Beszédfelismerő modellépítési kísérletek akusztikai, fonetikai szinten, kórházi leletező beszédfelismerő kifejlesztése céljából

Velkei Szabolcs, Vicsi Klára

BME Távközlési és Médiainformatikai Tanszék, Beszédakusztikai Laboratórium

1117 Budapest, Magyar tudósok krt. 2.

E-mail: vicsi@tmit.bme.hu, velkei@tmit.bme.hu

Kivonat: Cikkünkben a Beszédakusztikai Laboratóriumban kifejlesztett HMM alapú beszédfelismerő rendszert, a rendszer optimalizálását mutatjuk be, és a felismerési eredményeinket összehasonlítjuk a széles körben elterjedt Hidden Markov Model Toolkit (HTK) rendszerrel kapott eredményekkel. A kutatás folyamatos, most az első évben a fonetikai felismerési szintet fejlesztettük ki, optimalizáltuk az akusztikai és a fonetikai szinteket. Az összehasonlító kísérletek azt mutatták, hogy az általunk kifejlesztett beszédfelismerő eljárás akusztikai szintű optimalizálásával valamint az akusztikai–fonetikai modellek optimalizálásával növelni tudtuk a felismerési pontosságot, és gyorsítani tudtuk a feldolgozást.

1 Bevezetés

Munkánk célkitűzése a Beszédakusztikai Laboratóriumban egy középszótáras, általános magyar nyelvű, folyamatos beszédfelismerési technológia kidolgozása, valamint egy ahhoz tartozó nyelvi modell elkészítése, amelynek segítségével a rendszer meghatározott kötött témában, közepes szótárméret alapján működik, rögzített nyelvtani keretek között, kis-zajú környezetben.

A 2004 év elején kezdődött, és 3 évig tartó project keretén belül a Laboratóriumban új megoldásokat dolgozunk ki az akusztikai előfeldolgozásban, a statisztikai modellépítésben valamint fonetikai, fonológiai és morféma nyelvi szinteket vonunk be a felismerési folyamatba. A felismerési kísérletekhez HMM alapú saját fejlesztésű beszédfelismerő rendszert állítottunk össze, a szokásos Hidden Markov Model Toolkit (HTK) [6] rendszert összehasonlításra használtuk. Az első évben a fonetikai felismerési szint feldolgozását optimalizáltuk. Cikkünkben erről az optimalizálási folyamatról és a kifejlesztett MKBP 0.8 felismerőről számolunk be.

2 Akusztikai előfeldolgozási eljárás optimalizálása

Akusztikai előfeldolgozási eljárásra a beszédfelismerés immár több évtizedes fejlődése alatt számos eljárás született, melyek közül ma az alábbiak a legismertebbek [1]: szűrősoros elemzés (BFFP), lineáris predikció analízis (LP), perceptuális lineáris predikció (PLP), kepsztrális együtthatók vizsgálata (MFCC), spektrális torzításon alapuló rendszerek (SDM).

A mai legsikeresebb felismerők előfeldolgozási rendszere mell-kepsztrum (MFCC) vizsgálatot végez: valamilyen hallásmodell alapján (Mel, Bark) [4,8] kiszámított szűrők sorozata szolgáltatja a bemeneti vektort (szűrősoros elemzés):

$$\text{Pl: Bark szűrő: } 10\lg L(x) = 15.8 + 7.5(x+0.5) - \sqrt{17.5(1+(0.5x)^2)} \tag{1}$$

$$\text{Mel-szűrő: } \mathrm{Mel(f)} = 2595\log_{10}\left(1+\frac{f}{700}\right) \text{ háromszögszűrő} \tag{2}$$

(f: frekvencia)

Az információ jobb kinyeréséhez kepsztrális együtthatókat képzünk (MFCC):

$$c_i = \sqrt{\frac{2}{N}\sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j-0.5)\right)} \tag{3}$$

