FIATAL MŰSZAKIAK TUDOMÁNYOS ÜLÉSSZAKA XVII.

Kolozsvár, 2012. március 22–23.

BESZÉDFELISMERÉS HATÉKONYSÁGÁNAK VIZSGÁLATA KÜLÖNBÖZŐ NYELVTANOKKAL

CZAP László, PINTÉR Judit Mária

Abstract

There are different levels of speech recognition systems depending on their function and type. One of these is the continuous speech recognition system. The goal in this case is to recognize coherent sentences, which can be carried out by hidden Markov models. Several factors affect its performance. Agglutinative languages – like Hungarian – make it possible to add hundreds of suffixes to a given word. This multiplies the number of word forms contrary to isolating languages – like English – where the main grammatical features determined by prefixes and the sequence of words, and use few suffix words, so the number of word forms is much smaller. Consequently, an English language developed word-based recognizer cannot be adapted to the Hungarian language, and a continuous speech recognizer raises even more problems. During the work we made tests on a trained speaker-dependent hidden Markov model-based speech recognizer. We took into account different grammars for testing, then we analysed the results.

Key words:

speech recognition, grammar, hidden Markov model, HTK toolkit

Összefoglalás

A beszédfelismerőknek feladatától és típusától függően több szintjét különböztetjük meg. Ezek egyike a folyamatos beszédfelismerő. Ebben az esetben a cél: összefüggő mondatok felismerése, amihez megoldást nyújt a rejtett Markov modellek alkalmazása. Fejlődését számos tényező befolyásolja. Agglutináló nyelveknél, mint amilyen a magyar nyelv is, a toldalékoló jelleg szavanként akár több száz féle toldalék hozzáadását teszi lehetővé. Ez megsokszorozza a szóalakokat szemben az izoláló nyelvekkel - mint az angol - ahol a fő nyelvtani funkciókat a szavak sorrendje határozza meg, továbbá alig használ ragozott szavakat emiatt a szóalakok száma nagyságrenddel kisebb. Ebből adódóan egy angol nyelvre kifejlesztett szóalapú felismerő nem adaptálható a magyar nyelvre, folyamatos beszédfelismerőnél pedig még több probléma merül fel. Munkánk során egy betanított beszélőfüggő rejtett Markov modelleken alapuló beszédfelismerőn végeztünk vizsgálatokat, különböző nyelvtani eseteket figyelembe véve, majd elemeztük az eredményeket.

Kulcsszavak:

beszédfelismerés, nyelvtan, rejtett Markov modell, HTK toolkit

1. Bevezetés

A gépi beszédészlelést megvalósító alkalmazások tipikusan a gépi beszédfelismerők, amelyek pusztán beszéd-szöveg átalakítást végeznek anélkül, hogy a beszédben hordozott jelentést megérteni képesek lennének. A beszédfelismerés fejlődése során több módszert is kifejlesztettek ennek megvalósítására.[1] Vizsgálatainkban arra voltunk kíváncsiak, hogy a felismerő nyelvtana hogyan befolyásolja a felismerés hatékonyságát.

A HTK toolkit alkalmas eszköz a beszédfelismerésre. Szerszámkészletnek is tekinthető, amely rejtett Markov modell (HMM) építésére szolgál. A HMM alkalmas bármilyen sorozat modellezésére, és maga a HTK is alapjában véve hasonlóan általános célú. Mindazonáltal a HTK elsősorban mégis rejtett Markov modell alapú beszédfelismerők fejlesztésére szolgál, ami megfelelő nagyságú infrastruktúra háttérrel biztosít ezen feladat számára.

A "rejtett Markov modell" [2] kifejezésben a "rejtett" jelző arra utal, hogy csak a modell működésének az eredménye, a kimenet (azaz a generált szekvencia) jelenik meg a felszínen, a modell maga és a paraméterei belső változók maradnak.

