Nyelvfüggetlen fonémaszintű automatikus szegmentáló fejlesztése

TDK dolgozat, 2013.

Készítette:
Tulics Miklós Gábriel, Villamosmérnöki szak, MSc I. évfolyam
Konzulens:
Vicsi Klára, Távközlési és Médiainformatikai Tanszék

Tartalomjegyzék

1 Bevezetés

A beszédfeldolgozásban számos olyan terület van, ahol szükséges a folyamatos beszéd feldolgozása. Ilyenkor a beszéd fonémaszintű szegmentálására van szükség. Vannak olyan alkalmazások, ahol a pontos fonéma ismerete nem szükségszerű, csak a hang típusa a fontos, vagyis, hogy nazális, magánhangzó, zöngés, zöngétlen típusú-e a hang. Ezekben az alkalmazásokban a nyelvi tartalom nem alapvető, az akusztikai jellemzők a fontosak. A beszéd-szöveg átalakító rendszerek a beszédjelet szöveggé alakítják, viszont nem foglalkoznak a fonémák pontos időbeli információjával. Azonban erre a fajta szegmentációra van szükség, amikor a vizsgált jelenség függ a beszéd időzítésétől. Ilyen például a ritmus, vagy a magánhangzók helyének pontos meghatározása. Erre például akkor lehet szükség, ha olyan tulajdonságokra vagyunk kíváncsiak, amelyek egy-egy beszédhangra jellemzőek, vagy ha az artikulációs sebességet szeretnénk mérni. Ez a szegmentációs technika fontos segédeszköz a beszéd akusztikai paramétereinek a megjelenítésében, az audió-vizuális kiejtésoktató rendszerekben is [1, 2, 3, 4].

A csupán akusztikai ismeretekre támaszkodó (nyelvi jelentést figyelmen kívül hagyó) szegmentálási rendszerek alkalmasak lehetnek nyelvfüggetlen megvalósításra. Az irodalmakban megtalálhatóak ilyen rendszerek, ám azok két vagy három akusztikai-fonetikai osztályos (zöngés, zöngétlen, csend) felismerést valósítanak meg [5, 6, 7, 8]. Tanulmányomban egy kilenc osztályos nyelvfüggetlen folyamatos beszédszegmentáló rendszert mutatok be, amely automatikusan bejelöli az egyes fonetikai osztályok időbeli határait. A határok bejelöléséhez az egyes fonetikai osztályok akusztikai modellezését valósítom meg. Az optimális modellek létrehozásához a bemeneti beszédjel több fajta akusztikai előfeldolgozását végzem el, és azok teljesítményét pedig összehasonlítom. A fonémacsoportok osztályozását és szegmentációját a Szupport Vektor Gép alapú gépi tanuló eljárással valósítottam meg. A Szupport Vektor Gépekkel történő felismerés után szabályalapú, valamint statisztikai elven működő utólagos simítást, fonémahatár szegmentálási megvalósítást alkalmaztam. A nyelvfüggetlenség értékeléséhez három adatbázist használtam fel, a magyar MRBA [9], a német KIEL [10] és az angol TIMIT [11] adatbázisokat.

2 Elméleti összefoglaló

2.1 Fonéma osztályok ismertetése

A *beszéd szegmentálásán* a folyamatos beszédben elkülönített adott méretű egységeket értjük [12]. Ezek a *szegmentálási egységek* különböző hosszúságúak lehetnek, attól függően, hogy milyen szinten történik a feldolgozás, például: gondolat, mondat, szó, fonéma. Jelen dolgozatban fonéma szintű szegmentálási egységeket használtam.

A *fonéma* olyan legkisebb elemi beszédegység, amelyből egy nyelv szavai épülnek fel és az egyes szavakban jelentés megkülönböztető szerepe van.

A fonémák jelöléséhez egyezményes jeleket kell használnunk, hogy általános esetben a különböző hangokat nemzetközileg írásos formában jelölni tudjuk. Ezek a jelek ugyanazokat a hangokat jelölik a különböző nyelvekben. Két ilyen nemzetközi jelölésrendszer használatos: az *IPA* jelölések és a *SAMPA* jelölések. Az IPA jelölések képesek a fonémák illetve a fonéma variánsok jelölésére is. A szimbólumkészletet a nemzetközi Fonetikai Társaság (International Phonetic Association) alkotta meg 1889-ben [13]. Ezután fejlesztették ki, a számítógépes billentyűzet karakterkészletéhez igazított SAMPA szimbólumrendszert, amit magyar nyelvre is kidolgoztak 1996-ban. A SAMPA jelöléseket a magyar hangokra az 1. táblázatról olvashatjuk le [14]. A táblázatban a csillaggal jelölt fonémák a variánsokat jelölik.

Betű	SAMPA-jel	Betű	SAMPA-jel
á	A:	m, mm	m, m:
a	O	m*	M
o	o	n, nn	n, n:
ó	o:	n*	N
u	u	ny, nny	J,J:
ú	u:	j, ly, jj, lly	j,j:
ü	y	j*	X)
ű	y:	h, hh	h, h:
i	i	h*	h
í	i:	h*, ch*	x
é	e:	v, vv	v, v:
ö	2	f, ff	f, f:
ő	2:	z, zz	z, z:
e	E	sz, ssz	s, s:
b, bb	b, b:	zs, zzs	Z, Z:
p, pp	p, p:	s, ss	S, S:
d, dd	d, d:	dz	dz, dz:
t, tt	t, t:	dzs	dZ, dZ:
gy, ggy	d', d':	c, cc	ts, ts:
ty, tty	t', t':	cs, ccs	tS, tS:
g, gg	g, g:	l, ll	l, l:
k, kk	k, k:	r, rr	r, r:

1. táblázat: A magyar beszédhangok SAMPA jelölései

Az általános beszédfelismerő rendszerek egy adott nyelv fonémakészletét alkalmazzák a beszéd-szöveg átalakítás során, ugyanis azok akusztikai modelljei egy adott nyelvre készülnek el. Ehhez ismerni kell a vizsgált nyelv sajátosságait. Viszont, ha a fonémákat akusztikai jellemzőjük alapján rendeljük csoportokba kizárólag azok akusztikai tulajdonsága alapján (ahelyett, hogy fonéma osztályokra bontanánk őket), nyelvfüggetlenséget érhetünk el. Ez a fajta csoportosítás elegendő ahhoz, hogy minden európai nyelvet lefedjünk. Mivel a fonémaosztályok különböző akusztikai jellemzőkkel rendelkeznek, egymástól időben jól elkülöníthetőek. Így a fonémacsoport határok elegendő információt nyújtanak ahhoz, hogy megtaláljuk a beszédfolyamon belül a fonémák (fonémacsoportok) pontos időbeli

információját. Ugyanakkor probléma merülhet fel ezzel az egyszerűsítéssel, ha a beszédsorozatban két egymást követő fonéma ugyanabba az akusztikai osztályba esik, így nem lehet őket elkülöníteni. Ez a probléma viszont nem lényeges olyan alkalmazásoknál ahol a szótagok felismerése elegendő, mint például a prozódia felismerése vagy a ritmus észlelése. A kilenc fonetikai osztály tartalmaz még egy „csend" osztályt is kiegészítésképpen a beszéd nélküli szakaszok modellezésére. Ennek a hosszabb szövegek szegmentálásakor van nagy szerepe, hiszen a beszéd maga tartalmazhat csend szakaszokat is. Az általam alkalmazott fonémacsoportokat és azok jelöléseit az 2. táblázat tartalmazza.