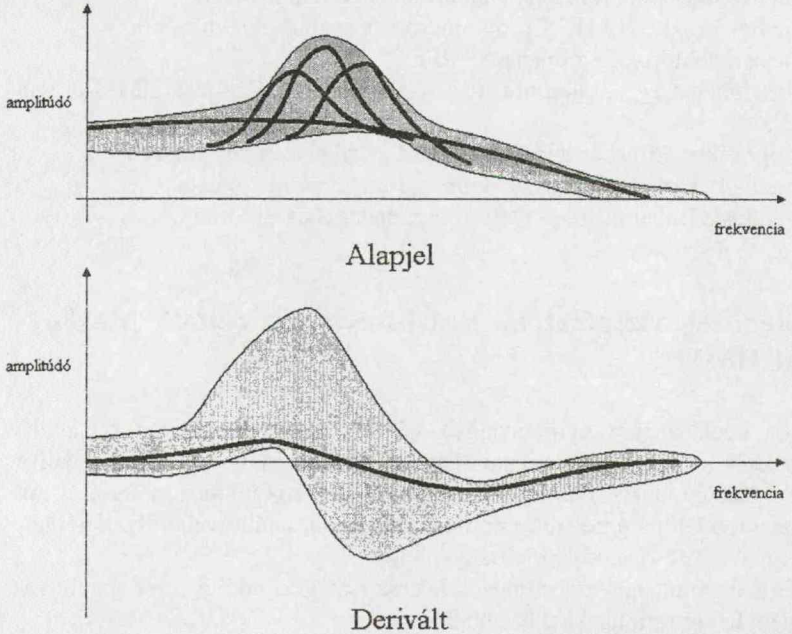
(c: kepsztrális együtthatók, m: szűrők energiái)

A kepsztrális együtthatók információtartalma igen magas, azonban az emberi szem számára nem hordoz jól felismerhető jegyeket. Ennek ellenére (vagy éppen ezért) az MFCC nagy népszerűségnek örvend, az egyik legtöbbet alkalmazott módszer. Ennek okai a következők lehetnek: Kizárólag a Mel- illetve Bark szűrősorokkal végzett akusztikai előfeldolgozáskor, a szűrősorok kimenetei ugyan hasonlóak a különböző ejtésekben, de egymáshoz viszonyítva spektrálisan el vannak tolódva (nagy a varianciájuk), a szűrősorral betanított Markov-modellekkel a betanítástól teljesen eltérő jeleket is lehet ismerni. Ebből kifolyólag a szűrősorral képzett eredmény vizuálisan ugyan jól meghatározott képet ad (emelkedés, süllyedés, maximumok), de mindenképpen kevés az információ. A (3.) képletben közölt módszer csökkenti a varianciát.

Felvetődött a kérdés, hogy miért ne készíthetnénk olyan akusztikai elemzést, melyben az eredeti szűrősoros adatok frekvenciatérbeli deriváltjai kapnak szerepet. A deriváltak a szomszédos szűrők különbségeként nyerhetők ki. Ez az eljárás nem teljesen idegen a szakirodalomban, de ma a gyakorlatban nem használatos. A frekvencia deriváláskor létrejött adatkiemelést ábrázolja a 2. ábra.

A felső szürke kitöltéssel stilizált eloszlás az adatok Bark szűrősoron való áteresztésekor keletkezett. Az eloszlást létrehozó bemeneti adatok egy része zöld színnel ki lett emelve. A képen jól látható, hogy a modell a piros függvényhez tartozó bemeneti vektorokat is nagy valószínűséggel képes felismerni, pedig a betanító minták ettől szignifikánsan különböztek. Az alsó ábrán az eddigi Bark szűrősor frekvenciabeli

deriváltjaiból képzett modell bemeneti vektor-eloszlása látható. Megfigyelhető, hogy mindenképpen szükséges egy minimális egyezés a derivált kilengésénél. Ezáltal a csúcs nem tűnt el, mint az előbbi esetben, és a modell csak olyan eseteket ismer fel,



2. ábra: Bark szűrősor és Bark szűrősor deriváltjainak eloszlása

amikor a vizsgált kritikus helyen valóban van energiamaximum. Továbbá, míg az igen népszerű MFCC számításigénye N^2-el arányos, az új módszeré csak N-el.

Találati százalékos eredmények 20 kHz mintavételezési frekvencia mellett 3 és 5 állapotú, fonéma alapú diszkrét Markov-modellekkel az 1. táblázatban találhatók.

