Multidiszciplináris tudományok, 10. kötet. (2020) 4 sz. pp. 66-79 **https://doi.org/10.35925/j.multi.2020.4.10**

BESZÉDFELISMERŐ HATÉKONYSÁGI VIZSGÁLATA KÜLÖNBÖZŐ ÁLLAPOTSZÁMÚ REJTETT MARKOV-MODELLEKKEL

Pintér Judit Mária
tudományos főmunkatárs, Automatizálási és Infokommunikációs Intézet
3515 Miskolc, Miskolc-Egyetemváros, e-mail: pinterjm@uni-miskolc.hu

Absztrakt

A beszédfelismerés az a folyamat, melynek során a beszédfelismerő gép azonosítja a kiejtett beszédjeleket és átalakítja ezeket szöveggé, vagy más, számítógép által feldolgozható adattá. Beszédjelek alatt természetesen érthetünk akusztikus vagy akár vizuális jeleket is (gesztikulációk, arcmimika, szájmozgás). Az általam betanított beszédfelismerő viszont az akusztikus jeleket fogja figyelembe venni, azaz magát a beszédhangot, amit szöveggé fog átalakítani. Vizsgálataim a lényegkiemelés alapját képező rejtett Markov-modellek állapotszámának optimális megválasztását helyezik középpontba.

Kulcsszavak: *lényegkiemelés, állapotszám, rejtett Markov-modell, hatékonyság vizsgálat, beszédfelismerés*

Abstract

Speech recognition is the process with which a speech recognition machine identifies the pronounced speech signals and converts them to text or other computer-processable data. By speech signals, of course, we can also mean acoustic or even visual signals (gestures, facial expressions, mouth movements). The speech recognizer I have taught, on the other hand, will take into account the acoustic signals, i.e. the speech itself, which it will convert to text. My research focuses on choosing the optimal number of states for the hidden Markov - models that form the basis for highlighting the essentials.

Keywords: *speech enhancement, state number, hidden Markov - model, efficiency study, speech recognition*

1. Bevezetés

A beszédfelismerés meglehetősen tág témakör. Szűkebb értelemben a tartalom felismerését értjük alatta, tágabb értelemben azonban alkalmazások egész sora használ egészében vagy komponenseként beszédfelismerőt. A beszéd használata az ember-gép kapcsolatban, azaz amikor gépekkel beszélünk, része annak a törekvésnek, hogy a számítógépeket (és más információs rendszereket) többféle bemeneti és kimeneti eszközön tudjuk elérni, úgynevezett multimodális interakcióval. [1]

A tervezett felismerő elkészítésének elsődleges célja a navigáció megkönnyítése és hatékonyabbá tétele volt ipari eszközök esetén. Bizonyos funkciók előhívása, vagy több ugyanabból a lépésekből álló utasítások, kiértékelési folyamatok végrehajtása kiváltható lenne egészen rövid kifejezéseket tartalmazó szóbeli paranccsal. Figyelembe kell venni azt is, hogy a felületet nem csak egy adott személy fogja nagy valószínűséggel használni, ezért a szóbeli navigációnak ugyanolyan jó hatékonysággal kell majd működnie, minden egyes ember esetén. A navigációt megvalósító beszédfelismerő egy kötött szótáras; mintafelismerő; kapcsolt szavak felismerésére alkalmas; beszélőfüggetlen tulajdonságokkal bíró modul.

2. Rejtett Markov-modell

Az a fogalom, hogy valami Markov-tulajdonságú azt jelenti röviden, hogy adott jelenbeli állapot mellett, a rendszer jövőbeni állapota nem függ a múltbeliektől. Másképpen megfogalmazva, ez azt is jelenti, hogy a jelen leírása teljesen magába foglalja az összes olyan információt, ami befolyásolhatja a jövőbeli helyzetét a folyamatnak. [2] E modelleket a tudomány számos más területén – fizika, statisztikai folyamatok, internet, matematika, biológiai modellezés, gazdasági elemzések, szerencsejátékok - is alkalmazzák. A Markov-modellek bonyolultabbak a döntési fa modelleknél, de lényegesen kevesebb programozói ismeretet és kisebb adatmennyiséget igényelnek, mint a szimulációs modellek. A Markov-modell magában foglalja a döntési fa lényeges tulajdonságait, és ezen felül már az események bekövetkezésének idejét is figyelembe tudja venni.

A "rejtett Markov-modell" [3,4] kifejezésben a "rejtett" jelző arra utal, hogy mi csak a modell működésének az eredményét, a kimenetet (azaz a generált szekvenciát) ismerhetjük, a modell maga és a paraméterei számunkra ismeretlenek. Így mi csak a kimenetből következtethetünk a modell felépítésére és a működését leíró paraméterekre (az átmeneti és a kibocsátási valószínűségekre).

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, amely valószínűséggel a modell ezt az elemet ilyen kiejtéssel generálhatta. Ha ezek között a valószínűségek között van pontosan egy kiemelkedő, akkor a felismerés sikeres, és a kiemelkedő valószínűséghez tartozó szótári elem lesz az eredmény. (A rejtett Markov-modell érzékeny a túltanulásra.) Tehát az ilyen modellekre épülő beszédfelismerés tisztán statisztikai alapú. A HMM előnye, hogy elég egyszerűen kiterjeszthető nagyszótáras, folyamatos beszéd felismerésére.