Osztály neve	**Osztály jele**	**Osztályba sorolt fonémák (SAMPA jele)**
Mély (és közép) magánhangzók	mv	O, A:, E, o, o:, u, u:, 2, 2:
Magas magánhangzók	hv	e:, i, i:, y, y:
Zöngétlen spiránsok	s-	f, s, S, h,
zöngés spiránsok	s+	v, z, Z
Nazálisok és likvidák	na	r, l, j, m, n, J
Zöngés zárszakasz	b+	+*
Zöngés felpattanás	vo	b, b:, d, d:, g, g:, dz, dz:, dZ, dZ:, d', d':
Zöngétlen zárszakasz	b-	-*
Zöngétlen felpattanás	uv	p, p:, t, t:, k, k:, ts, ts:, tS, tS:
Csend	si	

2. táblázat: Nyelvfüggetlen fonémaosztályok jelölése

2.2 Support Vector Machine osztályozó

Szupport vektor gépek (SVM) olyan kernel gépeket tartalmazó gépi tanuló eljárás, amely a bemenő adatokat elvi szinten egy bázisfüggvény segítségével a jellemző térbe transzformálja, így az adatok egy jobb reprezentációját biztosítja az eredeti térnél. A jellemzőtérből a következő transzformációs lépésben a kerneltérbe jutunk. A kerneltérben lévő szabad paraméterek száma független a jellemzőtér szabad paramétereinek számától. Ez a kezünkbe adja annak a lehetőségét, hogy végtelen dimenziós jellemzőtérrel dolgozzunk, miközben a kerneltérbeli reprezentációnk véges dimenziójú [25].

Az SVM általános esetben kétosztályos osztályozóként működik, amelyet döntési problémák megoldásához szoktak alkalmazni. Ez azt jelenti, hogy a hálózat két osztály megkülönböztetésére képes. Ha több osztály között szeretnénk dönteni, akkor több SVM-et kell megtanítanunk úgy, hogy minden egyes alkalommal egyetlen osztályt különítünk el a többi osztály uniójával. A matematikai modellből, és a tanítási algoritmusból következik, hogy egyszerre nem fog több osztály mellett dönteni a rendszer.

A folyamat során van egy tanítómintákat tartalmazó halmazunk. Minden tanítóminta egy x vektornak felel meg, és minden ilyen x vektorhoz tartozik egy d elvárt válasz, amely lehet -1 és +1, azaz $d_i \in \{-1,1\}$. A hálózat az aktuális *x*-re ad egy *y* választ. Adottak tehát az $\{\mathbf{x_i}, \mathbf{d_i}\}_{i=1}^{P}$ mintavektorok, és elvárt válaszok. Az összes tanítómintára felírhatjuk az alábbi egyenletet:

$$d_i\left(\boldsymbol{w^T}\,\varphi(\underline{x}) + \, b\right) \geq \, 1 - e_i, \;\; i = 1, 2, \ldots, P,$$

ahol w az **x** vektorokhoz tartozó (mindhez ugyanaz) súlyok vektora, míg b egy skalárral történő eltolás. Az egyenlőtlenség jobb oldalán látható e_i az x_i minta távolsága az elválasztó hipersíktól (2D esetben egyenestől). *P* tanítópontok minden lehetséges mintapárjához tartozó kernel függvény értékét tartalmazza. A $\varphi(\underline{x})$ függvény a bázisfüggvény, amely a jellemző térbe történő transzformációt írja le. Az SVM-eknél is alkalmazott kernel-trükk azonban sosem számol ezzel a függvénnyel, így nem kell meghatároznunk. A trükk lényege, hogy a kernelfüggvény felírható

$$\mathrm{K}\left(\mathbf{x_i}, \mathbf{x_j}\right) = \, \varphi(\mathbf{x_i})^T \varphi(\mathbf{x_j})$$

alakban. A kernelfüggvény megválasztása ennek ellenére tapasztalati úton működik.

A leggyakrabban használt kernel függvények az alábbiak:

Lineáris: $K(\boldsymbol{x}, \boldsymbol{x_i}) = \boldsymbol{x_i^T x}$

Polynomiális: $K(\boldsymbol{x}, \boldsymbol{x_i}) = (\boldsymbol{x_i^T x} + 1)^d$

Gauss (RBF): $K(\boldsymbol{x}, \boldsymbol{x_i}) = exp\{-\|\boldsymbol{x} - \boldsymbol{x_i}\|^2/\sigma^2\}$,

ahol σ konstans.

Tangens hiperbolikusz: $K(\boldsymbol{x}, \boldsymbol{x_i}) = \tanh\left(k\boldsymbol{x_i^T x} + \theta\right)$,

ahol k és θ megfelelően megválasztott konstansok, mert nem minden kombináció eredményez magfüggvényt.

A tanítás során egy optimalizálációs problémára keressük a megoldást. Az optimális hipersikot akkor kapjuk, ha a hibásan osztályozott tanítóminták számát minimalizáljuk. A kifejezés a következőképpen néz ki:

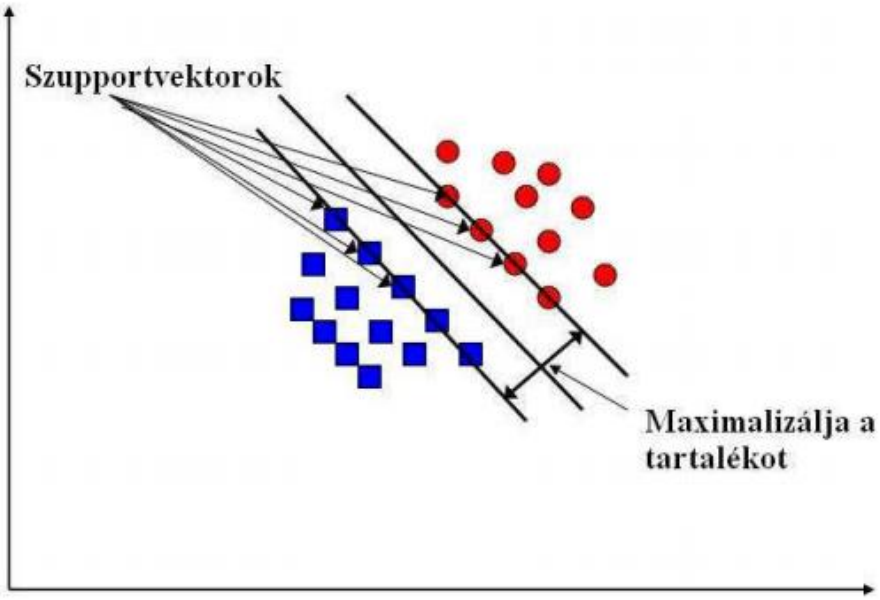
$$\mathrm{J}(\mathbf{w})(\frac{1}{2}\,\boldsymbol{w^T w} + C\sum_{i=1}^{P}\xi_i)$$

C egy korrekciós tényező. Arról ad képet, hogy mennyi tanítóminta eshet biztonsági sávon belülre, akár jó, akár rossz osztályba. Kicsi érték esetén a sáv széles, tehát megnöveli a sávban található tanítóminták számát, míg nagy érték esetén az ellentéte történik. A minimalizálást ezen kifejezés alapján Lagrange multiplikátoros eljárással oldhatjuk meg, amely a következő egyenletre vezet:

$$L(\boldsymbol{w}, e\,, \boldsymbol{\alpha}, \boldsymbol{\gamma}, b) = \frac{1}{2}\boldsymbol{w^T w} + C\sum_{i=1}^{P} e_i - \sum_{i=1}^{P}\alpha_i\,\{d_i(\boldsymbol{w^T x_i} + b) - 1 + e_i\} - \sum_{i=1}^{P}\gamma_i e_i$$

A kritériumot w, b és e_i − k szerint minimalizálni, míg α_i-k és γ_i-k szerint maximalizálni kell. Először a Lagrange kritériumot deriváljuk azok szerint, ami szerint minimalizálni akarunk. Majd a három egyenletet nullával egyenlővé téve, kapunk három összefüggést. Melyeket az eredeti Lagrange kritériumba visszahelyettesítve, kapunk egy duális feladatot, amely megoldható. Az eredményben lesznek 0 értékű α_i-k. Ezek jelölik azon x_i vektorokat,

amelyek nem vesznek részt a hipersík meghatározásában. Minden olyan x_i, amelynek 0-tól különbözik az α_i paramétere, egy szupport vektor (1. ábra) [16].



