?	
Eső-ereszkedő	FD	Gyere ide!	
Lebegő	FL	Ez Anna, és ...	
Emelkedő	RI	Nem?	

Előbbit oktávugrások ellen szűrjük, és logaritmikus tartományban lineárisan interpoláljuk a zöngétlen helyeken. Mindkét jellemző értékét 25 pontos átlagoló szűrővel szűrjük 10 ms keretidő mellett, majd első és második deriváltjaikat is kiszámítjuk.

5. Validálás

Az intonációs osztályozóként használandó modalitásfelismerő előzetes tesztelése az ép hallású bemondások betanításból kihagyott 1/3-án történt. Az egyes mondatokból olyan csoportokat képeztünk, amelyek a beszédterápiás eszközben egy-egy konrét feladatnak felelnek meg. Az eredmények tévesztési mátrix formájában a 2. táblázatban láthatóak (%-os értékekel megadva). Az eső ereszkedő osztályt (FD) az optimalizálás során az esőbe (FA) olvasztottuk be.

2. táblázat. *Tévesztési mátrix az ép hallású gyermekek által produkált intonáció gépi osztályozásában.*

Referencia	Osztályozás [%]				
	DE	FA	AF	FL	RI
DE	**97.67**	2.33	0.00	0.00	0.00
FA	1.61	**82.26**	8.06	6.45	1.61
AF	0.00	0.00	**93.10**	3.45	3.45
FL	2.56	2.56	2.56	**92.31**	0.00
RI	0.00	0.00	0.00	0.00	**100.0**

6. Az intonációs osztályozás tesztelése

Az intonáció osztályozására használt modalitásfelismerő végső tesztelése a hallássérült, és emiatt beszédhibával is rendelkező gyerekektől származó felvételeken

történt. Az osztályozás szerepe ebben az esetben a kiejtés intonáció szempontjából történő értékelése, a kiejtést akkor tekintjük helyesnek, ha a modalitásfelismerő a kívánt intonációt ismeri fel. Ezek a tesztek egyben megfelelnek a modalitásfelismerő beszédterápiás rendszerben történő használatának. A teszteredmények az 3. táblázatban láthatók. Felhívjuk a figyelmet arra, hogy az eredmények nem a modalitásfelismerőt minősítik (arra ugyanis a 2. táblázat vonatkozik), hanem azt mutatják, hogyan alakult a gyermekek által helyesen vagy helytelenül kiejtett intonációinak aránya az egyes intonációtípusokéra a gépi osztályozás esetében.

3. táblázat. *Beszédhibás gyermekek által produkált intonáció osztályozása modalitásfelismerővel.*

Kívánt	Osztályozás [in]				
kiejtés	DE	FA	AF	FL	RI
DE	**33.0**	35.0	0.0	32.0	0.0
FA	9.5	**62.3**	0.0	28.1	0.0
AF	15.5	15.5	**53.5**	15.5	0.0
FL	16.9	32.3	0.0	**50.7**	0.0
RI	0.0	10.0	0.0	30.0	**60.0**

A tesztek alaposabb kiértékelésének érdekében emberi hallgatók is értékelték a beszédhibás gyermekek által használt intonációt szubjektív lehallgatási tesztek keretében. A 21 hallgató ugyanazokra az intonációosztályokra osztályozott, mint a gépi rendszer azzal a kivétellel, hogy a szubjektív hallgatók teljes bizonytalanság (UC) esetén kihagyhatták az adott elem értékelését. Az eredmények a 4. táblázatban láthatók.

