Az alacsony szintű beszédfelismerés mesterséges feljavítása magasabb szintű modellellenőrzéshez

Németh András[1,2], Balázs László[1], Gyepesi György[1]

[1] Alkalmazott Logikai Laboratórium,
Hankóczy J. u. 7. 1022 Budapest,
{xandrew, bazsi, ggyepesi}@all.hu
[2] Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi és Információelméleti Tanszék,
Magyar tudósok körútja 2. 1117 Budapest,
xandrew@cs.bme.hu

Kivonat: Jelen cikkben ismertetünk egy módszert, mellyel a HMM+GM felépítésű klasszikus beszédfelismerő rendszer teljesítményét feljavítjuk a rendelkezésre álló szöveges átirat segítségével. A feljavítás mértéke az eredeti rendszer teljesítményétől a szinte tökéletes fonéma felismerés szintjéig folytonosan változtatható. Így mérhetővé válik, hogy a különböző, a fonéma felismerő rétegre épített magasabb szintű modellek milyen alacsony szintű hiba esetén milyen teljesítményt nyújtanak. Az itt ismertetett kutatásban a hibatűrő szókeresés és a fonéma n-gramm modellek viselkedését vizsgáltuk meg.

1 Bevezetés

A beszédfelismerő rendszerek jellegzetes felépítési módja, hogy az emberi beszédet több, egymásra épített statisztikai modellel jellemzi, majd az így kapott teljes beszédmodell segítségével történik a felismerés. A legalsó szintet a vizsgált nyelv fonémáinak akusztikus modelljei alkotják. Általánosan elfogadott a beszédfelismeréssel foglalkozók körében, hogy tisztán fonéma szinten jó minőségű beszédfelismerés nem készíthető (erre különben az ember sem képes). Ezért a fonéma szint fölé a konkrét feladattól függően különböző magasabb szintű nyelvi modelleket, pl. fonéma n-gramm statisztikákat, szótárakat, nyelvtanokat illesztenek.

A különböző megoldások hatékonyságát a teljes összetett modell hatékonyságával szokás mérni. Arra ugyan van lehetőség, hogy csak az alacsonyabb szintű modellek használatával végezzünk méréseket, de a magasabb szintű modellek önálló értékelése a klasszikus validációs módszerekkel nem megoldott.

Jelen cikkben ismertetünk egy módszert, melyben a magasabb szintű modellek értékelése az alacsony szint teljesítményének függvényében történik. Így a különböző, magasabb szintű modellekre vonatkozó kísérletek eredményei összehasonlíthatóvá válnak akkor is, ha más fonéma szintű modellel dolgoznak. A beszédfelismerési technológiáknál minden esetben nagyon fontos a felhasználási területhez igazítás, így

a legalkalmasabb magas szintű modell is alkalmazásfüggő. Ha a különböző magas szintű modellekhez a fenti típusú értékelő függvények a rendelkezésünkre állnak, akkor megalapozottabban választhatjuk ki az adott körülmények között (az adott helyzetben elérhető fonéma felismerési hiba ismeretében) a legalkalmasabb magasabb szintű modelleket.

2 A HMM+GM beszédfelismerő architektúra és feljavítása

A felismerés bemenetét egy feature vektor sorozat adja, melynek minden eleme a beszédjel egy kis szakaszának valamilyen akusztikai jellemzőit tartalmazza. A fonémák akusztikai modellje a legtöbb esetben egy Hidden Markov modell, melynek állapotaihoz a feature vektorok valamilyen folytonos eloszlását rendeljük. Leggyakrabban Gauss-mixture típusú eloszlásokat használunk.

A rossz minőségű modellek esetében is nagyon jól működő Viterbi align eljárással minden egyes feature vektorra megmondható, hogy ott a modell mely állapotban van a legnagyobb valószínűségű, az ismert elhangzó szövegnek megfelelő trellis mentén.

Generáljunk egy új feature vektor sorozatot úgy, hogy minden pozícióban az align szerint ott található állapot eloszlásából „húzunk” egy vektort, azaz véletlenszerűen válasszunk az adott eloszlásnak megfelelően. Ezzel lényegében azt érjük el, hogy a kapott feature vektor olyan lesz, mintha a modellünknek tökéletesen megfelelő beszédből nyertük volna ki. Ha ezt a feature file-t adjuk a felismerő bemenetére, akkor érthető módon nagyon jó, méréseink szerint kevesebb mint 5%-os fonéma hibájú felismerést tapasztalunk.

Ha az így kapott feljavított feature file és az eredeti feature file különböző együtthatós, egy összegű lineáris kombinációját tesszük a felismerő bemenetére, akkor a hiba az együtthatók függvényében folytonosan változtatható a kiindulási hibaarány és a szinte tökéletes felismerés között. Az így kapott állítható minőségű alacsony szintű felismerésre ráépítve a vizsgálni kívánt magasabb szintű modellt megkapjuk a bevezetésben előrebocsátott tulajdonságú értékelő függvényt.

3 Hibatűrő keresés és az N-gramm modellek teljesítménye a fonéma hiba függvényében

A hibatűrő keresés a keresett szó és a felismert fonémasorozat edit distance távolsága ill. a felismert sorozatban előforduló fonéma-trigrammok halmazának és a keresett szó trigramm halmazának összehasonlításával történik ([2]). A két módszer nagyon hasonló eredményt ad, és mindkettő esetében a teljesítmény közel lineárisan függ a fonéma hibától.

Az n-gramm modellek esetében egy nagy nyelvi korpuszból (pl. [1]) kinyert fonéma trigramm előfordulási valószínűségek segítségével irányítjuk a felismerést a nyelvre jellemző hangsorozatok preferálása felé. (Lásd pl. [3]).

Ebben az összefoglalóban példaként megadjuk a keresési teljesítmény függését a fonéma hibától. A teljes cikk mindkét modell részletes leírását és értékelését tartalmazni fogja.

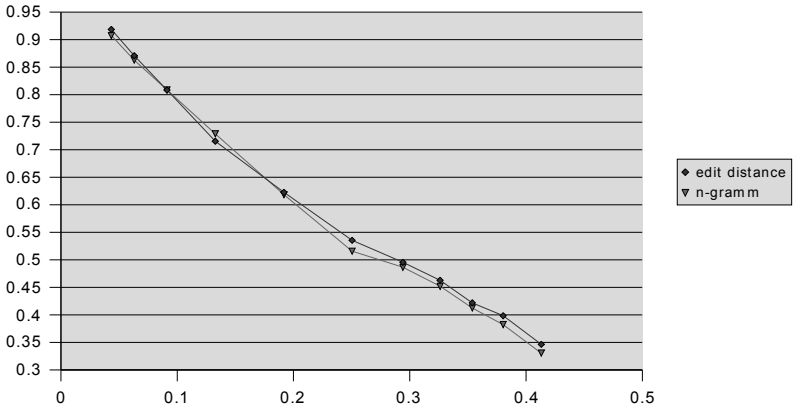

Fig. 1. Az edit distance alapú és a trigramm alapú keresés teljesítménye f-measure-ben a fonéma felismerési hiba függvényében

Bibliográfia

1. P. Halácsy, A. Kornai, L. Németh, A. Rung, I. Szakadát, V. Trón: Creating open language resources for Hungarian. In Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004), 2004
2. K. Ng: Subword-based Approaches for Spoken Document Retrieval Ph.D. Thesis,MIT, 2000
3. C. D. Manning, H. Schtze (ed.): Foundations of Statistical Natural Language Processing. MIT Press, 1999