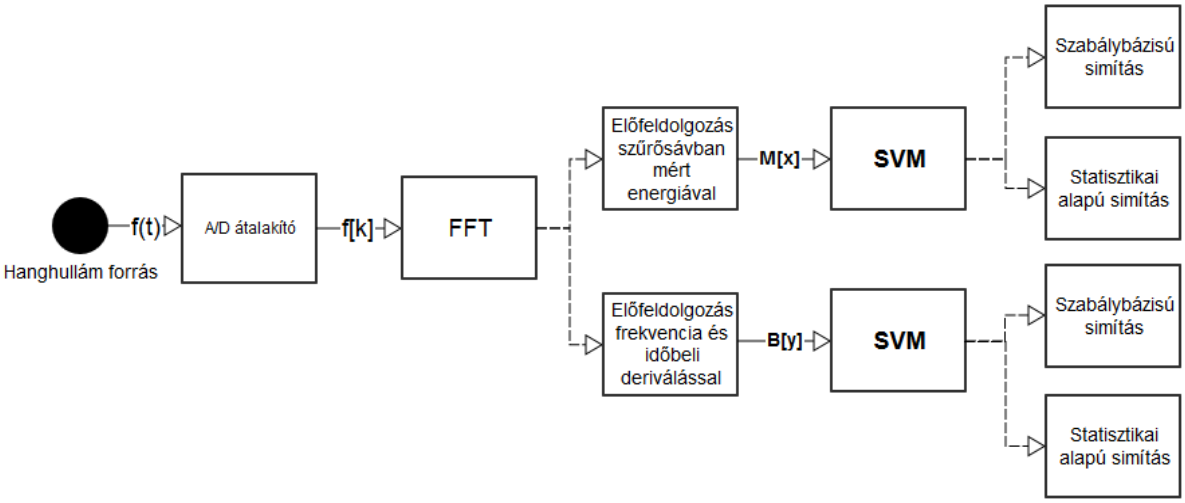
1. ábra: Két osztály, amely egy hipersíkkal elkülöníthető

Az SVM tanítása során a duális feladat megoldása numerikus probléma, amelyre a kvadratikus programozás nyújt segítséget.

3 A munka célja és a vizsgálati módszerem ismertetése

A munkám célja egy olyan szegmentáló kifejlesztése, amely nyelvtől függetlenül képes a 2.1-es fejezetben felsorolt kilenc akusztikai osztály elkülönítésére és időbeli meghatározására. Habár a felismerési feladat mindegyik akusztikai osztály időbeli meghatározására irányul, a gyakorlati alkalmazásokban a magánhangzók detektálása a legfontosabb. Ez hordozza ugyanis a beszéd ritmusbeli információját, valamit a dallam (alaphang menet) jellemzőit. Ezért a felismerési kiértékelése során a magánhangzók felismerési eredménye nagy fontosságú. A két magánhangzó típus megkülönböztetése pontosabb időbeli követést tesz lehetővé (egymás után elhelyezkedő, eltérő típusú magánhangzók megkülönböztetése), ez például a ritmus vizsgálatánál fontos tényező lehet.

A vizsgálati módszer legfontosabb lépése a vektorok generálása, amely szükséges az SVM-nek az osztályok elkülönítésére. Ez gyakorlatilag lényegkiemelés a hullámformából, más szóval akusztikai előfeldolgozás. Munkám során két típusú előfeldolgozást végzek, az egyik módszerben energiaértékek abszolút értékei képezik a tulajdonságvektoraim elemeit, a második módszerben viszont frekvencia és időbeli deriváltakkal dolgozom. Ennek az eljárásnak egy nagy előnye lehet, hogy a derivált értékek miatt kevésbé érzékeny az intenzitásváltozásra, nem kell normálni a felvételeket előtte A beszédjelet mindkét esetben spektrális sávokra bontottam, az első esetben mel-sávos felbontás történik, a másodikban bark-sávos. Mindkét felbontási eljárás az emberi hallás sávfelbontó mechanizmusához illeszkedik, így az eredményeket meghatározó lényegi különbség nincs közöttük. A két eltérő feldolgozási módot a rendelkezésre álló eszközök miatt kellett alkalmaznom. Az előfeldolgozási eljárásokról bővebben a 4-es fejezetben. Az eljárás blokkdiagramját a 2-es ábra szemlélteti.



2. ábra: Munkafolyamat blokkvázlata

Esetemben a hanghullám forrás és az analóg-digitális átalakítás részt a beszédadatbázisok képzik. A felhasznált adatbázisokról a következő fejezetben írok röviden. Az előfeldolgozás gyakorlatilag a digitális jel ablakozásával kezdődik, az FFT, azaz (Fast Fourier Transform) a gyors Fourier transzformáció viszont mindkét előfeldolgozási eljárásban közös, így azt az ábrán nem kezelem külön. A két előfeldolgozás kimeneti vektorainak jelölése a sávszűrők típusára utal. Az SVM kimenete viszont még csak az akusztikai besorolást végzi 10 ms-os lépésenként (25 ms-os ablakmérettel), szükség van olyan eljárásokra, amelyek meghatározzák az elhangzott fonémák pontos határait és összemérhetők az adatbázisban szegmentált hanganyaggal. Ezek az algoritmusok fonémacsoport szintű szegmentálást valósítanak meg az SVM kimenetét felhasználva. Ezeket *simító eljárásoknak* nevezem. Az egyik ilyen eljárás a szabálybázisú simítás, a másik a statisztikai alapú eljárás. Ezekről bővebben a 6. fejezetben írok. Az eljárásokkal kapott eredményeket tévesztési mátrixokba foglalom össze, valamint vizsgálom a szegmentálás pontosságának arányát is, különleges figyelemmel a magánhangzókra.

3.1 A használt adatbázisok rövid leírása

3.1.1 MRBA

Az MRBA név a „magyar referencia beszédadatbázis” szavakból ered [18]. Ez egy általános célú, otthoni/irodai környezetben felolvasott folyamatos szöveget tartalmazó adatbázis, amely alkalmas PC-s beszédfelismerők betanítására, tesztelésére. A korpusz a magyar nyelv leggyakoribb kettős hangkapcsolatainak a 98,8%-a le van fedve. 300 beszélő 12 különböző mondatot és 12 különböző, a mondatoktól független szót olvasott fel. A nyelvi változatosság lefedése érdekében négy különböző tájegységben lévő városban (Budapest, Szeged, Győr, Miskolc) is készültek hangfelvételek. Az életkor és a nemek szerinti eloszlást a 3. táblázat mutatja.

Korcsoport	Férfi beszélők	Női beszélők
16 év alatt	0,9%	3,3%
16-30 év	46,1%	27,7%
31-45 év	5,7%	6,%
46-60 év	3,9%	5,1%
60 év felett	0,9%	0,4%
Összesen	57,5%	42,5%

3. táblázat: A beszélők életkor és neme szerinti megoszlása

Az adatbázis teljes anyaga annotált, harmada (100 beszélő) beszédhangszinten kézileg szegmentált és címkézett. A hanganyagok formátuma: 16 kHz mintavételi frekvencia mellett 16 bit [24].

3.1.2 TIMIT

Amerikában a legjelentősebb adatbázisok közé tartozik a TIMIT nevű korpusz, amit ugyancsak gépi beszédfelismerés céljából hoztak létre [19]. Akusztikai modellek felépítésére alkalmas amerikai angol nyelvre. Az adatbázis személyfüggetlen fonetikai beszédfelismerők betanítására és tesztelésére szolgál, szómodellek felépítésére alkalmatlan, mivel szűkített szótárkészletet használ. Vannak viszont fonetikailag gazdag mondatai, amelyek kiválóan alkalmasak akusztikai (beszédhang) modellek létrehozására. Így az adatbázis egy része betanításra, a másik része tesztelésre ad lehetőséget. A mintavételi frekvencia itt is 16 kHz (16 biten), a felvételeket csendes szobában rögzítették, a bemondás módja pedig felolvasás. Az adatbázis beszédhang szinten szegmentált [24].

3.1.3 KIEL

A KIEL jelenleg a világ egyik legnagyobb adatbázisa (és folyamatosan bővül), ami rendelkezésre áll egy adott nyelvből [20]. A korpusz nyelve a német. Felolvasott szöveg mellett kvázi spontán dialógusokat is tartalmaz. Az adatbázis kiválóan alkalmazható prozódiai vizsgálatokhoz. A rendelkezésemre álló felvételeken 20 beszélő hangja szerepel. A mintavételi frekvencia 16 kHz (16 biten), az adatbázis beszédhang szinten szegmentált [24].

