Prozódiai információ felhasználása a beszédfelismerés hatékonyságának növelésére

SZASZÁK GYÖRGY, VICSI KLÁRA

BME Távközlési és Médiainformatikai Tanszék
{szaszak, vicsi}@tmit.bme.hu

Lektorált

Kulcsszavak: beszédfelismerés, prozódia, szóhatár-detekció, prozódiai szegmentálás

Cikkünkben lényegében a mondat-, tagmondat- és szóhatárok prozódia alapú detektálását mutatjuk be rejtett Markov modelles módszerrel. Az így elkészült prozódiai szegmentálót beszédfelismerőbe építve a felismerési hipotéziseket újrasúlyozzuk annak alapján, hogy mennyire illeszkednek a detektált dallammenetre. Ultrahangos leletező alkalmazásban egyszerűsített nyelvi modellel a felismerési hatékonyság 3,82%-os javulását értük el.

1. Bevezetés

A prozódia – vagy szupraszegmentális szerkezet – az emberi beszéd szerves részét képezi, így például a hallgató számára segíti a közlés értelmezését azáltal, hogy a beszédet tagolttá teszi, kiemeli a fontos vagy új információt tartalmazó részeket. E funkcióján túlmenően hordozza a modalitást (kijelentő, kérdő stb.), illetve lehetőséget ad a beszélőnek érzelmei kifejezésére is, ami jelentős részben szintén a szupraszegmentumok révén valósul meg.

A beszédtechnológia műszaki oldaláról közelítve ma már elképzelhetetlen jó minőségű beszédszintézis a megfelelő prozódia – azaz a megfelelő hangsúlyozás és a természetes dallammenet – modellezése nélkül. A beszédfelismerésben azonban korábban szinte egyáltalán nem foglalkoztak a prozódiával, jóllehet a szupraszegmentumoknak nem csak a jelentést tagoló vagy árnyaló, hanem bizonyos esetekben a jelentés egy részét illetően egyfajta „hordozó” szerepük is van. Mindezt a gépi beszédfelismerés során is figyelembe kell vennünk, ha nem szeretnénk a közlésből lényeges információt elveszíteni. Így például bizonyos beszédinformációs rendszerekben alapvető követelmény lehet, hogy különbséget tudjunk tenni kérdések és kijelentések között akkor is, ha adott esetben a kérdő- és a kijelentő mondat ugyanazon szóláncból épül fel, különbség közöttük csak modalitásukban van [1]. Hagyományos beszédfelismerő rendszerrel – amely csupán az elhangzott beszédhang-szekvenciára koncentrál – ez a feladat megoldhatatlan.

Ezen a „magától értetődő” felhasználási területen túl a szupraszegmentális jellemzők alakulásának nyomon követése a beszédfelismerési feladat során egyéb haszonnal is járhat, amennyiben a beszédfolyam prozódiai eszközökkel történő tagolása segítségünkre van a beszédfelismerésben. A mondatok, tagmondatok, szószerkezetek vagy akár az egyes szavak határainak ismerete hasznos lehet a keresési tér csökkentésében, ha ugyanis biztosak vagyunk benne, hogy valahol a beszédfolyamban szóhatárt találunk, akkor azokat a felismerési hipotéziseket, amelyek az adott ponton nem tartalmaznak szóhatárt, kizárhatjuk. Ezáltal gyorsul és pontosabb lesz a felismerés, amelynek során – különösen a toldalékoló nyelvek esetében, amilyen a magyar nyelv is – sokszor problémát, de legalábbis korlátozó tényezőt jelent a valós idejű működés követelménye. További segítséget adhat a prozódia a lényeges információ automatikus kiemeléséhez a közlésből, illetve segítheti a szintaktikai elemzést is [3].

A nemzetközi porondon többen próbálkoztak már a prozódiai információ felhasználásával a beszédfelismerésben, elsősorban angol és német nyelven. Veilleux és Ostendorf például olyan algoritmust dolgoztak ki [10], amelyek az egyes hipotézis-gráfok közül az N darab legvalószínűbbet újrasúlyozzák (úgynevezett N-best rescoring) a prozódiai információk ismeretében, majd ezután az újrasúlyozott gráfokkal számítják a felismerés végeredményét a hagyományos módon. Hasonló munka készült a német nyelvre is [2]. Gallwitz és munkatársai ennél tovább lépve integrált gépi beszédfelismerőt készítettek [1], amely a beszédláncra vonatkozó, illetve a prozódiai információt egységesen kezeli és követi a felismerés során. A szerzők is végeztek már kísérleteket magyar nyelvre a prozódia beszédfelismerésben történő felhasználhatóságát illetően [12].