1. Táblázat. Felismerési találatok frekvenciatérbeli és időbeli deriválás esetén.

Találat	Bemeneti vektor	Állapotszám
44.36%	23 Bark szűrő	3 állapot
46.1%	23 Bark szűrő	5 állapot
46.15%	23 Bark szűrő + 23 Bark időbeni derivált	3 állapot
64.38%	23 Bark frekvenciatérbeli derivált	3 állapot
70.16%	23 Bark frekv. derivált + 23 időbeni derivált	3 állapot
72.32%	23 Bark frekv. derivált + 23 időbeni derivált	5 állapot
71.38%	13 MFCC+13 Delta+13 Acc. komponens	5 állapot

A Betanítás Babel magyar nyelvű beszédadatbázissal történt [5].

Referencia-felismerő

A referencia-felismerő ajánlását a COST249 később annak lejárása utána COST278-as munkacsoport dolgozta ki [7] amelynek paraméterei az alábbiak:

- a modellek akusztika fonémák halmaza a megfelelő nyelven,
- a modellek készítése a HTK programcsomag segítségével történik,
- a bemeneti vektorok 39 dimenziós MFCC,
- minden fonéma egy 3 állapotú 'ballról-jobbra haladó' típusú HMM-el van modellezve,
- az ortografikus átírást és a kiejtési szótárt az adatbázis tartalmazza,
- diagonális kovariancia-mátrixú Gauss paraméterek használata.
- A betanítás a Babel magyar nyelvű beszédadatbázissal történt.

3 Modellépítési vizsgálatok, kvázi-folytonos rejtett Markov-modellek (QCHMM)

A 80-as évek végére nyilvánvalóvá vált, hogy diszkrét (vektorkvantált) Markov-modellekkel a felismerés nem javítható tovább, ezért ki kellett dolgozni a folyamatos valószínűségi mezőkre értelmezett modelleket, azok tanítási módszerét. Az alapelv ugyanaz: a modell paramétereire optimális értékeket találni valamilyen iterációs algoritmus segítségével. A módszer hátránya, hogy

- nagybonyolultságú algoritmusok jelennek meg az eddigi négy alapműveletet igénylő algoritmusokkal szemben,
- új probléma merül fel: a gaussi változók szórásainak figyelése és esetleges módosítása,
- a robusztus matematikai módszer sok időbe kerül – lassú programot eredményez.

Nyilvánvalóan ez nem probléma amennyiben lehetőség van nagyteljesítményű vektorszámítógépek használatára, de ennek hiányában kifejlesztettünk egy kvázi-folytonos rejtett Markov-modelleket (QCHMM) használó programot, mely a fenti két vetélytárs jó tulajdonságait igyekszik ötvözni, nevezetesen a nagyobb pontosságot a kisebb futási idővel.

Ezen problémák megoldásához a kvázi-folytonosság a kulcs, mely a következőt jelenti: a meglévő N felbontású diszkrét mezővel nem a tanítóanyag mintáinak eloszlását kell maximális hűséggel visszaadni, hanem azok alapján egy becslést adni a folytonos valószínűségi mezőre. Ehhez egy simítási algoritmusra van szükség, amelyet egyszer vagy többször végigfuttatva az eddig betanított modellen a modell tanítóanyagra való adaptálódását lehet csökkenteni (megszüntetni). A használt algoritmus egy paraméterezhető simítófüggvény (továbbiakban blur), melynek a következőket lehet megszabni:

- az élsimító mátrix (1 dimenzió miatt jelen esetben csak vektor) méretét,
- a mátrix elemeinek értékét, ahol minden sorban összértékben 1-nek kell lennie (a teljes valószínűségi mező mindig 100%)

Megoldandó problémák:

- kvantálási lépcsők számának optimális megválasztása,
- megfelelő súlyozású simítófüggvény megválasztása,
- a tartomány minél jobb eseménytérbeli kihasználtsága.

Az optimális kvantálási együttható, valamint az ehhez tartozó simítófüggvény mérési úton lett kiszámítva. Többféle adatbázisrész felismerési aránya nagyszámú véletlenszerűen választott paraméterezésű felismerő kimenetén össze lett vetve, majd a statisztikailag így megismert kétváltozós probléma maximumkereséssel lett megoldva.