Egy beszédfelismerési feladat rejtett Markov-modellekkel matematikailag az alábbiak szerint fogalmazható meg:

$$\text{szófelismert}=\text{argmax}_{\text{minden szóra}}\{P(\text{szó}|X)\} \quad (1)$$

Vagyis azt a szót (vagy más beszédelemet) keressük, amelyre az X adott akusztikai megfigyelés-sorozat valószínűsége a legnagyobb. Számunkra azonban az X megfigyelés-sorozat ismert, ezért Bayes tétele alapján átalakítva a fenti összefüggést az alábbiak szerint írhatjuk:

Bayes- tétel:

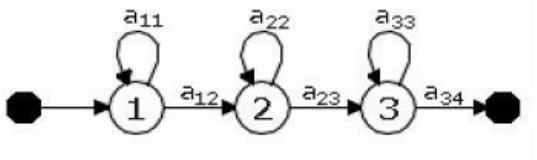
$$\text{szófelismert}=\text{argmax}_{\text{minden szóra}}\{P(X|\text{szó})\,P(\text{szó})\} \quad (2)$$

Ebben az alakban a P (X) tagot a nevezőből elhagytuk. A P (X|szó) valószínűséget az akusztikai, a P (szó) valószínűséget pedig a nyelvi modell adja meg. Az akusztikai modellnek tehát arról kell informálnia, hogy adott akusztikai megfigyelés az egyes szavakra milyen valószínűségű, a nyelvi modellnek pedig arról, hogy az egyes szavak előfordulásának mekkora a becsült valószínűsége.

A szótárban szereplő minden egyes kulcsszóhoz létre kell hozni a hozzá tartozó rejtett Markov–modellt. Beszédhangok esetén általában háromállapotú lineáris struktúrájú modellt (ún. balról – jobbra) szokás választani (1. ábra). Magát a modellezést például diádok (a diád olyan fonémakapcsolat, ami két hangból tevődik össze, és az első hang felétől a második hang feléig tart) esetén három vagy több állapot végzi, valójában azonban két további szélső állapotot is találunk, amelyek az egyes beszédelem-modellek összefűzését biztosítják.

Felismeréskor a rendszer számára minden keret érkezésekor két lehetőség áll fent, vagy állapotot változtat, vagy helyben marad, bizonyos valószínűséggel. Ezeket nevezzük állapot-átmeneti valószínűségeknek, melyek becslése a tanítás során történik. Ez a mechanizmus biztosítja az időbeli illesztést a modell és az aktuális keret között. A rendszer az adott (belső) állapotból két keret érkezése között

egy megfigyelést bocsát ki, mely tulajdonképpen egy hasonlósági mérték az adott állapotra jellemző jellemzővektor-eloszlás és az aktuálisan érkezett, a külső megfigyelést reprezentáló jellemzővektor között. Lényegében azt mondhatjuk, hogy e hasonlósági mérték a mérőszáma a megfigyelt jellemzővektor és a modellállapot spektrális illeszkedésének. Egy állapotra jellemző jellemzővektor-eloszlást általában sűrűségfüggvényével adunk meg, amelyről feltételezzük, hogy normális (*Gauss*) eloszlások lineáris kombinációjából áll elő. Ezt szokás kibocsátási valószínűségnek is nevezni [2].



1. ábra. *3 állapotú lineáris modell*

A mi esetünkben a kulcsszavak hosszúsága nagyon eltérő, ezért nem egyszerű feladat a megfelelő HMM megválasztása. Kiindulásként egy 8 állapotú modellt választottam. Korábbi vizsgálataim során, ahol diád alapú felismerőn végeztem vizsgálatokat a 8 állapotú modell bizonyult optimálisnak.

A kulcsszavak közül a legrövidebb szavak 2 hangból tevődnek össze, ezért egyértelmű, hogy ennél kisebb állapotszámú rejtett Markov-modellel nem érdemes vizsgálni a felismerő hatékonyságát.

3. Az alkalmazott lényegkiemelési eljárás bemutatása

A lényegkiemelés a hallás bemutatása során megismert biológiai jellemzők ismeretében valósítható meg. A HTK toolkit keretrendszer tartalmaz olyan modult, amivel elvégezhető az átalakítás. Munkám során ezzel valósítottam meg a lényegkiemelést. A folyamat a jelből megkísérli meghatározni a beszéd tartalmát hordozó mennyiségeket, azaz a fontos információkat, és kiküszöbölni a felismerés szempontjából érdektelen információkat (zaj, fázis, torzítások). A digitalizált (beszéd) jelből egy diszkrét idejű, adott dimenziójú lényegvektor-sorozatot alkot.