2. A prozódiai információ kinyerése a beszédből

A prozódiai jellemzők közül az alapfrekvencia, az energiaszint, és az időtartamok objektív mérésére van lehetőség. Ezen paraméterek közül az alapfrekvencia és az energiaszint értékeit találtuk jellemzőnek a hangsúly detektálása szempontjából korábbi vizsgálataink alapján [8]. E két prozódiai tényező értékeit a beszédjelből a Snack programcsomag [7] segítségével nyertük ki, majd előfeldolgozásnak vetettük alá.

Az alapfrekvencia számításához a Snack programban implementált AMDF-alapú algoritmust használtuk 25 ms ablakmérettel, 10 ms keretidővel. Ezt követően okta-

vugrást korrigáló szűrőt alkalmaztunk, amelyet 5 pontos átlagoló (mean) szűrő követett, majd az alapfrekvencia-értékek logaritmusát alapul véve lineáris interpolációt végeztünk annak érdekében, hogy az alapfrekvencia görbe többé-kevésbé folytonos legyen. Nem történt interpoláció olyan zöngétlen szakaszokon, amelyek hossza a 250 ms-ot meghaladta, illetve akkor sem, ha a zöngétlen szakasz utáni első zöngés keret alapfrekvenciája meghaladta a zöngétlen rész előtti 3 utolsó keret alapfrekvencia-értékei átlagának 1,1-szeresét.

Minderre azért volt szükség, hogy a 250 ms-nál hosszabb, ezért igen nagy valószínűséggel beszédszünetet tartalmazó szakaszokon az alapfrekvenciát ne interpoláljuk, mivel egyrészt a szünetet magát is szeretnénk a későbbiekben detektálni, másrészt ilyen hosszú szakaszon az interpolálás már túl durva közelítés lenne. Az alapfrekvencia-érték zöngétlen szakasz utáni emelkedését a zöngétlen szakasz előtti utolsó három érték átlagánál pedig azért nem engedjük magasabbra, mert ekkor valószínűbb, hogy a kérdéses szakaszon mondat, tagmondat vagy szószerkezet határa volt, és emiatt indít magasabbról az alapfrekvencia. E fenti értékeket tapasztalati úton állítottuk be, de a jövőben célszerű lehet ezeket a beszélőtől (beszédtempó, artikulációs sebesség stb.) függően meghatározni, ehhez azonban további vizsgálatok szükségesek, így a továbbiakban ezzel egyelőre nem foglalkozunk.

Az energiaszint-értékeket szintén a Snack programmal számítottuk, a keretidő ismét 10 ms volt, az alkalmazott ablakméret 25 ms, melyet szintén átlagoló (mean) szűrés követett. Mivel az energiaszint folytonosan számítható a beszédjelre, ezért itt interpolációra értelemszerűen nem volt szükség.

Ezután mind az alapfrekvencia, mind az energiaértékekhez első és másodrendű deriváltjaikat is kiszámítottuk. A deriváltak közelítésére alkalmazott (1) regressziós képletben a figyelembe vett környezetet három lépcsőben fokozatosan növelve valójában 3-3 első és másodrendű deriváltat képeztünk, rendre ±10, ±25 és ±50 keretnek megfelelően ablakolt (*W* az (1) képletben) minták alapján, így a véglegesen kapott jellemzővektor összesen 14 elemet tartalmazott: az eredeti, feldolgozott alapfrekvencia- és energiaértéket, és ezek mindegyikéhez 3-3 első- és másodrendű deriváltat. A deriváltak számítására használt képlet ([9] alapján):

$$d_t = \frac{\sum_{i=1}^{W} i(c_{t+i} - c_{t-i})}{2\sum_{i=1}^{W} i^2}, \qquad (1)$$

ahol d_t a t időpontban értelmezett derivált, c_{t-i} és c_{t+i} az eredeti (deriválandó) együtthatók, W pedig az ablakméret keretszámban.

3. A prozódiai információ felhasználása a beszédfelismerésben

A prozódia vizsgálatával és modellezésével célunk a beszéd durva felszegmentálása mondat- és tagmondat-határokon, illetve a szavak, szószerkezetek határainak minél pontosabb meghatározása. Az így nyert információt ezután a beszédfelismerőben felhasználva a felismerés hatékonysága javítható, illetve új, a beszédfelismerő felhasználási lehetőségeit kibővítő – a bevezetőben már említett – egyéb funkciók is megvalósíthatóak lesznek.

Munkánk során nagyban támaszkodunk arra a tényre, hogy a magyar nyelv kötött hangsúlyozású, azaz a hangsúly mindig a hangsúlyos szó első szótagjára esik [2]. Mindez azért rendkívül fontos, mert így lehetőségünk nyílik a prozódiai információ egységes kezelésére anélkül, hogy a beszédet szavak vagy beszédhangok szintjén is ismernünk kellene a prozódiai struktúra értelmezéséhez. Természetesen végcélunk ezzel együtt az, hogy a beszédhangok sorozatát jellemző spektrális, illetve a prozódiát befolyásoló szintaktikai információt egységesen kezeljük és használjuk fel a beszédfelismerésben.