4 Az előfeldolgozási eljárások ismertetése

4.1.1 Az alkalmazott pszichoakusztikus skálák

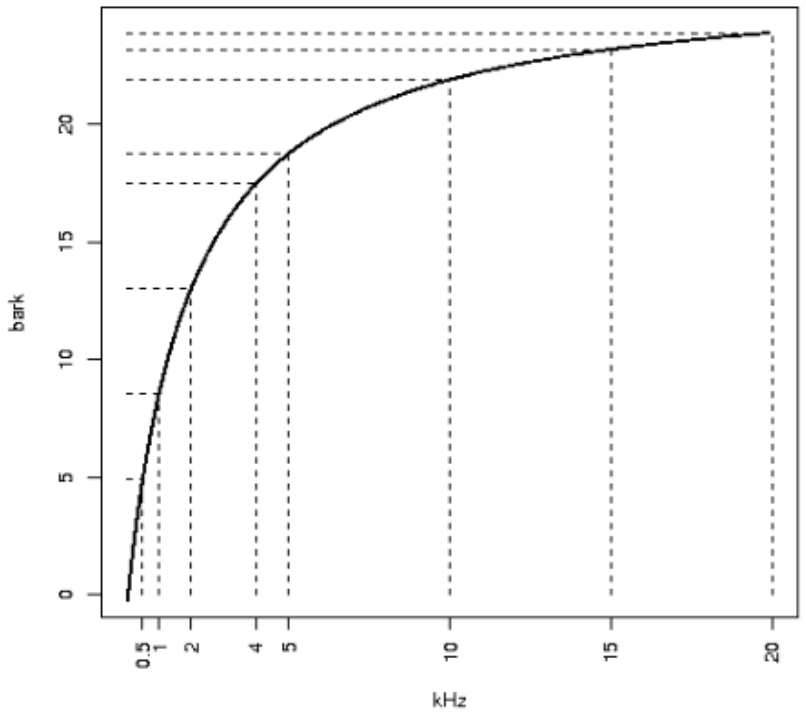
Az előfeldolgozási lépések során két, az emberi halláshoz igazodó spektrális sávfelbontást alkalmazok. Az egyik eljárás a színkép energiájának abszolút értékeivel dolgozik, míg a másik azoknak frekvenciabeli és időbeli deriváltjaival. A két eltérő előfeldolgozási lépés azért volt szükségszerű, mert a rendelkezésemre álló eszközök a két eltérő ábrázolásmódot (abszolútérték, illetve derivált) két eltérő frekvenciabeli sávfelbontással valósítják meg.

Az emberi fül átviteli karakterisztikája jól modellezhető egy sávszűrőkből összeállított szűrősorral [21]. Az első esetben a szűrősor alapját a mel-skála szolgálja, ami tulajdonképpen átléptékezi a frekvenciában azonos távolságra lévő hangokat az ember észlelése alapján azonos távolságra lévő hangokká. Tehát a hangok egyenletes hangmagasságérzetét fejezi ki. A skála kiindulási alapjaként egy 1 kHz-es, 40 dB-es hang szolgál, amelynek érzékelt magasságát 1000 mel-ben állapították meg [17]. A skálát úgy állították be, hogy az értékek megduplázása a hangmagasságérzetet is megduplázza, és a 131 Hz-hez 131 mel feleljen meg. Ekkor a 0-16 kHz közötti frekvenciatartományt a 0-2400 mel értéksorral jellemzik. Ha frekvenciát akarunk *mel* – be konvertálni, azt a következő képlettel lehetséges:

$$\mathrm{Mel(f)} = 2595 \log_{10}\left(1 + \frac{f}{700}\right).$$

Az emberi fül felbontása a frekvenciatartományban az úgynevezett kritikus sávokkal írható le [23], ez képezi a második eset alapját. Ezek a kritikus sávok megfeleltethetőek a csiga frekvenciafelbontó képességének, és fontos szerepet játszanak a percepcióban: ha a fülünket egyszerre több hang éri, és ezek a kritikus sávon belül vannak, akkor az intenzitásuk egyszerűen összeadódik és nem érzékeljük őket különálló hangokként. A kritikus sávok frekvenciahatárai kimérhetők, a határokhoz érve a hangosságérzet megnő. A sávok szélessége függ a sávközép-frekvenciától. Az emberi hallásra 24 kritikus sáv jellemző, a 20 és 15500 Hz közötti hallástartományban, ezek adják az úgynevezett Bark szűrők sorszámát. Így a kritikus sávok szolgálnak a Zwincker által 1961-ben kialakított, Heinrich Barkhausen tiszteletése

elnevezett Bark-skála alapjául. A sávszélesség 500 Hz alatt közel állandó, 100 Hz, e fölött egyre nagyobb értékeket vesznek fel. A 3. ábra szemlélteti, hogy az *x*-tengelyen jelölt értékek között egyre nagyobb a távolság, míg a nekik megfelelő Bark-értékek között nagyjából azonos a különbség.



3. ábra: A frekvenciaérték (kHz) és a fül által észlelt hangmagasság (Bark) függvénye

Egy bark értéke megközelítőleg egyenlő száz mellel. A konverziót frekvenciáról (Hz) bark-ba a következő képlet adja meg [23]:

$$Bark = 13 \cdot \arctan(0{,}00076\, f) + 3{,}5 \cdot \arctan((\frac{f}{7500})^2) .$$

4.2.1 Előfeldolgozás szűrősávokban mért energiával

Mintafelvételek segítségével olyan vektorhalmazokat hoztam létre, amelyeket az SVM betanításához, majd teszteléséhez használtam. Egy tulajdonságvektor egy 25 ms hosszúságú keretet jellemez. A tulajdonságvektorok kiszámítása 10 ms-onként történt.

A tulajdonságvektorunk ebben az esetben az MFCC értékek (Mel-frequency cepstral coefficients). Az MFC-együtthatók kiszámítása tulajdonképpen egy lényegkiemelési eljárás, amely során a beszédjelből, az információtartalom szempontjából releváns paramétereket kiemeljük, lényeges információt nem hordozó (vagy redundáns) részeket eldobjuk.

Az MFC - együtthatókat a következőképpen kapjuk meg:

- A beszédjelet rövid, ablakozott szeletekre osztjuk
- Az adott szakaszt Fourier-transzformáljuk (FFT)
- Meghatározzuk a teljesítményspektrumot
- Elvégezzük a szűrősoros elemzést, azaz az összetevőket mel sávok szerint összegezzük
- Logaritmáljuk a teljesítményeket mindegyik mel-frekvenciában
- Végül a mel-frekvenciás logaritmikus teljesítményeket diszkrét-koszinusztranszformáljuk a következő képlet alapján

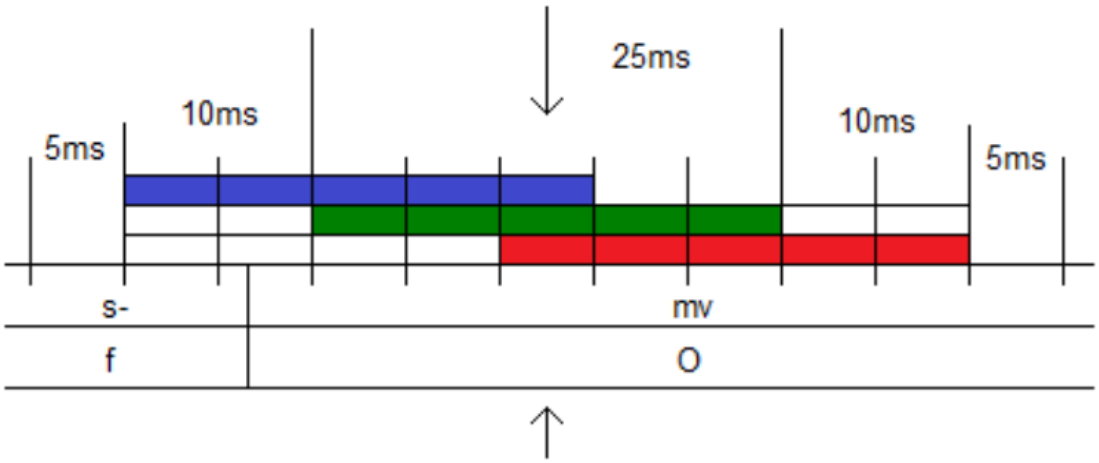
$$c_i = \sum_{j=1}^{N} P_j \cos(i\pi / N(j - 0.5)),$$

ahol c_i a kapott MFC *i*-edik együtthatónk, N a szűrők száma (beszédtechnológiában ez gyakran 24 darab), P pedig a *j*-edik szűrő teljesítménye.