4. táblázat. *Beszédhibás gyermekek által produkált intonáció osztályozása szubjektív lehallgatási tesztek során.*

Kívánt	Osztályozás [%]					
kiejtés	DE	FA	AF	FL	RI	UC
DE	**89.0**	1.0	1.5	5.5	0.5	2.5
FA	17.0	**75.0**	1.5	0.5	0.0	6.0
AF	11.4	2.5	**79.6**	0.5	1.0	5.0
FL	44.0	3.5	10.5	**33.5**	0.0	8.5
RI	17.0	1.0	0.5	3.0	**70.0**	8.5

A szubjektív lehallgatási tesztek és az automatikus osztályozás eredményeit összevetve az osztályozási teljesítmények jól párhuzamba állíthatók, kivéve az ereszkedő (DE) és a lebegő (FL) intonációtípusokat. Ennek oka az, hogy a szubjektív lehallgatók valószínűleg ódzkodtak a kissé szofisztikált lebegő kategória

használatától, és akkor is ereszkedő intonációra döntöttek, ha az intonáció valójában bizonytalan, lebegő volt (mintegy alkalmazkodtak a beszédhibás beszélő beszédmódjához). Ugyanerre vezethetők vissza a szubjektív lehallgatás során tapasztalt nagyobb elfogadási hajlandóság, illetve arra is, hogy a szubjektív lehallgatók nyelvtani információra is támaszkodhattak a lehallgatás során, jóllehet természetesen azt az utasítást kapták, hogy a grammatikai vonatkozásoktól tekintsenek el.

Az eredményeket részletesen összehasonlítva azt tapasztaltuk, hogy a szubjektív lehallgatók legalább 50%-a által a kívánttal megegyezőnek elfogadott intonációt a gépi osztályozás csupán az esetek 9%-ában nem fogadta el. A gépi osztályozás tehát szigorúbb, de véleményünk szerint elfogadható osztályozást valósít meg, ami kívánatos is a helyes kiejtés elsajátításában, hiszen a helyes, és nem a még elfogadható kiejtésformák megerősítése az elsődleges cél.

Hivatkozások

[1] Kompe, R.: Prosody in Speech Understanding Systems. LNAI 1307, Springer (1997)
[2] Fujisaki, H., Ohno, S.: The Use of a Generative Model of F0 Contours for Multilingual Speech Synthesis. 4^{th} Int. Conf. on Signal Proc., Vol. 1 (1998) 714–717
[3] Hunyadi, L.: Hungarian Sentence Prosody and Universal Grammar. Peter Lang (2002)
[4] Szaszák, Gy., Vicsi, K.: Using Prosody in Fixed Stress Languages for Improvement of Speech Recognition. In: A. Esposito et al. (eds.): Verbal and Nonverbal Communication Behaviours. Springer. (2007) 138-150
[5] Hirose, K. et al.: Continuous Speech Recognition of Japanese Using Prosodic Word Boundaries Detected by Mora Transition Modeling of Fundamental Frequency Contours. ISCA Tutorial and Research WS on Prosody. Red Bank, USA (2001) 61-66
[6] Veilleux, N. M., Ostendorf, M.: Prosody/parse scoring and its application in ATIS. In: Proc. of ARPA Human Language Technology Workshop '93 (1993) 335-40
[7] Král, P., Klečková, J. Cerisara C.: Sentence Modality Recognition in French based on Prosody. In: Proc. of World Academy of Science, Engineering and Technology, Vol. 8 (2005) 185-188.
[8] Vicsi, K., Szaszák, Gy.: Using Prosody for the Improvement of ASR: Sentence Modality Recognition. Interspeech 2008, ISCA Archive. http:www.isca-speech.org/archive/ (2008)
[9] Vicsi, K.: Computer-Assisted Pronunciation Teaching and Training Methods Based on the Dynamic Spectro-Temporal Characteristics of Speech. In: Divenyi, P. L. et al. (eds.): Dynamics of Speech Production and Perception. IOS Press (2006) 283-304
[10] James, E.: The acquisition of prosodic features of speech using a speech visualizer. IRAL, 14(3) (1976) 227-243
[11] Vicsi, K., Csatári, F., Bakcsi, Z., Tantos, A.: Distance score evaluation of the visualized speech spectra at audio-visual articulation training. In: Proc. Eurospeech (1999) 1911-1914
[12] Narusa, J.: Computer-aided spoken language training with enhanced visual and auditory feedback. In: Proc. Eurospeech (1999) 183-186