3.1. Automatikus prozódiai szegmentáló betanítása

A prozódiai szegmentálás során az egyes mondatépítő szintaktikai elemek dallamtípusát szeretnénk felismerni és járulékosan a szintaktikai egységek határát a lehető legnagyobb pontossággal meghatározni. A szintaktikai egységek határai egyben szóhatárok is lesznek, amelyek egybeeshetnek tagmondatok vagy mondatok határaival is.

A felismerni szándékozott dallamtípusok kiválasztásánál tekintettel kell lennünk arra, hogy az egyes dallamok egymástól élesen elkülöníthetőek legyenek és felöleljék a leggyakrabban előforduló dallamváltozatokat. Az éles különbségtétel követelménye miatt mindössze 6 alapvető dallamtípust különíttettünk el a prozódiai felismeréshez. A szünet adja a 7. felismerendő „dallamtípust". A felhasználandó dallamtípusokat az *1. táblázatban* foglaltuk össze.

A prozódiai szegmentáló betanításához a tanítómintákat a BABEL beszédadatbázis [6] szöveganyagából választottunk ki (22 beszélő által bemondott 1600 mon-

1. táblázat
A felismerésre kiválasztott dallamtípusok

Címke	Dallam	Megjegyzés
me	változó	Mondat eleje.
fe	(emelkedő-) eső v. eső-ereszkedő	Erősen hangsúlyos szintaktikai egység.
fs	eső-ereszkedő	Mellékhangsúlyos szintaktikai egység.
mv	ereszkedő	Mondat vége.
fv	emelkedő	Folytatást jező szintaktikai határ.
s	ereszkedő vagy lebegő	Hangsúlytalan szakasz. Szünetet is kitölthet az F_0 interpolációja miatt.
sil	–	Szünet.

dat). Ezt a szöveganyagot kézzel, majd félautomatikusan felszegmentáltuk a táblázatban szereplő dallamtípusok szerint. A kézi szegmentálás az alapfrekvencia- és az energiakontúr alapján történt, a lehallgatás során kapott szubjektív ítéletet is figyelembe véve.

Az elkészült prozódiai szegmentáló rejtett Markov-modell alapú, a keretidő a már említett 10 ms, a Markov-modellek lineárisak, állapotai száma (optimalizálás után) 11. A prozódiai szegmentálót a HTK szoftvercsomag [9] felhasználásával valósítottuk meg.

3.2. Az automatikus prozódiai szegmentálás menete

Az automatikus prozódiai szegmentálás menete a beszédfelismerésben is használt algoritmusokkal az ott ismeretes lépések szerint történik, azaz a lényegkiemelést (előfeldolgozást) követi a dekódolás. Az előfeldolgozás a 2. szakaszban megismertek szerint történik, azaz az alapfrekvencia- és az energia-jelet az ott ismertetett módon nyerjük ki és dolgozzuk fel. A dekódolás során a Viterbi-algoritmus fut, amely a „rövid" jellemző vektorok, a modellek csekély száma és az alkalmazott nyelvtan miatt igen gyors.

A dekódolás során ugyanis a szintaktikai egységeknek megfelelő dallamtípusokra vonatkozóan szigorú nyelvtant vezetünk be, amely megadja, hogy azok milyen sorrendben követhetik egymást. Az ily módon létrehozott megszorítások tapasztalataink szerint a szegmentálás minőségét lényegesen javítják, ugyanakkor ehhez képest elhanyagolható azoknak az eseteknek a száma, amikor a szigorú, nem minden kivételes esetet leíró nyelvtan miatt történik tévesztés.

A nyelvtan a HTK-ban is alkalmazott jelöléseket alapul véve (vö. [9], p.163.)

