FIATAL MŰSZAKIAK TUDOMÁNYOS ÜLÉSSZAKA XVII.

Kolozsvár, 2012. március 22–23.

GÉPI BESZÉDFELISMERÉS

VARGA Attila K.

Abstract

Speech recognition is a method that can translate spoken words into text. In other words, recognition of speech is a multileveled pattern recognition task [1], in which acoustical signals are examined and structured into a hierarchy of sub-words, words, phrases, and sentences. The speech signal consists of the uttered digit along with a pause period and background noise. Pre-processing reduces the amount of processing required in later stages. Generally, pre-processing involves taking the speech samples as input, blocking the samples into frames, and returning a unique pattern for each sample, as described in the following steps.

A speech-to-text system converts speech into normal language text. Modern systems are capable of understanding continuous speech. It is much easier for the program to understand words when we speak them separately, with a distinct pause between each one. Commercial systems for speech recognition have been available off-the-shelf since the 1990s. The paper focuses on the basis of speech recognition.

Key words:

Speech-To-Text, Speech Recognition, Acoustical Signals, Speech Samples

Összefoglalás

A beszédfelismerés az a folyamat, mely során a beszédfelismerő gép azonosítja a kiejtett beszédjeleket és átalakítja ezeket szöveggé, vagy más, számítógép által feldolgozható adattá. Ezzel hatékony kommunikáció alakítható ki a számítógéppel.

A beszédfelismerő alkalmazások segítségével a számítógépek képesek értelmezni az emberi beszédet. A beszédfelismerés és a hangfelismerés fogalmát sokszor összekeverik, pedig a beszédfelismerés a beszélt nyelv értelmezését jelenti, míg a hangfelismerés a beszélő személy azonosítását, annak hangja alapján.

Az emberekhez hasonlóan a gépi felismerőnek is szüksége van tanulásra. Mind a nyelvi, mind az akusztikus információt valamilyen formában előre be kell vinni a rendszerbe. Ha egy nyelv szókészletének egy részével és hangjainak paramétereivel és kiejtési szabályaival betanítunk egy gépi felismerőt, akkor lehet esély arra, hogy önálló szavakat vagy hosszabb kifejezéseket gépi úton felismertessünk.

Kulcsszavak:

Speech-To-Text, akusztikus hullámok, gépi beszédfelismerés, akusztikus jelek, beszédminták

1. Bevezetés

Az emberi beszéd a beszédhangok kibocsátásán alapszik (akusztikus hullámok). A gépi beszédfelismerés szintjén a beszédjelek fonémák elektronikus képe. Ha pontosabbak akarunk lenni, akkor igazából a beszéd nem csupán fonémák sorozata, hanem fontos a kontextus is, azaz a hangsúlyozás, hanglejtés, stb.

A beszélők közötti legfőbb különbség, hogy mindenki másképp ejti ki a hangokat, sőt ugyanaz a személy sem képes kétszer pontosan ugyanazt a hangot produkálni. Ezen felül, a beszédhibás emberek esetén további problémák is adódhatnak a gépi beszédfelismerés során. Ez a lényegkiemeléssel javítható, mely kapcsán lehetséges eljárás például a lineáris predikció, vagy a Fourier transzformáció. Zajos környezet

esetén zajkiszűrő algoritmusok használatával javítható a felismerés, illetve több hangrögzítő eszköz egyidejű alkalmazása. [2]

2. A gépi beszédfelismerés folyamata

A hatékony beszédfelismerés kritikus része minden hangutasítással működő rendszernek. A hangfelismerés legfontosabb mérőszáma a felismerés pontossága, azaz pontosan lett-e megértve az elhangzott hangüzenet. Egyszerűbb megérteni valaki beszédét, ha a hallgató ismeri azt, aki beszél hozzá és már hozzászólt annak beszéd stílusához, főleg akkor, ha a beszélőnek egyedi kiejtése vagy erős akcentusa van. Ez a megállapítás különösen igaz a számítógép alapú hangfelismerésnél, mely révén a hangfelismerés pontossága fokozható. [3]

A szavakat, amelyeket a beszédfelismerő rendszernek le kell tudnia fordítani, szótárba kell foglalni. Azoknál a beszédfelismerő rendszereknél, amelyek számítógépes szöveg bevitelre képesek, szintén több ezer szavas szótár kell, hogy rendelkezésre álljon. A beszédfelismerő rendszer megoldások másik véglete azon rendszerek, amelyeket arra terveztek, hogy a felhasználó például a feltett kérdésekre igennel vagy nemmel válaszoljon. Ezek a kis szótáros rendszerek.

A beszédfelismerő rendszerek sok problémával szembesülnek az emberi nyelv és a kommunikáció komplexitása miatt, melyek befolyásolják a program teljesítményét. A fejlettebb beszédfelismerők ezeket a problémákat már képesek kezelni.

1. ábra. A beszédfelismerés folyamata

A természetes emberi kommunikáció előnye az automatikus gépi feldolgozással ellentétben, hogy a kommunikációban résztvevő személyek rendelkeznek háttértudásról (témáról, másodlagos jelentésekről, a testbeszédről, stb.). A gépi beszédfelismerés az alábbi fázisokra bontható:

1. Hangfelvétel

Beszédjel továbbítása a beszédfelismerő rendszernek.

2. Előzetes szűrés vagy lényegkiemelés

A beszédjelből megkísérli meghatározni a beszéd tartalmát hordozó mennyiségeket, és kiküszöbölni a felismerés szempontjából érdektelen információkat (zaj, fázis, torzítások). Törekvés az invariáns elemek kiemelésére. A kimenet gyakorlatilag egy adott dimenziójú lényegvektor-sorozat.