Végeredményben tehát a laboratóriumban kifejlesztett fonémaszintű felismerőnk, a továbbiakban MKBF 0.8, 16 kHz mintavételezésű, 17 Bark frekvenciatérbeli derivált + 17 időbeni derivált + 17 időbeni második derivált + energia bemeneti jelvektor mellett, 4-5 állapotú kvázi-folytonos, 24 lépcsős, rejtett Markov-modellekkel (QCHMM) fonéma, illetve trifon alappal dolgozik.

Kulcsfontosságú lépés az eredmények összehasonlítása más eljárások eredményeivel. Ennek elvégzéséhez a már megismert HTK 3.2 Markov-modellre épülő fejlesztő szoftver nyújtott segítséget, mely egy teljes körű beszédfelismeréshez kialakított alkalmazás. A teszthez a Babel adatbázis anyaga került feldolgozásra a következő módosítással:

- 20 kHz mintavételezés helyett 8 kHz-es újra mintavételezés.

A HTK a szerzők ajánlása szerinti legjobb paraméterezéssel tanulta meg a modelleket:

- előszűrés: $s_n' = s_n - k \cdot s_{n-1}$, ahol $k=0.97$,
- Hamming ablak: $s_n' = \left\{0.54 - 0.46\cos\left(\frac{2\pi(n-1)}{N-1}\right)\right\} s_n$
- 1,2,4,8 illetve 16 Gauss-keverék használata az eloszlás modellezéséhez
- maximum 20 iterációs mélység
- 13db MFCC és ezek első és második időbeli deriváltjai (összesen 39 elem)
- a paraméterek részletesebb megismerése a HTK könyvéből [6] lehetséges.

Az MKBF 0.8 felismerőnk felismerési eredményeinek összehasonlítása a HTK 3.2 alapú referenciafelismerővel a 2. táblázatban találhatók. A táblázat alapján kivehető legfontosabb felfedezés azonban az, hogy a QCHM modellek esetében a felismerési arány gyakorlatilag nem romlott a lépcsők számának drasztikus csökkentésével.

2. táblázat: Az MKBF 0.8 sajátfejlesztésű felismerő felismerési eredményeinek összehasonlítása a HTK 3.2 alapú referenciafelismerővel

HTK 3.2	
Paraméterszám	találati arány
1 Gauss	52.3%
2 Gauss	58.8%
4 Gauss	61.2%
8 Gauss	62.3%

MKBF 0.8	
Paraméterszám	találati arány
6 lépcső/ 0 blur	67.61%
12 lépcső/ 0 blur	64.98%**
24 lépcső/ 4 blur	68.188%
48 lépcső/6 blur	69.55%

32 Gauss használatát a HTK a kevés mintaanyagra hivatkozva elvetette

A laboratóriumban a 80-as években végzett kutatások alapján kimutatták, hogy az emberi fül dinamikában csak meglepően nagy intenzitás-különbségeket (5-6 dB) képes észrevenni. A beszéd intenzitásának tartománya körülbelül 35 dB, így 6-7 lépcsővel jellemezhető az emberi fül 'kvantálása'. A szakirodalomban található SISI-teszt [2], erről a jelenségről tanúskodik.

Tehát a felismerésnél kapott eredmények összhangban vannak a dinamikára vonatkozó szubjektív akusztikai vizsgálatokkal. Az is megfigyelhető, hogy ugyanaz a lépcső/simítás≈7 arány mellett lettek a felismerések a legjobbak. Ez érthető, hiszen a simítás durva közelítéssel megfeleltethető a kvantálási lépcsők csökkentésének is.

Kiértékelési módszerek

Az előforduló hibákat értelmezni kell, erre jó módszer lenne, ha minden esetben meg lehetne mutatni, hogy a felismerő mikor milyen típusú hibát vétett. A hibák típusai a következők lehetnek:

- csere: valamely szimbólum tévesztése egy másikkal,
- beszúrás: két szimbólum közé újabb szimbólum(ok) kerül(nek),
- törlés: valamely szimbólum hiányzik.