$$Phrase = [sil] < [me]\{fe \mid fv[s]\}[mv][sil] > sil \quad (2)$$

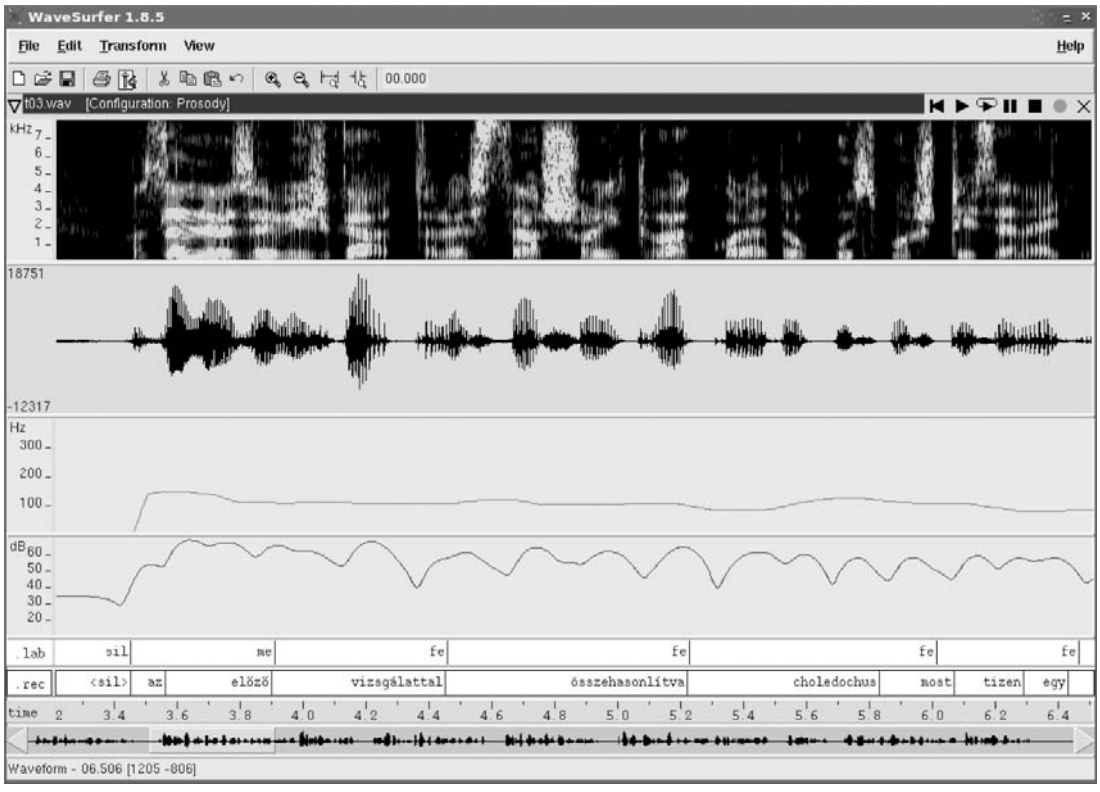
összefüggéssel írható le, melyben a '<>' szimbólumok egy vagy több, a '{}' nulla, egy vagy több ismétlődést jelölnek. A '|' szimbólum kizáró vagy kapcsolatot, a '[]' opcionálisan elmaradó eseményeket jelöl. Az ily módon formalizálva lejegyzett sorozatot tekintjük a prozódiai mondatmodellnek. A prozódiai szegmentálás eredményeként a felismert dallamtípusok kezdő- és végidőpontjukkal együtt ismeretté válnak. Az *1. ábrán* látható példa egy így nyert prozódiai szegmentálást jelenít meg.

3.3. Prozódiai szegmentáló beépítése beszédfelismerőbe

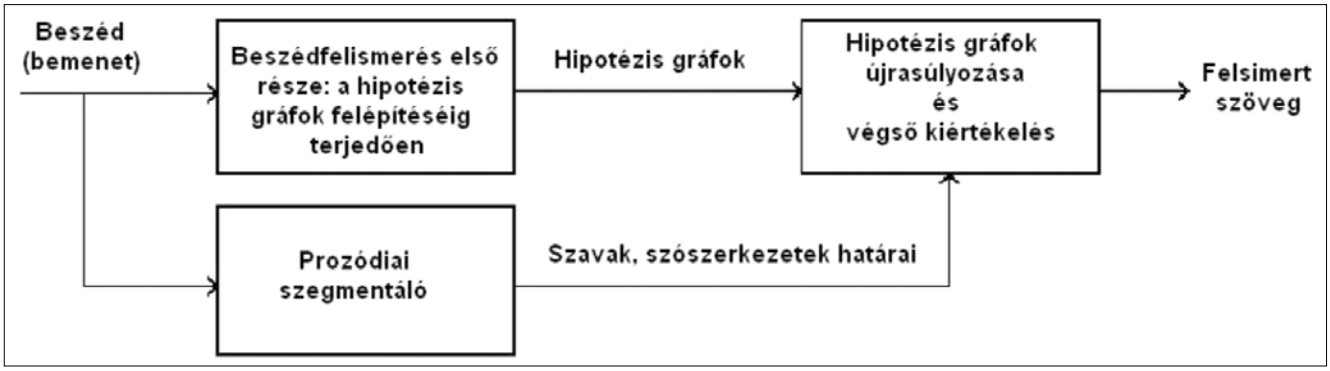
A prozódiai szegmentáló kimenetét felhasználhatjuk a beszédfelismerőben a keresési tér csökkentésére, így jobb felismerés és gyorsabb működés remélhető a beszédfelismerőtől. Ehhez a beszédfelismerés folyamatába kell avatkoznunk. Erre egy lehetőség az a pont, ahol a dekódolás részeként a hipotézis-gráfok felépítése – azaz voltaképpen a felismerés lehetséges eredményeinek felmérése és valószínűségeik kiértékelése – történik.

