Gépi beszédfelismerők betanítása – Mennyi kézi szegmentálásra van szükségünk?

Mihajlik Péter, Tatai Péter, Gordos Géza
BME,Távközlési és Médiainformatikai Tanszék
mihajlik@tmit.bme.hu

Kulcsszavak: automatikus beszédfelismerés, beszélőfüggetlen telefonos felismerő tanítás, beszédadatbázis, kézi – gépi szegmentálás, fonetikus átírás

A mai beszélőfüggetlen beszédfelismerők betanításához nagy mennyiségű beszédadatra van szükség. Mivel a felismerés elemi egységei tipikusan a beszédhangok, feladatunk ezen szegmentumok és pozíciójuk meghatározása a tanítóanyagban, hogy a megfelelő hangrészletekkel a megfelelő beszédhang-modellek betaníthatók legyenek. A tanító-adatbázis hangokra szegmentálása történhet kézi vagy automatikus módszerekkel, illetve ezek kombinációival. Míg a világtrend szerinti "főáramlatban" szinte csak (implicit) gépi szegmentálást használnak, néha nagy mennyiségű kézi szegmentálást tartalmazó beszédadatbázisok is kifejlesztésre kerülnek, mint pl. a nemrégiben elkészült MTBA (Magyarnyelvű Telefon-Beszéd Adatbázis)[1].

Fontosnak éreztük eldönteni, hogy a felismerés pontosságát javítandó, melyik utat válasszuk; ezért az említett adatbázison kutatásokba kezdtünk. Először mind az 500 beszélő kézi szegmentálását felhasználva tanítottuk a rendszert, környezetfüggő beszédhangmodelleket alkalmazva. Így 1000-es szótárméretű, izoláltszavas felismerési feladat esetén, független teszt adatbázison 6.85%-ra tudtuk szorítani a felismerési hibát. Ezután megpróbáltuk mindössze 10 beszélő kézi szegmentálásának felhasználásával betanítani a felismerőt. Összetett tanítási módszerünk a továbbiakban már csak az (500 beszélős) adatbázis annotált szövegeit és hullámformákat igényelte. Kiejtési változatokat is tartalmazó fonetikus átiratokat fonetikai szabályok alapján automatikusan állítottuk elő[2]; ezek segítségével a "kényszerített felismerés" módszerét alkalmazva automatizáltuk a lehallgatást és a szegmentálást is. A végeredményül kapott felismerési hiba mindössze 7.02% lett.

Összegzésként megállapíthatjuk, hogy mélyebb nyelvi információkat is felhasználó tanítási módszerünket alkalmazva gyakorlatilag ugyanannyi a felismerési hiba, mint a kizárólag kézi szegmentáláson alapuló esetben. Ugyanakkor a szükséges kézi munka mennyiségét az eredeti *50-ed részére* csökkentettük.

[1] http://alpha.ttt.bme.hu/speech/MTBAhun.htm

[2] Mihajlik, P., Révész, T. and Tatai, P., Phonetic transcription in automatic speech recognition, Acta Linguistica Hungarica, Vol. 49, pp. 407–425, 2002