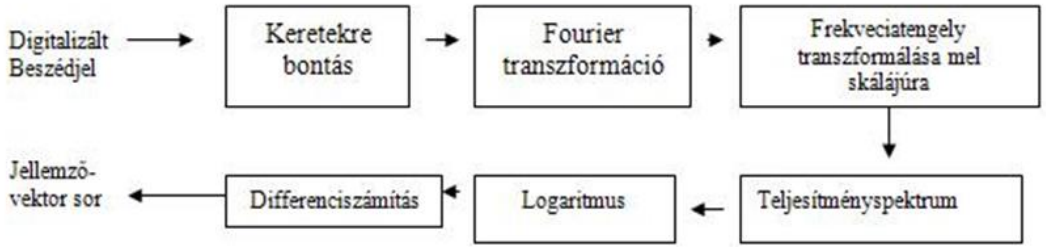
Fontos információ alatt itt a mel-frekvenciás kepsztrális komponenseket (*Mel-Frequency Cepstral Coefficient / MFCC*) értjük [5]. Az *MFCC* jellemzők a beszéd lényegi tartalmának kinyerésére szolgálnak (2. ábra) és napjainkban a beszédfelismerő rendszerek jelentős többsége e jellemzők (vagy ennek kicsit módosított változatai) segítségével próbálják reprezentálni a beszéd lényegi információtartalmát.

A *digitalizálás* folyamata nem csak a mintavételezésből áll. A mintavételezett jel ez után kvantáláson esik át: az egyes minták nem vehetnek fel tetszőleges értéket, mivel binárisan ábrázoljuk őket, tehát csak diszkrét értékeik lehetnek. A bitmélység, azaz hogy hány biten reprezentáljuk, a kódolás során az egyes mintákat, jellemzi a kvantálás részletességét. Legelőször normalizálni kell a hangjeleket: mivel véges értéken ábrázoljuk a mintákat, és torzítás nélkül szeretnénk visszaállítani az eredeti jelet, azt adott jelszint tartományba kell szűkíteni még digitalizálás előtt. Ez után következhet a mintavételezés és kvantálás. Utóbbi a legtöbb esetben 8 vagy 16 bit mintánként. A nagyobb bitmélységgel reprezentált hang nagyobb adatmennyiséget is jelent, ami növeli a feldolgozási időt és a szükséges tárhelyet, viszont csökkenti a kvantálási zajt. A kvantálás lehet lineáris és logaritmikus.

A gyors Fourier-transzformáció (FFT), majd gördülő spektrum generálása: a digitalizált jelből N mintányi kereteket kivéve, majd azokon az FFT algoritmust elvégezve a keret spektrális tulajdonságait leíró vektort kapunk. Ha a kereteket „csúsztatjuk” a minták mentén, és N-nél kisebb mintával odébb

kezdve újabb N méretű kereten is elvégezzük az FFT-t, akkor a spektrális viselkedés időbeli lefolyását reprezentáló vektorsorozatot kapunk, ezt nevezzük *spektrogramnak*.

A kapott vektorsorozatnak a korábban ismertetett *Bark-skála szerinti szűrése*, majd az egyes sávokba eső komponensek energiaértékének maghatározása pedig olyan jellemzővektor-sorozatot eredményez, mely az emberi fül számára lényeges információt tartalmazza, sokkal tömörebb Ezeket a fájlokat nevezzük Mel Frequency Ceptrum Coefficient, azaz MFCC fájloknak. A paraméterek (például N, és a bark szűrők száma, valamint a keretek csúsztatásának mértéke) változhatnak, azaz a lényegkiemelés több tömörítési mélységben is reprezentálhatja a hangmintát. Sokszor a szűrők energiájának értékei mellett azok megváltozását is tárolja az MFCC fájl, így több információt hordoz az időtartománybeli jellemzőkről.



2.ábra. *A lényegkiemelés folyamat*

Ahhoz, hogy végre tudjuk hajtani a konvertálást, több paraméter beállítása is szükséges, melyek közül kiemelten a továbbiakban az állapotszámok változtatásának a hatékonyságra gyakorolt hatását vizsgáljuk és a többi paramétert változatlanul hagyjuk. Továbbá a vizsgálatok elvégzéséhez elengedhetetlen a nyelvtani szabály és a szótár megalkotása.

4. A nyelvtan fájl

A *HTK* egy nyelvtani definíciókat létrehozó "nyelvet" nyújt számunkra, aminek a segítségével egyszerű, vagy akár összetett nyelvtani szabályokat is alkothatunk a feladatunknak megfelelően. Ezen szabályok egyfajta reguláris kifejezéseket adnak meg, amelyek állhatnak fonémakapcsolatok sorozatából és metakarakterekből. A definíciók megadhatók egyszerű *txt* fájlokban is.

A létrehozott nyelvtan fájl a vezérléshez szükséges parancsszavakat tartalmazza. Mivel a tanító és tesztelő minták egyaránt tartalmaznak típus szerint szavakat és parancsokat is, ezért a nyelvtan fájlban nem definiálhatjuk a parancs szintaktikáját. Így a nyelvtan fájl megengedi, hogy a szavak tetszőlegesen követhessék egymást, ismétlődések is létrejöhetnek. A nyelvtan fájl által a *HParse* modul egy *SLF* fájlt hoz létre a kiterjesztett *Backus- Naur forma* (EBNF) [6,7] metaszintaxisát felhasználva (környezet független nyelvtanok leírására használható metaszintaxis) átalakítja a *HTK* által értelmezhető nyelvtani szintaktikájú formátumra. Az *SLF* fájl generálása automatikus mivel elengedhetetlen része a rendszerépítés folyamatának. A fájl tartalmazza a parancs kifejezések csomópontjainak listáját, és az ívek listáját, amik a szavak közti átmenetet (sorrendet) reprezentálják.