A prozódiai szegmentálás ismeretében a hipotézisgráfok újrasúlyozhatók, a felismerés további folyamatába pedig már az újrasúlyozott gráfok kerülnek, így a felismerés végeredményének kiértékelését már a prozódia alapján nyert információ is befolyásolja. Utolsó lépésként az újrasúlyozott hipotézis gráfon kell a maximális pontszámú utat megkeresnünk, amihez gyors keresőalgoritmusok állnak rendelkezésünkre.

1. ábra A prozódiai szegmentálás kimenete
„Az előző vizsgálattal összehasonlítva a choledochus most 11 milliméteres..." mondatrészletre.



A felső sávban a spektrogram, az alatta levő sávokban rendre az időfüggvény (hullámforma), az interpolált alapfrekvencia, az energia görbéje, a prozódiai szegmentálás, végül az elhangzott szöveg látható bejelölt szóhatárokkal.



2. ábra Prozódiai szegmentálóval kiegészített beszédfelismerő felépítése

A felismerés menetét a *2. ábrán* követhetjük végig.

3.4. A hipotézis-gráfok újrasúlyozása

Mint az előzőekben láttuk, a hipotézis-gráfok újrasúlyozása a prozódiai szegmentálás alapján történhet. Az alapötlet az, hogy azokat a szavakat és szóláncokat (a hipotézis-gráfból kinyerhető szósorozatokat, ezek egy-egy lehetséges felismerést írnak le), amelyek esetén a szavak határai időben egybecsengenek a prozódiai szegmentálás által jelzett határokkal, valamilyen módon részesítsük előnyben a felismeréskor, azaz a hozzájuk rendelt valószínűségi súlyt növeljük. Hasonlóképp, azokban az esetekben, amikor a prozódiai szegmentáló által megadott határok szavak belsejébe esnek, az eredetileg hozzárendelt súlyokat csökkenthetjük.

Problémát okoz azonban, hogy a prozódiai szegmentáló sem működik hibamentesen, azaz bizonyos százalékkal téves ítéletet hoz a szintaktikai egységek határait illetően. Spontán beszédben még gyakoribbak azok a jelenségek, amelyek az automatikusan futó algoritmust megzavarhatják (gyakoribbak a szótévesztések, javítások, előfordulhat, hogy a prozódia eltorzul, ha a beszélő „mondat közben meggondolja magát” és máshogyan folytatja közlendőjét, a hevesebben kifejezett érzelmek is befolyásolhatják a prozódiát stb.). A prozódiai szegmentáló teljesítményének kiértékelésével korábban részletesen foglalkoztunk [8], jelen esetben pedig a hibaszázalék pontos számszerű ismerete nélkül is beláthatjuk, hogy az újrasúlyozás során valamennyire a prozódiai információt is fenntartással kell kezelnünk.

Ügyelnünk kell továbbá arra is, hogy a prozódiai információ – éppen szupraszegmentális jellegéből adódóan – időben kevésbé pontosan lokalizálható, mint az egyes beszédhangok – így akár az egyes szavak – határai. Gondoljunk például arra, hogy ha egy adott szintaktikai egység utolsó beszédhangjaként zöngétlen hangot (különösen is, ha zöngétlen réshangot) találunk, a dallamban számunkra az utolsó „biztos” támpontot a legutóbbi magánhangzó jelenti. Ez máris egy beszédhanghossznyi bizonytalanságot jelent, amit a prozódiai szegmentáló a beszédhangsor ismeretének hiányában nem tud feloldani.

Éppen ezért a prozódiai szegmentáló által megjelölt határokat intervallummá transzformáljuk, azaz megengedünk bizonyos ΔT csúszást a prozódiai szegmentáló által megállapított határhoz (t_B) képest. Az intervallumon belül a ténylegesen előrejelzett határtól való távolság függvényében értelmezzük a határ adott időpontban történő elhelyezkedésének valószínűségét, pontosabban egy azzal arányos pontszámot (L_B) az alábbiak szerint:

$$L_B(t) = \begin{cases} A\cos(\frac{\pi}{2\Delta T}t) + C, ha\ t \in [t_B - \Delta T, t_B + \Delta T] \\ 0 \quad egyébként \end{cases}, \tag{3}$$

ahol A és C konstansok. A kísérleteinkben ΔT értéke 10 keret, azaz 100 ms volt. A cosinus függvényt az egyszerűség kedvéért választottuk, de lényeges, hogy minél távolabb van a határ az előre jelzettől, annál kisebb pontszámot rendelünk hozzá.