A szótár minden egyes eleméhez tanulással - approximációs eljárással - el kell készíteni egy-egy Markov modellt, majd a felismerés során a kiejtett elemhez ki kell számítani minden modell esetén azt a valószínűséget, amellyel a modell ezt az elemet ilyen kiejtéssel generálhatta. Ha ezek között a valószínűségek között van pontosan egy kiemelkedő, akkor a felismerés valószínűleg sikeres, és a kiemelkedő valószínűséghez tartozó szótári elem lesz az eredmény. (A rejtett Markov modell érzékeny a túltanulásra.) Tehát az ilyen modellekre épülő beszédfelismerés statisztikai alapú.

2. A betanított beszédfelismerő

A fentebb már említett HTK szoftvercsomag alkalmazásával egy 747 hangfájllal, azaz 7470 szóval betanított folyamatos beszélőfüggő (a hanganyagok egy személytől származnak) beszédfelismerőnek a hatékonyságát vizsgáltuk egy 225 hangfájlból, azaz 2250 szóból álló tesztanyaggal. Akusztikus lényegkiemelésre az MFCC (Mel-Frequency Cepstral Coefficients) jellemzőket használtuk. A felismerő alapegységének – a kísérleti eredményeknek megfelelően [3] - a diádot választottuk. A diád olyan fonémakapcsolat, ami két hangból tevődik össze, és az első hang felétől a második hang feléig tart. Így mindkét végükön illeszthetők a szomszédos hangpárokhoz, a kontextusfüggés jól meghatározható. Megjegyzendő, hogy a beszélő az adatbázis létrehozásakor törekedett a gondos artikulációra kerülve a modoros kiejtést.[1]

A beszédfelismerőn megelőzve a nyelvtani vizsgálatokat elvégeztünk néhány tesztelést, hogy növeljük annak hatékonyságát. A tesztelések alapján az alábbi optimalizált beállításokat választottuk az alkalmazott beszédfelismerőnél [4]:

- optimalizálás alapján 8 állapotú HMM modellt alkalmaztuk;
- kettős Gauss eloszlásúak a HMM modellek, a súlyok minden egyes állapot esetén 0,1 és 0,9 eloszlásúak;
- *HVite* modul *p* paraméterének értéke: -30 (a beszúrások miatti büntetés mértéke);
- Hosszú felpattanó zárhangok átírása (pl.: *bb* -> *b*, *tt* -> *t* stb.);
- Ezeket a javításokat átvezetve a hosszú felpattanó zárhangok átírása utáni állapotra, a beszédfelismerés kiindulási eredményeihez jutunk. [4]

$$\text{helyesség} = (\text{helyesen felismert alakzatok} / \text{összes tesztelő alakzatok}) * 100\% \quad (1)$$

$$\text{pontosság} = \{(\text{helyesen felismert alakzatok} - \text{téves beszúrások}) / \text{összes tesztelő alakzat}\} * 100\% \quad (2)$$

A beállítások alkalmazása után a kiindulási helyességi (1) és pontossági (2) értékek az 1. táblázat tartalmazza.

1. táblázat A kiindulási helyességi és pontossági értékek

helyesség	pontosság
85,82%	84,71%

3. A vizsgált nyelvtani esetek

A különböző esetekben az adott nyelvtani szabálynak megfelelően átalakításokat végeztünk a tanító anyagokban, az érintett diádokat átírtuk a kiejtésnek megfelelően (pl. *éljen* -> *éjjen* szóban előforduló *lj* hangpár esetén teljes hasonulás jön létre ezért *jj*-t ejtünk). Az egyes teszteléseket az alábbi nyelvtani szabályok alkalmazásával végeztük el [1]:

- **zöngétlenedés:** a zöngés mássalhangzó helyett a megfelelő zöngétlent ejtjük a rá következő zöngétlen hatására (pl. *megbízta* *zt*->*szt*, *dobta* *bt*->*pt*);
- **zöngésedés:** zöngétlen mássalhangzó helyett a megfelelő zöngés ejtjük a rá következő zöngés hatására (pl. *vasgolyó* *sg*->*zsg*, *juhászbojtár* *szb*-> *zb*);
- **teljes hasonulás** (pl. *utca* *tc*->*cc*);
- **egyedi besorolást igénylő esetek:** Összességében 10 átalakításról van szó amit korábban a zöngésségének megfelelően az első és a második vizsgálatban is alkalmaztunk. Ezek maguk a képzés helye szerinti részleges hasonulások is, és nem hagyhatjuk figyelmen kívül a hagyomány érvényesülését sem, ami gyakran ellentmond a nyelvtani szabályoknak. Például a *ch* és *cz* hangot némileg módosult *h* illetve *c* hangnak ejtjük, de régi családneveknél, mint a *Zichy* vagy *Madách* a *ch* helyett *cs*-t ejtünk;
- **az összes nyelvtani eset együttes figyelembevétele.**

1. ábra Az egyes nyelvtani esetek alkalmazásánál mért százalékos javulások a kiindulási esethez viszonyítva

Az 1. ábrán jól látható, hogy az egyes esetekben hány százalékos javulást lehet elérni a nyelvtani átalakításokkal. Összességében a helyességnél 4,83 százalékos, a pontosságnál 5,18 százalékos javulást eredményezett a nyelvtani szabályok módosítása.

4. Összefoglalás

Minden esetben helyesség és pontosság javulás volt megfigyelhető. Az eredmények igazolják a nyelvtani beállítások megválasztásának fontosságát, hiszen a számítási idő növelése nélkül javítható a felismerés hatékonysága. Megjegyzendő, hogy bár a változtatások javulást eredményeztek, azok visszaalakítása már nem minden esetben egyértelmű, például *lábtól* -> *láptól* [1].

Az egyes nyelvtani jelenségeknél mutatkozó eredmények nem csak az adott szabály alkalmazásának jelentőségétől függenek, hanem a mintaszavakban előforduló esetek számától is. A gyakrabban előforduló példák nagyobb változást generálnak.

A hosszú felpattanó zárhangok átírását nem határozza meg nyelvtani szabály. A hosszú és rövid felpattanó zárhangok a zárfelpattanás előtti szünet (zöngétlenek), illetve fojtott zöngé (zöngések) időtartamában térnek el, ilyenformán az azonosan képzett rövid és hosszú magánhangzók mintájára is lehetne kezelni őket. A nyelvtani szabályok változtatása előtt elvégzett vizsgálat [4] alátámasztja, hogy átírásuk növeli a hatékonyságot (helyességet 0,02%-al, pontosságot 3,33%-al).

5. Köszönetnyilvánítás

A bemutatott kutató munka a TÁMOP-4.2.1.B-10/2/KONV-2010-0001 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Irodalomjegyzék

[1] Pintér Judit Mária: Beszédfelismerés hatékonyságának vizsgálata különböző nyelvtanokkal, TDK dolgozat, 2011, 4;16;18; 29-37. oldal

[2] http://en.wikipedia.org/wiki/Hidden_Markov_model, utoljára látogatva: 2012.január 28.

[3] Dr. Czap László: Audiovizuális beszédfelismerés és beszédsszintézis, PhD értekezés, 2004, 82. oldal

[4] Pintér Judit Mária: Akusztikus beszédfelismerés rejtett Markov- modellek alkalmazásával, Szakdolgozat, 2010, 40-53. oldal

Pintér Judit Mária, egyetemi hallgató

Munkahely: Miskolci Egyetem

Cím: 3515 Miskolc, Egyetemváros

Telefon: +36 (46) 565-111

E-mail: pinter.judit.maria@gmail.com

Dr. Czap László, egyetemi docens, dékánhelyettes

Munkahely: Miskolci Egyetem

Cím: 3515 Miskolc, Egyetemváros

Telefon: +36 (46) 565-111

E-mail: czap@mazsola.iit.uni-miskolc.hu