3. Keretezés

Ablak mérete ms-ban (10-30 ms), 50%-os fedésben. A lényegvektor-sorozat feldarabolása.

4. További szűrés

Nem minden esetben jelenik meg, van, amikor az előzetes szűrés elégséges. Sokszor itt megy végbe az időillesztés és a normalizálás.

5. *Összehasonlítás és illesztés (osztályozás)*

Felismerni a kijelentést. Ez legtöbbször valamilyen mintával való összehasonlítást jelent (például két jellegvektor összehasonlítása). Minden alkalmazott módszer egy valószínűségi és pontossági illesztést végez.

6. *Kimenet (pl. írott szöveg), feldolgozás.*

3. SpeechToText alkalmazások

Már találkozhatunk olyan mobil alkalmazással, ami képes hang alapú keresésre, figyelembe véve akár az éppen aktuális tartózkodási helyünket is. A Google Voice keretében megtaláljuk a voicemail transcription szolgáltatást is, ami automatikusan leíratot készít hangüzeneteinkből; a YouTube-on pedig kérhetünk már angol nyelvű gépi felíratot, amit Google Translate-tel le is fordíthatunk más nyelvekre. A felíratozó ugyanazt az algoritmust használja, mint a Google Voice szolgáltatás, ahol a felhasználónak küldött hangüzenetből a rendszer szöveget készít, s azt elküldi az illető Gmail fiókjába. Ugyanez a YouTube esetében annyit tesz, hogy a feltöltött videók alatt lévő hangsávban található beszédet elemezve, szöveges formátumot állít elő, melyet a videó alá időzít. Mind a felíratozás, mind az időzítés teljesen automatikus, bár a felismerő pontatlan lehet, ezért a fejlesztők azt javasolják, hogy a művelet után a manuális ellenőrzés se maradjon ki.

A másik legismertebb a Nuance, mely már szinte mindenkit felvásárolt a piacon, aki valamit ezen a területen alkotott, övé lett a Philips Speechmagic-je, a Spinvox és már a hang-biometriával foglalkozó Persay is. Legismertebb terméke a Dragon diktáló termékcsalád, ami a világnyelveken nagy pontossággal képes valós időben leíratot készíteni PC-n vagy iPhone-on.

Hasznos kis handsfree assistant szolgáltatást indított a Dial2do is, ami beszédfelismerő és szintetizáló technológiákat felhasználva teszi könnyebbé a kommunikációt, amikor kocsiban.

A webes szolgáltatásokon és mobil alkalmazásokon túl is van élet a technológia számára, hiszen ha nem is tökéletesek még, de már most is vannak intelligens Call-Centerek. Az autókban ülve hanggal vezérelhetjük a navigációs eszközünket, és az USA-ban a sportesemények élő közvetítésének valós idejű felíratozása sem fantázia már. Az automatikus tolmácsolásra viszont még várni kell néhány évet.

Az Apple 2011. nyarán kiadta a fejlesztőknek szánt iOS 5 béta ötödik verzióját, amelybe bekerült egy olyan fejlesztés is, amellyel gépelés helyett a hangunk válik szövegbeviteli eszközzé.

4. Összefoglalás

Kötetlen, folyamatos beszéd felismeréséhez vagy a nagy háttérzajban történő felismeréshez szükséges a nyelvi és tartalmi elemzés is, mint ahogy az ember is csak azt ismeri fel biztonságosan, amit megért. Mobil számítógépes és a rádiós rendszereknél a hangfelismerés hatékonyságának és az adatfeldolgozás sebességének növelésre nagy hangsúlyt fektetnek.

Bár jelenleg nincs olyan beszédfelismerő rendszer, ami univerzálisan mindenre használható, de az adott feladathoz igazítva már biztatóak az eredmények magyar nyelven is. Minél jobban ismerjük a felismerendő hanganyag akusztikai és tartalmi környezetét, annál magasabb pontosság érhető el az adott helyzetre történő tanítással, adaptálással, például néhány száz parancsszó felismerése már 99%-os pontossággal lehetséges (pl.: mobilon hanghívás funkció), a személyre tanított diktáló rendszerek is 90% fölött teljesítenek és a nagy szótáras általánosabb rendszerek sem elégszenek meg a 70%-nál gyengébb eredménnyel.

A hazai kutatások eredményeit látva, nem tűnik reménytelennek, hogy a fenti nemzetközi példákhoz hasonló magyar nyelven értő egy-két szolgáltatás béta verziója még az idén megjelenjen a hazai közösség számára is.

Köszönetnyilvánítás

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

Irodalom

- [1] Hu, M. K. *Visual pattern recognition by moment invariants*. IRE Transactions on Information Theory, Vol. 8. pp. 179-187.
- [2] Allen J. B. *How do human process and recognize speech?* IEEE Trans. Speech Audio Processing, vol. 2., no 4., pp. 567-577. In Workshop on Robust Methods for Speech Recognition in Adverse Conditions, 1994.
- [3] Dupont S., Luetin J. *Audio-Visual Speech Modeling for Continuous Speech Recognition*. IEEE Transactions on Multimedia, Vol. 2, No. 3., 2000.

Varga Attila K.

Miskolci Egyetem, Gépészmérnöki és Informatikai Kar,
Automatizálási és Kommunikáció-technológiai Tanszék
3515 Miskolc-Egyetemváros
Telefon: +36 46 565 140
E-mail: varga.attila@uni-miskolc.hu