Az MFC-együtthatókat természetesen gördülő jelleggel érdemes kiszámolni, a gördülő spektrumot ablakszélességnyi keretenként konvolváljuk a hallást modellező szűrősor átviteli függvényével (a számítási ablakméret: 150 ms), ennek eredményeképpen fix időközönként egy-egy 24 dimenziós tulajdonságvektort kapunk. Ha a beszédjelet 16 kHz-en mintavételezzük, akkor az utolsó 4 kritikus sávot figyelmen kívül hagyhatjuk, így 20

szűrőkimenetünk lesz. A diszkrét-koszinusztranszformáció révén 12 dimenziós jellemzővektorokat kapunk, e vektorok számelemei az MFC együtthatók.

Egy tulajdonságvektor állhat csupán a közvetlenül hozzá tartozó keret értéksorozatából, vagy ezen kívül hozzávehetjük a tőle időben balra és jobbra elhelyezkedő 2 vagy 4 db szomszéd keret értékeit is. Ezek mind egy 25 ms-es keretet jellemeznek. Egy ilyen, 2 szomszéd keretét is felhasználó tulajdonságvektor szemléltetése a 4. ábrán tekinthető meg.



4. ábra: Egy tulajdonságvektor előállításának szemléltetése

A program által használt Praat scriptben úgy állítottam be a paramétereket, hogy a mel-filter skálája a 100-as mel értéktől induljon, és 100-mel távolságonként vegyen fel egy értéket. Így a 16kHz-es felvételeinkhez a Praat 27 db mel értéket tud kiszámolni. Viszont, ha az MFCC értékeket szeretnék használni, akkor a 27 db paraméter mindössze 12 db-ra redukálódik.

Tehát a szűrősávokban mért energiával történő akusztikai előfeldolgozás során MFCC értékek lettek számolva, a fentebb leírt 10 ms-os lépésközzel és 25 ms hosszúságú ablakkal, 4 szomszédos keret figyelembevételével. Következésképpen ebben az előfeldolgozási eljárásban 5x12 darab, azaz 60 elemű lesz a tulajdonságvektorunk.

4.2.2 Előfeldolgozás frekvencia és időbeli deriválással

Az előfeldolgozás során a jellemzővektorunkat a következő módon számoljuk:

- A beszédjelet rövid, ablakozott szeletekre osztjuk
- Az adott szakaszt Fourier-transzformáljuk (FFT)
- Meghatározzuk a teljesítménypektrumot
- Elvégezzük a Bark-szűrősoros elemzést (18 kritikus sávra)
- A különböző sávok energiáit sávonként összegezzük
- A nyers bark energia értékekből frekvenciabeli deriváltakat képzünk $(E_k + 1) - E_k$ összefüggés nyomán (ahol a k a bark-szűrő sorszáma, *E* pedig az ebben mért energia)
- Meghatározzuk ezen értékeken időbeli deriváltjait $E_{t+1} - E_t$ képlet alapján
- Végül kiszámoljuk a második időbeli deriváltakat is $E_{t+1} - E_{t-1}$ formula szerint.

Mivel 18 sávszűrőt használunk így egy keretre 54 elemű jellemzővektort kapunk (18 frekcenciabeli derivált + 18 időbeli derivált + 18 második időbeli derivált). Viszont az előző előfeldolgozáshoz hasonlóan figyelembe vesszük az aktuális keretünk 4 szomszédos keretében kapott értékeit is, így a tulajdonságvektorunk teljes elemszáma 5x54, azaz 270-re adódik.

Tehát egy nagy különbség a két előfeldolgozási eljárás között, hogy míg az első esetben pillanatnyi abszolútértékeket számolunk, második esetben idő és frekvenciabeli deriváltak képzik a bemenő vektorokat.

5. Az előfeldolgozási eljárások összehasonlítása

Az SVM-al minden 10 ms-ban kapunk egy döntést. Ez eredmény bemenetként szolgál két utólagos simító algoritmushoz, ami szegmentálást állít elő, azaz bejelöli a fonémacsoport határokat. Szeretnénk megmérni az SVM felismerési pontosságát, még a simítás előtt, azért, hogy összehasonlítsuk a két előfeldolgozás hatását. Az eredményt tévesztési mátrixban adtam meg. A mátrix azt mutatja, hogy egy adott osztály tesztanyaga (függőleges tengely, sorok) milyen százalékban sorolódik be az egyes fonémaosztályokba (vízszintes tengely, oszlopok).

Tetszőleges számú osztályra fel lehet állítani a tévesztési mátrixot, melyben az általunk várt (valódi) és megfigyelt (modell által jósolt) esetek számát regisztráljuk. Egy adott csoport szempontjából, pozitívak azok az esetek, amelyek az adott kiszemelt osztályra vonatkoznak, negatívok azok, amelyek más osztályokra vonatkoznak. Ennek megfelelően négy féle módon számolhatjuk össze az osztályozott eseteket:

- TP (True Positive): azon pozitív minták száma, melyeket a modell helyesen azonosított
- FP (False Positive): azon pozitív minták száma, melyeket a modell rosszul azonosított
- FN (False Negative): azon negatív minták száma, melyeket a modell rosszul azonosított
- TN (True Negative): azon negatív minták száma, melyeket a modell helyesen azonosított

A tévesztési mátrix értékei alapján különböző mérőszámok segítségével lehet kiértékelni az osztályozási eredményünket, a felismerési pontosságot a következő összefüggés adja meg:

$$\frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}$$

Továbbá, a mátrix egy sorában szereplő értékek összege 1 kell, hogy legyen. Ha ez nem teljesül, annak egy magyarázata van, az, hogy az értékek számolásakor kerekít a program, amit a két tizedesnyi pontosság nem ad ki. A kiértékelés során SVM minden egyes 10 ms-os keretének címkézését összehasonlítjuk a referencia címkézéssel. Tehát az itt kapott értékek még nem tartalmaznak információt a fonémahatárok detektálásának pontosságáról, csupán az

osztályozás precizitásáról. A 4-es sorszámú táblázat szemlélteti a magyar MRBA adatbázison kapott eredményeket a szűrősávokban mért energiával végzett előfeldolgozással.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.77**	0.04	0.02	0.01	0.05	0.02	0.02	0.04	0.03	0.01
hv	0.09	**0.68**	0.02	0.01	0.05	0.03	0.02	0.08	0.03	0.00
s+	0.08	0.05	**0.47**	0.09	0.08	0.08	0.02	0.08	0.03	0.02
s-	0.04	0.01	0.05	**0.69**	0.04	0.01	0.03	0.00	0.11	0.03
na	0.23	0.12	0.03	0.01	**0.35**	0.12	0.03	0.07	0.02	0.02
b+	0.02	0.05	0.03	0.00	0.07	**0.66**	0.09	0.05	0.00	0.04
b-	0.01	0.01	0.01	0.02	0.01	0.03	**0.59**	0.00	0.05	0.27
vo	0.03	0.06	0.02	0.01	0.02	0.08	0.03	**0.64**	0.08	0.02
uv	0.00	0.00	0.01	0.13	0.00	0.00	0.08	0.01	**0.70**	0.06
si	0.01	0.00	0.00	0.01	0.00	0.01	0.06	0.00	0.05	**0.85**

4. táblázat: SVM osztályozás eredménye MRBA adatbázison szűrősávokban mért energiával végzett előfeldolgozással

Az összesített találati arány pedig 69%-ra adódott (jó minták/összes minta). Ezt az eredmény összevethető a bark-sávos eljárás eredményével. Ezt az 5. táblázat szemlélteti. Az MRBA adatbázison az összesített felismerés 73%.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.88**	0.03	0.00	0.01	0.04	0.00	0.01	0.00	0.00	0.01
hv	0.17	**0.70**	0.00	0.01	0.09	0.00	0.01	0.00	0.01	0.01
s+	0.15	0.08	**0.32**	0.11	0.22	0.05	0.02	0.00	0.02	0.03
s-	0.07	0.01	0.01	**0.75**	0.02	0.00	0.02	0.00	0.05	0.06
na	0.30	0.11	0.00	0.01	**0.50**	0.02	0.02	0.00	0.01	0.03
b+	0.04	0.04	0.00	0.00	0.32	**0.44**	0.07	0.02	0.01	0.05
b-	0.02	0.01	0.00	0.02	0.02	0.00	**0.57**	0.00	0.06	0.30
vo	0.27	0.20	0.01	0.01	0.25	0.04	0.01	**0.09**	0.06	0.06
uv	0.09	0.02	0.01	0.21	0.02	0.00	0.05	0.00	**0.44**	0.16
si	0.02	0.00	0.00	0.01	0.01	0.00	0.04	0.00	0.01	**0.90**