5. A szótár fájl

A szótár a kulcsszavakat tartalmazza némileg módosult formában (a nyelvtan fájlban szintén alkalmaztam az átírást), melynek oka, hogy a HTK szoftver, nem értelmezi az ékezetes karaktereket, ezért

néhány esetben a SAMPA (a nemzetközi fonetikai ábécé ASCII átírása a SAMPA) szabványt követtem. A mássalhangzók duplázódását minden esetben a ':' jelzi (pl.: vészleállítás -> vESleAl:i:tAs).

6. Különböző állapotszámú rejtett Markov-modellek tesztelése

A felismerés hatékonysága jelentősen függ a választott HMM állapotainak számától. Mivel a kulcsszavak hossza változatos (pl. *el, haladás, pozícionálás* stb.), ezért érdemes a felismerőt nagyobb állapotszámú modellel is tesztelni, mert elképzelhető, hogy javulni fog a felismerés. Minden egyes szótári elemnek létre kell hozni a modelljét, a globális átlagokat és szórást tartalmazó prototípus segítségével, ami a mi esetünkben 26 darab modellt jelent.

1. táblázat. *Értékek változása a HMM állapotszám megválasztásának függvényében*

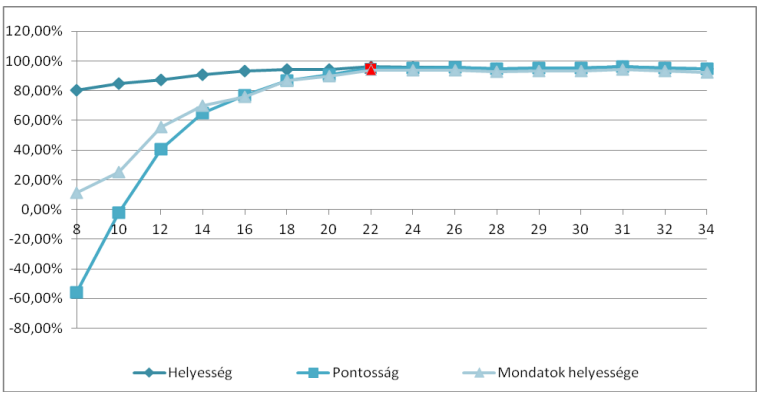
HMM állapotszám	Helyesség	Pontosság	Mondatok helyessége
8	80,42%	-55,94%	11,11%
10	84,62%	-2,10%	25,25%
12	87,41%	40,56%	55,56%
14	90,91%	65,03%	69,70%
16	93,01%	76,92%	75,76%
18	94,41%	86,71%	86,87%
20	94,38%	90,87%	89,83%
22	96,47%	95,06%	93,86%
24	95,73%	95,02%	93,79%
26	95,69%	95,69%	93,71%
28	94,92%	94,92%	92,53%
29	95,35%	95,35%	93,14%
30	95,33%	95,33%	93,09%
31	96,06%	96,06%	94,16%
32	95,21%	95,31%	93,04%
34	94,96%	94,96%	92,45%

Az 1. táblázatban az egyes állapotszámokra elvégzett tesztelések eredménye látható. A mondatok helyessége, ebben az esetben azt jelenti, hogy az összes hangminta közül hánynak a tartalmát ismeri fel hiba nélkül. A szavak típusúnál ez egy szót jelent, parancsoknál pedig a szavak összességét. Ahhoz, hogy érthetőek legyenek az eredmények fontos, hogy a *HTK* miként számolja ki a pontosság és a helyesség értékét. A két képlet mutatja meg ezt, ahol *H* a helyesen felismert szavak számát, *N* az összes szó számát, *I* pedig a tévesen beszúrt szavak számát jelenti.

$$\%Correct = \frac{H}{N} \times 100\% \qquad \%Accuracy = \frac{H - I}{N} \times 100\% \qquad (3,4)$$

A diagramon (3. ábra) láthatjuk, hogy a 8 és a 22 állapotszám között a pontosság és ezzel együtt a mondatok helyessége meredeken növekszik, majd a 22 és a 34 állapotszám között a görbék meredeksége a nullához közelít (konstansnak tekinthető). Az 1. táblázatban is láthatjuk, hogy csak néhány ti-

zed százalékban változnak állapotszámonként az értékek ezekben az esetekben, amik nem tekinthetők egyértelmű javulásnak vagy romlásnak. A 26. állapotszámtól a helyességi és pontossági értékek megegyeznek, ami azt jelenti, hogy a beszúrások száma (*I*) nullára csökkent.