Mindezek után rátérhetünk a hipotézis-gráfok tényleges újrasúlyozásának algoritmusára, amelyet az alábbiakban mutatunk be. Elöljáróban annyit jegyezzünk még meg, hogy a hipotézis-gráf éleihez szavak vagy szóláncok, csomópontjaihoz pedig a megfelelő kezdő- és végidőpontok vannak rendelve. Az újrasúlyozáshoz minden, a gráfban található szót vagy szóláncot kigyűjtünk, majd kezdő- és végpontjaira pontszámot számítunk, amely annál nagyobb (lásd (3)), minél közelebb van a prozódiai szegmentáló által jelzett határhoz a szó eleje, illetve vége:

$$Sc_{remum} = w_a L_B(t_{start}) + w_b L_B(t_{end}), \tag{4}$$

ahol t_{start} a szó gráf szerinti kezdő, t_{end} a szó gráf szerinti végpontjának felel meg (az időben), w_a és w_b pedig súlyok.

Ezt követően a szó valamennyi i keretére – az első és utolsó k darab keret kivételével – összegezzük $L_B(t_i)$ értékeket, ahol t_i az aktuális keretidő:

$$Sc_{punish} = \sum_{i=k+1}^{N-k-1} L_B(t_i), \tag{5}$$

A fenti képletben N a szóhoz tartozó összes keret száma, k=ΔT=100 ms pedig ésszerű választásnak kínálkozik.

A gráf éléhez tartozó új $Sc_{rescored}$ pontszám értéke pedig:

$$Sc_{rescored} = w_O Sc_{orig} + w_P (Sc_{remum} - Sc_{punish}), \tag{6}$$

ahol Sc_{orig} a gráf éléhez eredetileg tartozó, most felülbírált pontszám (élsúly), w_O és w_P pedig súlytényezők.

4. Kísérlet a prozódiai szegmentáló beszédfelismerő rendszerbe való illesztésére

A prozódiai szegmentáló elkészítésével célunk a beszédfelismerés hatékonyságának növelése volt, így a továbbiakban egy olyan kísérletről számolunk be, melynek során a prozódiai szegmentálót automatikus gépi beszédfelismerőbe építettük be. A 3.3. szakaszban már szó esett a beépítés módjáról, az előzőekben (3.4. szakasz) pedig áttekintettük azokat az algoritmikus változtatásokat, amelyeket a felismerés folyamatában szükséges eszközölnünk.

A kísérlethez magyar nyelvű, folyamatos beszédfelismerőt választottunk ki, amely az orvostudománybeli radiológiai leletezés területét, azon belül is a hasi és kismedencei ultrahangos vizsgálatok leletezésénél használt szótárkészletet öleli fel. A szótár elemszáma viszonylag csekély, mintegy 4000 szó. A területre bigram nyelvi modell is készült, azonban jelen kísérletben a bigram nyelvi modellt binarizáltuk, azaz csak azt tüntettük fel benne, milyen szavak után milyen szavak előfordulása megengedett a szövegben. Ennek célja egyúttal annak kipróbálása is, hogy képes-e a prozódiai információ minimális nyelvtani információ mellett a felismerés hatékonyságát javítani. Ezzel egyúttal a későbbiekben tervezett nagyszótáras alkalmazások felé is tekintünk, ugyanis nagy szótárméret esetén a nyelvi modell elkészítéshez rendkívül nagy szövegadatbázis kell, a nyelvi modell használata pedig rendkívül műveletigényes. Különösen igaz ez az agglutináló nyelvekre – így a magyarra is – amelyek esetén viszonylag kis tématerület esetén is relatíve nagy az előforduló szóalakok száma a toldalékoló jelleg miatt.

A felismerő HTK környezetben implementált, felépítését tekintve a „klasszikus” 39 MFCC együtthatót alkalmazó, a kibocsátási valószínűségeket 32 Gauss függvény szuperponálásával leíró, 10 ms keretidejű rendszer. A felismerő betanításához az MRBA adatbázis [11] mintegy 8 órányi, részben beszédhang szinten felszegmentált anyagát használtuk fel, összesen 37 beszédhang modell készült.

Ebbe a felismerőbe építettük bele a prozódiai szegmentálót, és vizsgáltuk a felismerési eredmény változását. A (4) és a (6) képletekben megadott súlyok értékeit tapasztalati úton az alábbiakra állítottuk be: w_a=0,5, w_b=0,5, w_O=1, w_P=2,5.