5. táblázat: SVM osztályozás eredménye MRBA adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

A második előfeldolgozási eljárással nemcsak az összesített találati arány növekedett (69%-ról 73%-ra) hanem, a magas (68%-ról 70%-re) és mély magánhangzók találati aránya is (77%-ról 88%-ra). A zöngés felpattanás osztálynál viszont a mel-sávos eset bizonyul sokkal jobbnak 64%-os találattal a bark-sávos 9%-os eredményéhez képest. Viszont ha olyan alkalmazást fejlesztünk ahol a vizsgált jelenség függ a beszéd időzítésétől, mint például a ritmus mérése, akkor ez lényegtelen, hiszen a magánhangzók detektálása a fő cél.

Továbbá, ugyanezekkel az előfeldolgozási eljárásokkal a további két adatbázison kapott tévesztési mátrixokat a következő két táblázatok tartalmazzák. A 6-os táblázat az angol TIMIT-re mel-sávos esetben, az 7-es pedig ugyanezen az adatbázison bark-sávosra.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.84**	0.05	0.01	0.01	0.03	0.00	0.01	0.01	0.02	0.00
hv	0.38	**0.42**	0.02	0.02	0.05	0.02	0.01	0.05	0.03	0.00
s+	0.23	0.01	**0.10**	0.30	0.03	0.07	0.06	0.06	0.12	0.02
s-	0.02	0.01	0.01	**0.66**	0.00	0.00	0.03	0.00	0.24	0.02
na	0.47	0.02	0.04	0.02	**0.17**	0.11	0.05	0.06	0.04	0.01
b+	0.02	0.01	0.01	0.01	0.04	**0.36**	0.31	0.05	0.05	0.13
b-	0.02	0.01	0.00	0.04	0.01	0.03	**0.47**	0.00	0.12	0.29
vo	0.07	0.04	0.04	0.03	0.07	0.04	0.07	**0.17**	0.45	0.04
uv	0.11	0.02	0.01	0.16	0.00	0.01	0.09	0.01	**0.54**	0.06
si	0.04	0.01	0.00	0.03	0.01	0.01	0.06	0.00	0.09	**0.75**

6. táblázat: SVM osztályozás eredménye TIMIT adatbázison szűrősávokban mért energiával végzett előfeldolgozással

A TIMIT adatbázisra az összesített találati arány 55% -ra adódott első esetben. Ha az 6-os táblázat magánhangzók felismerését vizsgáljuk, azt lehet észrevenni, hogy legnagyobb mértékben a mély magánhangzókat magasra téveszti és fordítva. Ugyanezen az adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással az összesített találati arány 57%, a részeredmények bizakodásra adnak okot (7. táblázat).

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.90**	0.05	0.00	0.01	0.02	0.00	0.01	0.00	0.01	0.00
hv	0.50	**0.39**	0.00	0.02	0.06	0.01	0.01	0.00	0.01	0.01
s+	0.33	0.01	**0.05**	0.38	0.08	0.02	0.05	0.00	0.04	0.03
s-	0.04	0.01	0.00	**0.82**	0.02	0.00	0.03	0.00	0.06	0.03
na	0.57	0.03	0.00	0.03	**0.27**	0.02	0.03	0.00	0.02	0.03
b+	0.11	0.02	0.00	0.03	0.09	**0.21**	0.41	0.01	0.02	0.10
b-	0.08	0.01	0.00	0.06	0.01	0.01	**0.60**	0.00	0.05	0.17
vo	0.15	0.04	0.02	0.07	0.12	0.04	0.10	**0.06**	0.34	0.05
uv	0.17	0.02	0.00	0.29	0.01	0.00	0.11	0.00	**0.32**	0.07
si	0.04	0.00	0.00	0.05	0.02	0.01	0.17	0.00	0.02	**0.69**

7. táblázat: SVM osztályozás eredménye TIMIT adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

A mély magánhangzók találata 84%-ról 90%-ra ugrott, a magasaké, viszont 3%-al csökkent. Ha figyelembe vesszük, hogy esetleges jövőbeni munkában az angol magánhangzókat pontosabban tudjuk besorolni az akusztikai osztályokba, a két magánhangzó csoport tévesztése csökkenthető, mindkettőben 90% körüli (vagy a feletti) értéket kaphatunk. Ámbár, ha csak a magánhangzók detektálása a cél és nem érdekel minket, hogy milyen magánhangzó, (tehát összevonjuk a két csoportot) akkor ezt már ezzel az eljárással is 92%-os biztonsággal meg tudjuk határozni ([(90+5)+(50+39)]/2).

Végül a KIEL korpuszon szintén 55%-ra adódott az összesített találati ráta szűrősávokban mért energiával történő akusztikai előfeldolgozás során (8. táblázat).

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.56**	0.10	0.03	0.00	0.11	0.06	0.01	0.09	0.03	0.02
hv	0.19	**0.54**	0.01	0.00	0.09	0.05	0.00	0.09	0.02	0.01
s+	0.00	0.02	**0.32**	0.09	0.02	0.20	0.02	0.13	0.15	0.05
s-	0.02	0.03	0.08	**0.53**	0.03	0.04	0.04	0.03	0.20	0.02
na	0.02	0.08	0.02	0.02	**0.35**	0.31	0.02	0.08	0.03	0.06
b+	0.01	0.03	0.00	0.02	0.05	**0.43**	0.14	0.06	0.06	0.21
b-	0.00	0.01	0.00	0.05	0.02	0.29	**0.35**	0.03	0.06	0.18
vo	0.01	0.02	0.01	0.01	0.00	0.02	0.01	**0.32**	0.59	0.01
uv	0.01	0.01	0.02	0.14	0.01	0.02	0.02	0.07	**0.62**	0.09
si	0.01	0.00	0.00	0.00	0.01	0.04	0.02	0.01	0.03	**0.87**

8. táblázat: SVM osztályozás eredménye KIEL adatbázison szűrősávokban mért energiával végzett előfeldolgozással

Összevetve ezt a bark-sávos előfeldolgozással azt vesszük észre, hogy itt is növekedett az összesített találati arány (65%). Mély magánhangzók esetében 14%-ot növekedett a találat, magas magánhangzók esetében viszont 2%-ot csökkent (9. táblázat). Zöngés felpattanás esetben itt is és az angol adatállományon is drasztikus csökkenést észlelhetünk.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.70**	0.10	0.00	0.00	0.15	0.01	0.01	0.00	0.01	0.02
hv	0.26	**0.52**	0.00	0.00	0.17	0.01	0.00	0.00	0.01	0.01
s+	0.04	0.04	**0.32**	0.17	0.13	0.15	0.02	0.00	0.08	0.04
s-	0.05	0.03	0.02	**0.67**	0.05	0.01	0.04	0.00	0.11	0.03
na	0.08	0.12	0.01	0.03	**0.60**	0.06	0.02	0.00	0.02	0.06
b+	0.03	0.02	0.00	0.03	0.17	**0.27**	0.32	0.02	0.07	0.08
b-	0.06	0.02	0.00	0.06	0.07	0.06	**0.56**	0.01	0.08	0.10
vo	0.09	0.08	0.01	0.03	0.10	0.03	0.010	**0.08**	0.54	0.02
uv	0.07	0.04	0.00	0.19	0.04	0.01	0.04	0.01	**0.49**	0.12
si	0.02	0.00	0.00	0.00	0.01	0.01	0.08	0.00	0.01	**0.87**

9. táblázat: SVM osztályozás eredménye KIEL adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

Az eredményeket a 10-es táblázatban foglaltam össze.