3. ábra. *Az értékek változása a HMM állapotszám megválasztása függvényében diagramon ábrázolva*

Korábbi munkám során, ahol diádokon végeztem hasonló vizsgálatot, a helyességi és pontossági értékek nem ingadoztak több állapotszámon keresztül egy érték körül [2]. Egyértelműen kiválasztható volt az optimális állapotszám. Ebben az esetben viszont csak a 34 állapotszám vizsgálatánál látható romlás. Ennek oka az lehet, ha figyelembe veszem korábbi tapasztalataimat, hogy a diádok adott hosszúságú fonéma kapcsolatok, ezzel szemben a kulcsszavak hossza nagy változatosságot mutat, és a tesztelő beszédadatbázis nem tartalmaz ugyanannyi mintát minden szó esetén. Elképzelhető, hogy a kisebb állapotszámoknál a kevesebb hangból álló szavak miatt magasak a hatékonysági eredmények, és ha ugyanezekkel az állapotszámokkal vizsgálnánk csak a hosszabb szavakat, akkor az értékek jelentősen csökkennének, és fordított esetben ugyanúgy romlanának az eredmények. E két szélsőséges eset közötti átmenetet képezheti a 22 és 34 állapotszám közötti minimális ingadozás. Ha egyértelműen választanom kellene, hogy melyik esetet tekinteném leghatékonyabbnak, akkor a 22 állapotszám mellett döntenék, mivel az a legkisebb állapotszám közel azonos hatékonysági értékek mellett, így a számítási igény is kisebb.

A megállapítások alapján további vizsgálatokat lenne érdemes elvégezni a HMM állapotszámokra vonatkozóan úgy, hogy az egyes kulcsszavakhoz azok hosszától függően különböző állapotszámú modelleket rendelnénk. A következő fejezetben felhasználva az eddigi eredményeket megkíséreltem a megfelelő állapotszámot hozzárendelni a különböző hosszúságú szavakhoz.

7. Különböző állapotszámú rejtett Markov-modellek alkalmazása az egyes parancsszavakhoz

7.1 Szabály nélkül megválasztott HMM állapotszámok vizsgálata

A kulcsszavakat attól függően, hogy hány különböző hangból tevődnek össze, csoportokba sorolhatjuk. A csoportosításnál a mássalhangzók duplázódását nem vettem figyelembe, így az alábbi felsorolásban szereplő módon osztottam be a kifejezéseket:

- 2 hangból állók: le, öt egy,
- 3 hangból állók: fel, négy, hat, hét
- 4 hangból álló: stop, gyors, lassú, vissza, nulla, kettő nyolc

- 5 hangból álló: start, előre, hátra, váltó, három,
- 6 hangból álló: állomás, kilenc
- 7 hangból álló: haladás
- 8 hangból álló: fordítás, váltóban
- 11 hangból álló: vészleállítás
- 12 hangból álló: pozícionálás

A 2. táblázat az egyes vizsgálatok esetén megválasztott rejtett Markov-modellek állapotszámát tartalmazzák a különböző hosszúságú szavakra vonatkozóan.

2. táblázat. *Az egyes vizsgálatoknál felhasznált rejtett Markov-modellek állapotszámai*

	2	3	4	5	6	7	8	11	12
mix1	12	16	20	30	34	34	34	42	42
mix2	12	16	24	34	34	38	38	42	42
mix3	11	15	26	34	34	40	40	42	42
mix4	11	15	26	38	38	42	42	42	42
mix5	11	15	26	40	40	42	42	42	42
mix6	11	15	26	30	40	40	40	46	46
mix7	11	15	26	29	40	40	40	46	46
mix8	11	15	26	32	40	40	40	46	46
mix9	11	15	26	30	34	42	42	50	50
mix10	11	15	28	30	36	42	46	50	50
mix11	11	15	26	30	36	42	42	54	54
mix12	11	15	26	30	36	42	46	54	54
mix13	12	16	26	30	34	42	42	50	50
mix14	12	14	22	24	28	30	30	40	40
mix15	11	15	20	24	26	28	30	40	40
mix16	12	16	20	24	28	30	32	38	38
mix17	12	16	18	22	26	28	30	36	36
mix18	10	14	20	24	26	28	30	38	38
mix19	12	16	20	22	26	28	30	38	38
mix20	12	14	20	24	26	28	30	38	38

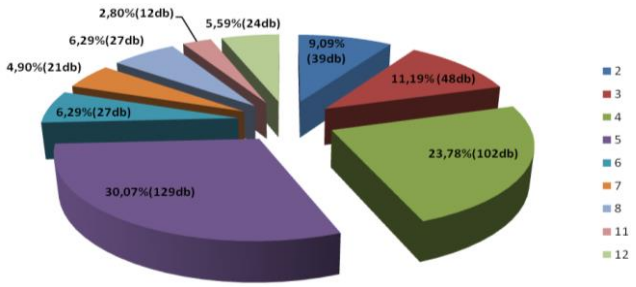
A táblázatban jól látható, hogy egyes csoportoknál megpróbáltam széles intervallumban megvizsgálni az állapotszámokat. Példaként ha megnézzük az 5 hangból álló szavak állapotszámának változását a vizsgálatok során, akkor azok 22 és 40 közötti intervallumba esnek. Korábbi tapasztalataimat figyelembe véve (diádok esetén a leghatékonyabb a 8 állapotszámú modell bizonyult [2]) a 2 hangból álló szavakhoz 8 vagy annál nagyobb állapotszámot rendeltem. Az elvégzett vizsgálatok eredményeit a 3. táblázat tartalmazza.