4.1. Eredmények

A kísérleti rendszerrel hasi és kismedencei ultrahangos leletek felismerését vizsgáltuk összesen 20 darab leletre. (Egy lelet nagyságrendileg kb. 10-20 mondatot tartalmaz.) A felismerést azonos körülmények között azonos (rögzített, majd visszajátszott) leletekre először az alaprendszerrel, majd a prozódiai szegmentálóval kibővített rendszerrel végeztük el. Az eredményeket a *2. táblázatban* mutatjuk be 6 darab, a teljes tesztanyag tekintetében reprezentatívan kiválasztott leletre.

A táblázatban megjelenített mérőszámok a helyesen felismert szavak aránya, illetve a szótévesztési arány javulása, mindkettő százalékosan értendő.

A táblázatból látható, hogy a prozódiai szegmentálóval kibővített rendszer teljesítménye összességében 3,82%-kal javult. A javulás mértéke leletenként változó, egyes esetekben 10% fölötti eredményt is kaptunk (lásd pl. 03-as azonosítójú lelet), ugyanakkor előfordul (lásd pl. 16-os azonosító), hogy a felismerés nem javul, hanem éppenséggel romlik a prozódiai információ figyelembe vételekor.

Az egyes leletbemondásokat megvizsgálva arra a következtetésre jutottunk, hogy a prozódiai szempontból jobban – ezzel együtt a „megszokott hétköznapi”, általánosan elvárható kiejtésnél nem gondosabban – bemondott leletek felismerése a prozódiai információ figyelembe vételekor jelentősebb mértékben javul. Azokban az esetekben, amikor a felismerés a kibővített rendszerrel nem javult, hanem romlott, a hibát jellemzően a prozódiai szegmentáló tévesztése okozta, ami a hipotézis gráfok újrasúlyozásakor eltorzította a felismerést. A hiba forrása esetenként a prozódiailag gondatlan beszéd, esetenként az alapfrekvencia-detektor tévesztése volt. Ez utóbbi történhet például egy kissé rekedt hangú beszélőtől származó lelet esetében.

Általános tapasztalatunk, hogy a prozódiai szegmentálás esetenként időben kevésbé olyan pontos, mint ami a szóhatárok helyének biztosabb megállapításához szükséges lenne. Úgy gondoljuk, ez utóbbi probléma legalább részben orvosolható, amennyiben a prozódiai szegmentálóban figyelembe vesszük az adott szót felépítő fonémasorozatot.

Korábban már említettük, hogy a prozódia követésének szempontjából számos szóvégi zöngétlen hang máris hosszának megfelelő, durván 50-100 ms, de akár 150 ms nagyságrendjébe eső bizonytalanságot is eredményezhet a szóhatár megítélésében, mivel az alapfrekvencia menetére ekkor általában nem támaszkodhatunk. Amennyiben lehetőségünk van figyelembe venni az aktuálisan elhangzó beszédhangokat is, időben korrigálni tudjuk a prozódiai alapján előrejelzett és a tényleges szóhatárok eltérését, vagy legalábbis tud-

2. táblázat
A helyesen felismert szavak arányának alakulása az alaprendszer és a kibővített rendszer esetén, illetve a szótévesztési arány javulása.

Lelet-azonosító	Helyesen felismert szavak [%]		A szótévesztési arány változása (relatív) [%]
	Alap rendszer	Kibővített rendszer	
03	71,2	78,9	10,9
07	78,8	80,6	3,6
08	84,6	84,6	0,0
10	70,8	72,2	2,0
16	68,3	66,7	-2,4
19	83,8	90,5	8,1
Összes lelet (20)	**75,99**	**78,89**	**3,82**

juk, hogy az adott helyen mennyiben támaszkodhatunk a prozódiai szegmentálás pontosságára. További kutatásaink során mindenképpen szeretnénk ilyen jellegű vizsgálatokat végezni, ezáltal a rendszert teljesebbé tenni.

Megjegyezzük továbbá, hogy a prozódiai szegmentáló által meg nem jelölt szintaktikai vagy szóhatárok a beszédfelismerést a hipotézis gráfok újrasúlyozási algoritmusából kifolyólag nem rontják, jóllehet érdekünkben áll minél több szóhatárt megtalálni, ezáltal több lehetőséget adva a prozódiai információt nem használó felismerés hatékonyságának növelésére. A prozódiai szegmentálótól nem várhatjuk el, hogy valamennyi szóhatárt megtaláljon (erre még gyakorlott szakember sem vállalkozhat pusztán az alapfrekvencia és az energiaértékek ismeretében), ugyanakkor bebizonyosodott, hogy a megtalált szóhatárok alapján a felismerés hatékonysága javítható.