	Adatbázis	Előfeldolgozás szűrősávokban mért energiával	Előfeldolgozás frekvencia és időbeli deriválással
Összesített találati arány	MRBA	69%	73%
	TIMIT	55%	57%
	KIEL	55%	65%
Magánhangzó találati arány	MRBA	79%	89%
	TIMIT	84.5%	92%
	KIEL	69.5%	79%

10. táblázat: SVM osztályozás eredményének összefoglalása

6 Simító eljárások

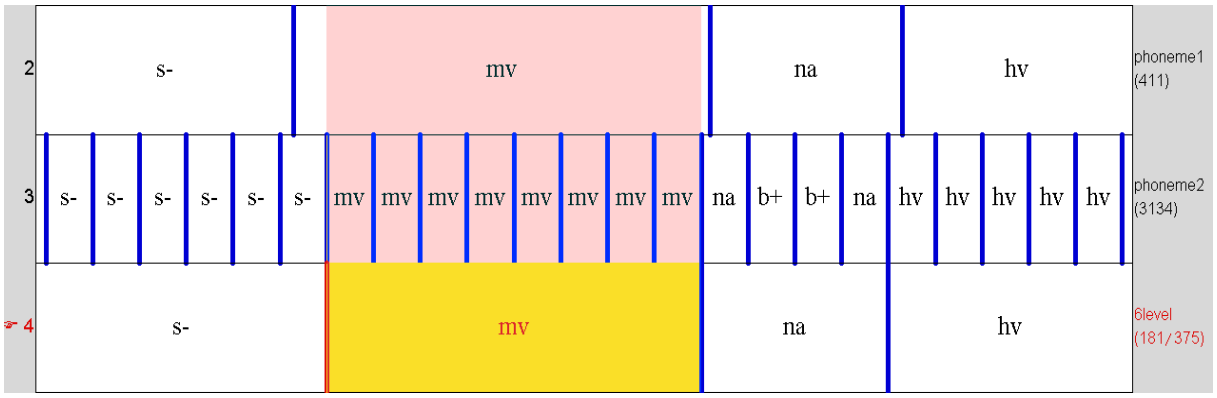
6.1 Szabálybázisú simító eljárás

Az SVM osztályozása egy, az akusztikai csoportoknak megfelelő címkézés. A szabálybázisú simító eljárás gyakorlatilag egy döntési logika implementálása az SVM osztályozás eredményeire, hogy fonémacsoport szintű szegmentálást valósítson meg, ami a következő algoritmust követ:

- Három egymást követő, ugyanolyan címkét viselő egység egy potenciális *hangszakaszt* alkot
- Ha egy 10 ms-os egység két hosszabb, különböző címkéjű hangszakasz között helyezkedik el, akkor összeolvad valamelyik szomszédjával
- A zárszakaszok zöngés vagy zöngétlennek lettek címkézve a többségi szavazás szerint
- A további egységek a hangszakaszokkal lettek egyesítve többségi szavazás szerint
- Azok a zárszakaszok, amelyeket nem követett felpattanó címke, ki lettek egészítve felpattanó címkével
- A magyar nyelv időtartambeli tulajdonságainak megfelelő módosítások elvégzése; például egy zárszakasz nem lehet 1000 ms.

Ezek a lépések hierarchikusak, ami annyit jelent, hogy a legalapvetőbb tulajdonságoktól a legspecifikusabb felé haladunk fontossági sorrendben.

Az 5. ábrában látható az SVM által 10 ms-os keretenkénti osztályozás (phoneme2 nevű sor), alatta a szabálybázisú eljárás eredménye, a felső sor pedig a kézileg elvégzett referencia szegmentálás.



5. ábra: Szabálybázisú hibajavítás szemléltetése Praat-ban

A vizsgálatot továbbá kiegészítem egy szegmentálás pontosságát mérő eljárással. Az eljárás megkeresi a referencia sorban a fonémahatárokat és egy állítható időablakban megnézi, hogy az összehasonlítandó sorban történik-e ebben az ablakban (a referencia fonémahatár két szomszédjában) fonémahatár beszúrás, avagy sem. Ebből százalékos értéket számít. A szegmentálási pontosságot magánhangzókra is külön megnéztem. Az vizsgálati időablakot 20 ms-nak választottam. Fontos továbbá megemlíteni, hogy a szegmentálási pontosságot nem fájlonként számoljuk, hanem azokat sorba illesztve, egyként kezeljük.

A következőkben bemutatom a szabálybázisú hibajavítás eredményeit a két előfeldolgozási eljárásra. A szabálybázisú módszer kimenete már ad információt a fonémahatárok detektálásának pontosságára is, ezt az eredményt is bemutatom. A kiértékelés konkrétan azt mondja meg, hogy hány százalékban illeszkedik a kimenet a referencia címkézésre. A 11-es táblázat mutatja mel-sávos esetben a tévesztési mátrixot az MRBA korpuszra.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.78**	0.03	0.02	0.01	0.03	0.01	0.02	0.03	0.04	0.02
hv	0.09	**0.69**	0.02	0.01	0.05	0.01	0.02	0.06	0.04	0.02
s+	0.10	0.05	**0.47**	0.09	0.07	0.08	0.02	0.06	0.04	0.02
s-	0.06	0.02	0.04	**0.71**	0.01	0.00	0.04	0.01	0.11	0.02
na	0.26	0.13	0.02	0.02	**0.32**	0.07	0.06	0.06	0.02	0.04
b+	0.03	0.05	0.02	0.00	0.06	**0.65**	0.09	0.05	0.00	0.04
b-	0.01	0.01	0.01	0.02	0.00	0.02	**0.66**	0.00	0.05	0.21
vo	0.04	0.06	0.02	0.00	0.03	0.09	0.03	**0.62**	0.08	0.02
uv	0.01	0.00	0.01	0.12	0.00	0.01	0.10	0.02	**0.71**	0.04
si	0.01	0.00	0.00	0.01	0.00	0.00	0.25	0.00	0.06	**0.65**

11. táblázat: Szabálybázisú simító eljárás eredménye MRBA adatbázison szűrősávokban mért energiával végzett előfeldolgozással

Mel-sávos előfeldolgozás esetén az összesített felismerés 65%, a fonémahatárok szegmentálási pontossága pedig: 79,75%. Ez az érték igazából akkor izgalmas, ha összevetjük ugyanezen adatbázis másik eljárás kimenetével (12. táblázat).

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.89**	0.03	0.00	0.01	0.04	0.00	0.01	0.00	0.00	0.02
hv	0.16	**0.71**	0.00	0.01	0.07	0.00	0.01	0.00	0.01	0.01
s+	0.19	0.10	**0.33**	0.11	0.20	0.01	0.01	0.00	0.02	0.03
s-	0.08	0.02	0.01	**0.78**	0.02	0.00	0.02	0.00	0.03	0.05
na	0.33	0.11	0.00	0.02	**0.48**	0.01	0.02	0.00	0.01	0.03
b+	0.15	0.11	0.01	0.00	0.38	**0.20**	0.06	0.02	0.01	0.07
b-	0.04	0.01	0.00	0.05	0.02	0.00	**0.54**	0.00	0.06	0.28
vo	0.32	0.21	0.01	0.00	0.21	0.02	0.01	**0.11**	0.05	0.05
uv	0.10	0.03	0.01	0.19	0.02	0.00	0.06	0.00	**0.46**	0.14
si	0.02	0.00	0.00	0.01	0.01	0.00	0.06	0.00	0.01	**0.87**

12. táblázat: Szabálybázisú simító eljárás eredménye MRBA adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

Az utóbbi összesített találati aránya 72%, 77,41% szegmentálási pontosság mellett. Így azt a következtetést tudjuk levonni, hogy összességben a bark-sávos megoldás vezetett jobb eredményre, még akkor is, ha a szegmentálási pontosságban 2 %-al elmarad. A

részeredmények is erről tanúskodnak. Számunkra fontos szempont, hogy nyelvtől függetlenül tudjuk minél pontosabban detektálni a magánhangzókat, bark-sávos módszer esetén, a magyar korpuszon 89%-ban tudjuk a mély magánhangzókat és 71%-ban a magasakat, szemben a mel-sávos előfeldolgozás esetén 78 és 69%-al.