A táblázatban a helyességi és pontossági értékek pár százalékos ingadozást mutatnak, a mondatok helyességi értéke viszont jelentősebb változást mutat. A pirossal jelölt esetekben (összesen 5 esetben) sikerült ~1,5% pontossági javulást elérnem a mondatok helyességénél pedig ~0,5%-ot.

A tesztelő anyag összetétele nagyban befolyásolja az eredményeket, ezért érdemes megvizsgálni, hogy az egyes szavak hányszor szerepelnek a tesztelőben, és így az egyes hangcsoportok tagjainak összelőfordulása is hogyan oszlik el.

3. táblázat. *A helyességi és pontossági értékek az egyes vizsgálatok esetén*

HMM állapotok	Helyesség	Pontosság	Mondatok helyessége
mix1	95,80%	95,80%	93,94%
mix2	95,10%	95,10%	92,93%
mix3	95,10%	93,01%	89,90%
mix4	95,10%	93,01%	89,90%
mix5	95,80%	95,80%	93,94%
mix6	95,10%	93,01%	89,90%
mix7	95,10%	93,01%	89,90%
mix8	95,10%	95,10%	92,93%
mix9	95,10%	93,01%	89,90%
mix10	95,10%	93,01%	89,90%
mix11	95,10%	93,01%	89,90%
mix12	95,10%	93,01%	89,90%
mix13	95,10%	93,01%	89,90%
mix14	96,50%	96,50%	94,95%
mix15	95,80%	95,80%	93,94%
mix16	96,50%	96,50%	94,95%
mix17	96,50%	96,50%	94,95%
mix18	95,80%	93,71%	90,91%
mix19	96,50%	96,50%	94,95%
mix20	96,50%	96,50%	94,95%



4.ábra. *A tesztelő anyag szavainak eloszlása hangcsoportok szerint*

A 4. ábrán jól látható, hogy az 5 hangból álló szavak 30,03%-t teszik ki a tesztelő mintáknak, a 4 hangból állók pedig 23,78%-t. Ez összesen 53,81%, ami több mint a fele a tesztelőknek. Ebből arra következtethetünk, hogy az állapotszám megválasztásnál a hatékonysági értékek változása ennek a két hangcsoportnak az állapotszám megválasztásaitól függ a leginkább.

Az előfordulások eloszlása alapján megadhatjuk az egyes vizsgálatoknál az állapotszámok súlyozott átlagát is (5. táblázat). Szintén pirossal jelöltem a mérések közül a legjobb eredményűeket.

Az állapotok igen változatosak. A 14-20 vizsgálatoknál a súlyozott állapotszám 21-24 között változik. Az előző fejezetben, amikor minden szóhoz ugyanazt a HMM-t alkalmaztam megállapítottam, hogy 22 állapotszám körül válnak a hatékonysági értékek maximálissá és több állapotszámon keresztül csak minimálisan változik. A pirossal jelölt értékek (a jelenlegi vizsgálatok közül a legjobbak) pedig a 14,16,17,19,20 vizsgálatoknál jelentkeztek. Tehát megállapíthatjuk, hogy ha a szavak hosszától függően választottuk meg az állapotszámot, és azok hatékonysága szintén a legjobb, akkor azok súlyozott átlaga közelíteni fog az előző fejezetben megválasztott leghatékonyabb állapotszámhoz. Valamint a 4 és 5 hangból álló hangcsoportokhoz is a 22 állapotszámhoz közeli érték lettek megadva ezekben az esetekben. Ez összességében jó kiindulási alapot nyújt a jövőben, hasonló vizsgálatok elvégzéséhez, mivel a szélsőséges hangcsoportok (pl. 2 hangból álló szavak, vagy 12 hangból álló szavak csoportja) további tesztelések alapján határozhatók meg.

4. táblázat. *Az állapotszámok súlyozott átlaga vizsgálatonként*

	HMM állapotszám súlyozott átlaga
mix1	26,13
mix2	28,73
mix3	29,22
mix4	30,90
mix5	31,63
mix6	28,73
mix7	28,43
mix8	29,34
mix9	28,92
mix10	29,77
mix11	29,38
mix12	29,63
mix13	29,12
mix14	23,58
mix15	22,90
mix16	23,29
mix17	21,69
mix18	22,53
mix19	22,34
mix20	22,71

Ha megfigyeljük az egyes tesztek esetén az állapotszámok változását, akkor több esetben szabályszerűségeket fedezhetünk fel. Az első öt állapotszám szabályosan kettővel növekszik. Ebben az esetben *4+n*4* képlettel adhatók meg a sorozat elemei, ahol *n* a hangok számát jelenti. A következő fejezetben több ilyen esetet fogok megvizsgálni, felhasználva az eddigi tesztekben fellelhető szabályszerűségeket.