5. Összegzés

Írásunkban azt vizsgáltuk, hogyan használhatók fel bizonyos szupraszegmentális jellemzők a beszédfelismerés segítésére. Bemutattunk egy automatikus prozódiai szegmentálót, amely az alapfrekvencia és az energiaszint értékei alapján próbálja meg felismerni a dallam 6 kiválasztott alaptípusát, illetve a szünetet. A prozódiai szegmentálót beszédfelismerőbe építve arra használjuk, hogy a dallamtípusok felismerése révén megtaláljuk az egyes szintaktikai egységek közötti határokat, amelyek egyúttal szóhatárokat is jelentenek.

Az elvégzett kísérletek tanúsága szerint a szóhatárok ismerete révén a felismerés hatékonysága a hipotézis gráfok újrasúlyozásával javítható (arról nem is beszélve, hogy egyes írásjelek, így a vessző kitételében is nagy segítséget adhat a szintaktikai határok ismerete). Mindezek mellett a prozódiai szegmentáló akár automatikus szintaktikai elemzőkben is felhasználható lehet.

A szerzőkről

Szaszák György 2002-ben végzett a Budapesti Műszaki és Gazdaságtudományi Egyetem Villamosmérnöki és Informatikai Karán. Ez évtől kezdődően a BME-TMIT Beszédakusztikai Laboratóriumában dolgozik, főbb kutatási területe a gépi beszédfelismerés, ezen belül beszédadatbázisok konstrukciója és feldolgozása, beszédfelismerés rejtett Markov-modellekkel, ejtésvariáció modellezés, szupraszegmentális jellemzők felhasználása a beszédfelismerésben, érzelmek felismerése akusztikai beszédjel alapján. Doktorjelöltként PhD dolgozata megvédésére készül.

Vicsi Klára a BME TMIT Beszédakusztikai Laboratórium vezetője. Beszédfelismerési témában írta meg PhD-jét 1992-ben. A MTA Mérnöki Tudományok Doktora lett 2004-ben, BME habilitációja pedig 2007-ben volt. Számos korábbi hazai és nemzetközi kutatás témavezetője és jelenleg is aktiv projektvezető a beszédakusztika, a gépi beszédfelismerés, beszéd adatbázis készítés és a pszichológiai akusztika területén. Előszeretettel foglalkozik beszédsegítő eszközök létrehozásában nagyothalló és beszédhibás személyek részére. A lektorált hazai és nemzetközi folyóiratokban, nemzetközi konferenciakiadványokban több mint 65 publikációja jelent meg. Szerzője számos beszédkutatással foglalkozó könyvrészletnek.

Irodalom

[1] Gallwitz, F., Niemann, H., Nöth, E., Warnke, V.: Integrated recognition of words and prosodic phrase boundaries. Speech Communication, Vol. 36, 2002, pp.81–95.

[2] Kassai Ilona: Fonetika. Tankönyvkiadó, Budapest, 1998.

[3] Kompe, R.: Prosody in Speech Understanding Systems. LNAI 1307, Springer Verlag, Berlin-Heidelberg, 1997.

[4] Kompe, R., Kiessling, A., Niemann, H., Nöth, H., Schukat-Talamazzini E.G., Zottman, A., Batliner, A.: Prosodic scoring of word hypothesis graphs. In: Proceedings of the European Conference on Speech Communication and Technology, Madrid, 1995, pp.1333–1336..

[5] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A.: Stochastic pronunciation modelling from hand-labelled phonetic corpora. In: Modeling Pronunciation Variation for ASR. Rolduc, 1998, pp.109–116.

[6] Roach, P.S. et al.: BABEL: An Eastern European Multi-language database. Int. Conference on Speech and Language, 1996.

[7] Sjölander, K., Beskow, J.: Wavesurfer – an open source speech tool. Proc. of the 6th International Conference of Spoken Language Processing in Beijing, China, Vol. 4, 2000, pp.464–467.

[8] Szaszák György, Vicsi Klára: Folyamatos beszéd szószintű szegmentálása szupra-szegmentális jegyek alapján. III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2005, pp.360–370.

[9] Young, S. et al.: The HTK Book (for version 3.3). Cambridge: Cambridge University, 2005.

[10] Veilleux, N.M., Ostendorf, M.: Prosody/parse scoring and its application in ATIS. Human Language and Language and Technology. Proc. of the ARPA Workshop, Plainsboro, 1993. pp.335–340.

[11] Vicsi K., Kocsor A., Tóth L., Velkei Sz., Szaszák Gy., Teleki Cs. , Bánhalmi A., Paczolay D.: A Magyar Referencia Beszédadatbázis és alkalmazása orvosi diktálórendszerek kifejlesztéséhez. III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2005, pp.435–438.

[12] Vicsi, K., Szaszák, Gy.: Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features. International Journal of Speech Technology, Vol. 8., No.4, 2005, pp.363–370.