Amennyiben a felismerő kimenetén a modellek sorozata található meg, a fenti módszer alkalmazható, ellenben a program a kimeneten adhat egyéb információkat is: például időzítési adatokat. Ebben az esetben lehetőség van utólagos módosításra, mivel a statisztikai alapon gyanús eredményeket szűrni lehet. Ilyen megfontolásból legyen a hiba definíciója a következő:

- Ha egy frame alatt a felismerő kimenetén különböző modellt detektálunk, mint a bemeneti szimbólumhoz tartozó modell, akkor az hibának tekintendő.

Ezzel a módszerrel egy könnyen algoritmizálható hibadefinícióhoz jutottunk, amelynél a hibapontszám továbbra is lineáris kapcsolatban áll a szimbólumhibákkal, de ez esetben hibapont adódik az offszethibákhoz is. A kiértékelésről és a modellek betanításáról részletesebb információkat közöl a szerző TDK dolgozata [3] .

4 Fonetikai modellek

A felismerésben használt fonetikai modellek háromféle típusból lettek összeválogatva:

- a modellek magját a magyar nyelvben használt fonémák rövid-hosszú párjai alkotják, illetve csekély számú fonémánál allofonok lettek megkülönböztetve, (zöngés zöngétlen h, stb.)
- a beszéddetektáláshoz használható beszéd-nem beszéd modell háromelemű: csend, rezonáns és zörej osztályba sorolja a hangokat,
- míg végül a pontosabb felismerés érdekében trifón modelleket alkalmazunk.

Trifón modellek esetén meg lettek különböztetve valódi trifón elemek és osztályozott fonémák. Második esetben ugyanis azokat a fonémákat, amelyekhez a tanító adatbázis nem tartalmazott elegendő információt valamennyi szóba jövő trifón betanításához, ott a kezdő és záró fonémák helyett fonémaosztályokat alkalmaztunk. A trifón modellek használatával a beszéd felismerése előreláthatólag 5-10% fog javulni.

5 Összefoglalás

Az összehasonlító kísérletek azt mutatták, hogy az általunk kifejlesztett beszédfelismerő eljárás (MKBF 0.8) akusztikai szintű optimalizálásával valamint az akusztikai–fonetikai modellek optimalizálásával növelni tudtuk a felismerési pontosságot, és gyorsítani tudtuk a feldolgozást. Természetesen a további szintek igen jelentős mértékben javítani fogják a felismerő pontosságát, de egy jobb kiindulással biztosabb az eredmény.

Köszönetnyilvánítás

A kutatás az OTKA T 046487 ELE és az IKTA 00056 pályázatok keretén belül készül.

Bibliográfia

1. Bechetti, C., and Prina Ricotti, L.: Speech Recognition (John Wiley and Sons LTD 1999.)
2. Pytel, J.: Audológia (Victoria kft., 1996)
3. Velkei, Sz.: Rejtett Markov-modell elméleti és gyakorlati optimalizálása folyamatos beszédfelismeréshez. TDK dolgozat. BME 2004, pp. 29-40
4. Vicsi, K. Matilla, M. and Berényi, P. (1990). Continuous Speech Segmentation Using Different Methods, Acustica, Vol. 71, 152-156. Video Voice, Micro Video, 210 Collingwood, Suite 100. PO Box 7357 Ann Arbor, MI 48107
5. Vicsi, K., A. Vig: Az első magyar nyelvű beszédadatbázis, Beszédkutatás'98, Tanulmányok az elméleti és alkalmazott fonetika köréből. MTA Nyelvtudományi Intézet, Budapest, pp. 163-178, 1998.
6. Young, S. et al.: The THK Book for HTK Version 3.2 (Cambridge University Engineering Department 2001-2002, http://htk.eng.cam.ac.uk/docs/docs.shtml)

7. Zgank, A., Kaịciịc, Z., Diehl, F., Vicsi, K., Szaszak, G., Juhar, J., Lihan, S., 2004. The COST 278 MASPER iniative - crosslingual speech recognition with large telephone
database. Proc. LREC 2004 Lisbon, Portugal.
8. Zwicker, E. and Terhardt, E. 1980. Analytical expressions for band rate and critical bandwidth as a function of frequency, J. Soc. Am. Vol. 68, 1523.