7.2 Képlet alapján megválasztott HMM állapotszámok vizsgálata

Figyelembe véve az előző értékeket és teszteket létrehoztam a 6. táblázatban felsorolt képleteket. Nem egyszerre végeztem el a vizsgálatokat, hanem figyelembe vettem folyamatosan az eredményeket és azok alapján próbáltam tovább formálni a képleteket változtatva a növekményen vagy a kiindulási értéken, hogy további hatékonyság javulást tudjak elérni. A képletekben az *n* a hangok számát jelenti.

A táblázatokban (6.- 7. -8. táblázat) pirossal jelöltem azokat az eseteket, ahol a legjobb eredmények születtek, ami helyesség és pontosság esetén is 96,50% a mondatok helyessége pedig 94,95% (8. táblázat). Ezek az értékek megegyeznek az előző fejezetben elért legjobb eredményekkel, így nem sikerült további javulást elérnem, de érdemes megvizsgálnunk ezeknél a teszteléseknél (és persze az összes többinél is, hogy legyen mihez viszonyítanunk), hogy szabályos növekményeknél az állapotszámok súlyozott átlagai hasonlóan az előzőekben tapasztaltakhoz a 22 állapotszám körül ingadoznak-e.

5. táblázat. *Az egyes tesztekhez tartozó képletek*

	Képlet
teszt1	3+4*n
teszt2	10+4*n
teszt3	5+3*n
teszt4	1+5*n
teszt5	10+3*n
teszt6	8+2*n
teszt7	4+4*n
teszt8	14+2*n
teszt9	2+4*n
teszt10	12+2*n
teszt11	5+4*n
teszt12	15+2*n
teszt13	8+3*n
teszt14	6+3*n
teszt15	3+4*n
teszt16	16+2*n

6. táblázat. *Az egyes vizsgálatoknál felhasznált rejtett Markov-modellek állapotszámai*

	2	3	4	5	6	7	8	11	12
teszt1	11	15	19	23	27	31	35	47	51
teszt2	18	22	26	30	34	38	42	54	58
teszt3	11	14	17	20	23	26	29	38	41
teszt4	11	16	21	26	31	36	41	56	61
teszt5	16	19	22	25	28	31	34	43	46
teszt6	12	14	16	18	20	22	24	30	32
teszt7	12	16	20	24	28	32	36	48	52
teszt8	18	20	22	24	26	28	30	36	38
teszt9	10	14	18	22	26	30	34	46	50
teszt10	16	18	20	22	24	26	28	34	36
teszt11	13	17	21	25	29	33	37	49	53
teszt12	19	21	23	25	27	29	31	37	39
teszt13	14	17	20	23	26	29	32	41	44
teszt14	12	15	18	21	24	27	30	39	42
teszt15	11	15	19	23	27	31	35	47	51
teszt16	20	22	24	26	28	30	32	38	40

7. táblázat. *A helyességi és pontossági értékek az egyes vizsgálatok esetén*

	Helyesség	Pontosság	Mondatok helyessége
teszt1	95,80%	95,80%	93,94%
teszt2	95,10%	95,10%	92,93%
teszt3	93,71%	89,51%	87,88%
teszt4	96,50%	96,50%	94,95%
teszt5	96,50%	96,50%	94,95%
teszt6	94,41%	87,41%	85,86%
teszt7	96,50%	96,50%	94,95%
teszt8	96,50%	96,50%	94,95%
teszt9	95,80%	93,01%	90,91%
teszt10	96,50%	96,50%	94,95%
teszt11	95,80%	95,80%	93,94%
teszt12	96,50%	96,50%	94,95%
teszt13	96,50%	96,50%	94,95%
teszt14	96,50%	94,41%	91,92%
teszt15	95,80%	95,80%	93,94%
teszt16	95,78%	95,78%	93,90%

A 9. táblázatban láthatóak a vizsgálatonkénti rejtett Markov-modellek állapotszámainak súlyozott átlaga. Az értékek 18 és 31 közé esnek. A pirossal jelölt esetekben pedig 22 és 27 közé. A legtöbb pirossal jelölt érték inkább 24 körül ingadozik. Ez az eltérés a 22 állapotszámtól nem dönti meg az eddigi megállapításainkat, főleg ha figyelembe vesszük azt, hogy a 7. fejezetben nem lehetett egyértelműen meghatározni az optimális állapotszámot, hanem csak egy intervallumot ahol a hatékonysági értékek maximalizálódnak.

A táblázatban láthatók olyan esetek, ahol az értékek jobban megközelítik a 22-t (pl. 3,9,14 tesztek), de ezekben az esetekben valószínűleg az 5 és annál kevesebb hangból álló csoportokhoz hozzárendelt állapotszám túl kevésnek bizonyul összességében ezért nem sokkal, de romlik a hatékonyság.

Végeredményül azt egyértelműen kijelenthetjük, hogy ezeknél a vizsgálatoknál és az előbbieknél sem mutatkozott olyan eset, ahol a hatékonyság a legjobb lett volna, és a súlyozott állapotszám pedig drasztikusan eltért volna az előző fejezetben megállapított intervallum értékektől (ahol a beszédfelismerő hatékonysága maximalizálódott), tehát az állapotszámok súlyozott átlagának figyelembevétele jó kiindulási alapot ad hasonló vizsgálatokhoz.

8. táblázat. *Az állapotszámok súlyozott átlaga vizsgálatonként*

	HMM állapotszám súlyozott átlaga
teszt1	23,70
teszt2	30,70
teszt3	20,52
teszt4	26,87
teszt5	25,52
teszt6	18,35
teszt7	24,70
teszt8	24,35
teszt9	22,70
teszt10	22,35
teszt11	25,70
teszt12	25,35
teszt13	23,52
teszt14	21,52
teszt15	23,70
teszt16	26,35

Figyelembe véve a képletek alapján, és a ”véletlenszerűen” létrehozott HMM összetételeket, azok közül amelyeknél mind a helyesség és mind pedig a pontosság a legjobb volt, azt tekintem leghatékonyabbnak, ahol az állapotszámok súlyozott átlaga a legkisebb. Ez pedig az előző fejezet 17. vizsgálata, ahol a súlyozott átlag 21,69. A további teszteléseket - amik a zaj a beszédfelismerés hatékonyságára gyakorolt hatását vizsgálják – ezekkel az állapotszámokkal fogom elvégezni.

8. Összefoglalás

Munkám során a HTK toolkit szoftvercsomag alkalmazásával betanítottam egy kis szótáras beszédfelismerőt, amin a rejtett Markov–modellek állapotszámának változtatásával próbáltam elérni a lehető legjobb hatékonyságot. Az első esetben minden szó esetén ugyanolyan állapotszámú modellt alkalmaztam. Kiindulási modellek állapotszámnak a nyolcat választottam, majd folyamatosan növelve az állapotszámot elemeztem a létrejövő helyességi és pontossági értékeket, amik azt a következtetést tudtam levonni, hogy a beszédfelismerő hatékonysága a 22 állapotszámú HMM esetén éri el a maximumát. Továbbiakban vizsgálataimban a parancsszavak hosszától függően különböző állapotszámú modelleket rendeltem az egyes szavakhoz. 20 esetben vizsgáltam meg a hatékonyságot, és az egyes eseteknél a rejtett Markov-modellek állapotszámainak súlyozott átlagát. Megfigyelhető volt, hogy azokban az esetekben, ahol ezzel a módszerrel sikerült javítanom a hatékonyságon (helyesség: 96,50%, pontosság: 96,50%, mondatok helyessége: 94,95%), az állapotszámok súlyozott átlaga 22 érték körül mozogtak, ami az első esetben végzett vizsgálatok legjobb eredménye.

A szavak hosszától függő állapotszámok megválasztásánál, egyes eseteknél az állapotok változása között szabályszerűségek fedezhetők fel. Ezen szabályszerűségek alapján további teszteléseket végeztem, és képleteket alkalmaztam az állapotszámok meghatározásához. Több esetben születtek az előző legjobb eredménnyel egyenlő értékek. Ezekben az esetekben a HMM állapotszámok súlyozott átlaga 24 körül ingadozott. A 24 állapotszám érték alkalmazása a szavak hosszától független állapotszám megválasztásnál, beleesik abba a tartományba, ahol a hatékonyság maximalizálódik. Összességében azt a következtetést vonhatjuk le, hogy ha olyan esetekben, ahol az egyes szavakhoz, azok hosszától függően szeretnénk állapotszámot megválasztani és maximalizálni a beszédfelismerő hatékonyságát, akkor jó kiindulási alapot ad, ha megvizsgáljuk, hogy azonos HMM-k alkalmazásánál, melyik állapotszám értéknél lesz a leghatékonyabb a beszédfelismerő és a továbbiakban úgy megválasztani az eltérő állapotokat, hogy azok súlyozott átlaga ehhez az értékhez közelítsen.

9. Köszönetnyilvánítás

A cikkben ismertetett kutató munka az EFOP-3.6.1-16-2016-00011 jelű „Fiatalodó és Megújuló Egyetem – Innovatív Tudásváros – a Miskolci Egyetem intelligens szakosodást szolgáló intézményi fejlesztése” projekt részeként – a Széchenyi 2020 keretében – az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

Irodalom

[1] Németh, G., Olaszy, G.: *A magyar beszéd*, Akadémiai Kiadó, Budapest, 2010

[2] Pintér, J. M.: *Akusztikus beszédfelismerés rejtett Markov-modell alkalmazásával*, szakdolgozat 2010

[3] Brooks, S., et al., eds.: *Handbook of Markov Chain Monte Carlo*, CRC press, 2011. **https://doi.org/10.1201/b10905**

[4] Deng, L.: *A generalized hidden Markov model with state-conditioned trend functions of time for the speech signal*, Signal Processing 27.1 (1992): 65-78. **https://doi.org/10.1016/0165-1684(92)90112-A**

[5] Muda, L., Begam, M., Elamvazuthi, I.: *Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques*, Journal of Computing, 2010, 2:138-143.

[6] Knuth, D. E.: *Backus normal form vs. backus naur form*, Communications of the ACM 1964, 7(12):735-736. **https://doi.org/10.1145/355588.365140**

[7] Farrel, A.: *Routing Backus-Naur Form (RBNF): A syntax used to form encoding rules in various routing protocol specifications*, RFC 2009, 4:5511. **https://doi.org/10.17487/rfc5511**