A nyelvfüggetlenség eredményességéről akkor tudunk nyilatkozni, ha az idegen nyelvű korpuszokon is jól szerepel a technológia, azaz összevethető a magyar nyelvű eredményeinkkel. A 12-es táblázat ábrázolja az angol adatbázisra kapott eredményt, szűrősávokban mért energiával végzett előfeldolgozás esetén.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.85**	0.05	0.01	0.01	0.03	0.00	0.01	0.01	0.02	0.01
hv	0.38	**0.42**	0.02	0.02	0.04	0.01	0.02	0.04	0.04	0.01
s+	0.25	0.03	**0.09**	0.31	0.03	0.06	0.05	0.06	0.11	0.02
s-	0.03	0.02	0.00	**0.66**	0.00	0.00	0.03	0.01	0.23	0.01
na	0.50	0.03	0.04	0.02	**0.16**	0.07	0.06	0.05	0.05	0.02
b+	0.03	0.01	0.01	0.02	0.02	**0.31**	0.44	0.04	0.05	0.06
b-	0.03	0.01	0.00	0.04	0.00	0.01	**0.68**	0.00	0.13	0.08
vo	0.10	0.04	0.03	0.02	0.05	0.05	0.06	**0.18**	0.45	0.02
uv	0.12	0.02	0.00	0.14	0.00	0.01	0.10	0.01	**0.57**	0.02
si	0.04	0.01	0.01	0.03	0.01	0.01	0.30	0.01	0.10	**0.50**

12. táblázat: Szabálybázisú simító eljárás eredménye TIMIT adatbázison szűrősávokban mért energiával végzett előfeldolgozással

Az összesített találati ráta 52%, ami 4%-al elmarad a bark-sávos megoldástól (13. táblázat).

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.91**	0.05	0.00	0.01	0.01	0.00	0.00	0.00	0.01	0.00
hv	0.50	**0.40**	0.00	0.03	0.05	0.00	0.00	0.00	0.01	0.01
s+	0.37	0.02	**0.04**	0.40	0.08	0.00	0.02	0.00	0.03	0.03
s-	0.05	0.01	0.00	**0.85**	0.01	0.00	0.02	0.00	0.04	0.02
na	0.60	0.03	0.00	0.04	**0.25**	0.00	0.02	0.00	0.02	0.03
b+	0.16	0.06	0.00	0.06	0.07	**0.17**	0.39	0.01	0.02	0.09
b-	0.09	0.02	0.00	0.12	0.01	0.01	**0.55**	0.00	0.05	0.14
vo	0.20	0.06	0.01	0.06	0.10	0.03	0.09	**0.06**	0.34	0.04
uv	0.19	0.02	0.00	0.28	0.01	0.00	0.10	0.00	**0.34**	0.05
si	0.06	0.02	0.00	0.05	0.01	0.00	0.19	0.00	0.02	**0.64**

13. táblázat: Szabálybázisú simító eljárás eredménye TIMIT adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

A szegmentálási pontosság: 75,38% (mel-sávos esetre) és 72,27% (bark-sávos esetre).

A kimenetet látva és összevetve a magyar korpusszal az összesített felismerést, első ránézésre kudarcélményünk lehet, viszont ha a részeredményeket nézzük, akkor világossá válik a siker. A magánhangzókat mel és bark-sávos esetben is jól elkülöníti a többi akusztikai osztálytól, bark-sávos esetben jobban.

A német adatbázison kapott eredményt első esetben a 14. táblázatban van összefoglalva.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.57**	0.10	0.02	0.00	0.11	0.03	0.02	0.08	0.04	0.03
hv	0.19	**0.54**	0.01	0.00	0.08	0.04	0.01	0.08	0.03	0.01
s+	0.01	0.04	**0.33**	0.09	0.02	0.16	0.04	0.15	0.13	0.03
s-	0.03	0.04	0.05	**0.55**	0.03	0.02	0.05	0.04	0.18	0.01
na	0.04	0.08	0.02	0.03	**0.36**	0.19	0.08	0.08	0.03	0.09
b+	0.01	0.03	0.00	0.02	0.04	**0.40**	0.29	0.06	0.06	0.07
b-	0.01	0.02	0.00	0.06	0.02	0.17	**0.52**	0.03	0.07	0.11
vo	0.01	0.03	0.00	0.01	0.00	0.02	0.01	**0.36**	0.55	0.01
uv	0.02	0.02	0.00	0.10	0.01	0.02	0.06	0.16	**0.57**	0.05
si	0.01	0.00	0.00	0.00	0.01	0.01	0.27	0.01	0.04	**0.65**

14. táblázat: Szabálybázisú simító eljárás eredménye KIEL adatbázison szűrősávokban mért energiával végzett előfeldolgozással

Az összesített felismerés 52%, hasonlóan, mint az angol korpuszon, viszont részeredményekben sokkal gyengébb, mint a magyar vagy az angol eredmény, szűrősávokban mért energiával végzett előfeldolgozással. A mély magánhangzókat 57%-ban, a magasakat 54%-ban ismeri fel, viszont az utóbbi esetben 19%-ban mély magánhangzóra téveszt. A szegmentálási pontosság: 79,24%. bark-sávos esetre a tévesztési mátrixot a 15-es táblázat írja le.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.71**	0.10	0.00	0.01	0.15	0.00	0.00	0.00	0.01	0.02
hv	0.26	**0.53**	0.00	0.01	0.17	0.00	0.00	0.00	0.01	0.01
s+	0.10	0.08	**0.33**	0.17	0.16	0.02	0.03	0.01	0.07	0.03
s-	0.06	0.04	0.01	**0.71**	0.05	0.00	0.02	0.00	0.09	0.02
na	0.08	0.13	0.01	0.04	**0.63**	0.01	0.01	0.01	0.02	0.07
b+	0.06	0.06	0.00	0.02	0.20	**0.15**	0.35	0.02	0.07	0.06
b-	0.07	0.03	0.00	0.12	0.06	0.02	**0.51**	0.01	0.08	0.09
vo	0.11	0.10	0.00	0.02	0.08	0.02	0.02	**0.11**	0.54	0.01
uv	0.07	0.04	0.00	0.16	0.04	0.00	0.05	0.02	**0.51**	0.11
si	0.02	0.01	0.00	0.00	0.01	0.00	0.09	0.00	0.01	**0.85**

15. áblázat: Szabálybázisú simító eljárás eredménye KIEL adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

Ebben az esetben az összesített találati arány 65%-ra adódik és a részeredmények is jobbak, mint a mel-sávos párjánál. A 80%-os magánhangzó detektálást ugyan nem éri el, viszont mély magánhangzók esetén a 71%-os egyezés, nem számít annyira rossznak, főleg ha figyelembe vesszük, hogy tizedét magas magánhangzóra döntötte. A szegmentálási pontosság 77,98%-ra adódott, ami nem sokban marad el mel-sávos párjától.

Az eredményeket összefoglaltam a 16-os táblázatban. A magánhangzó pontosság megadásakor összevontam az mv és hv csoportokat.

	Adatbázis	Előfeldolgozás szűrősávokban mért energiával	Előfeldolgozás frekvencia és időbeli deriválással
Összesített találati arány	MRBA	65%	72%
	TIMIT	52%	56%
	KIEL	52%	65%
Magánhangzó találati arány	MRBA	79.5%	89.5%
	TIMIT	85%	93%
	KIEL	70%	80%
Összesített szegmentálási pontosság	MRBA	79.75%	77.41%
	TIMIT	75.38%	72.27%
	KIEL	79.24%	77.98%
Magánhangzó szegmentálási pontosság	MRBA	78.05%	77.68%
	TIMIT	70.18%	67.66%
	KIEL	80.91%	78.82%

16. táblázat: Szabálybázisú simító eljárás összesített eredményei

Arra a következtetésre juthatunk, hogy a barl-sávos előfeldolgozással számolt jellemzővektorok javították az SVM osztályozását és a szabálybázisú simító eljárás után is kedvező eredményt kaptunk. A szegmentálási pontosság mel-sávos esetben jobbnak bizonyult úgy összesített esetben, mint a magánhangzókra nézve, azonban ez a különbség nem számottevő. Az akusztikai osztályok felismerésének nyelvfüggetlenség jelentős, elsősorban olyan alkalmazásoknál használható az eljárás ahol vizsgált jelenség függ a beszéd időzítésétől, amilyenek a prozódiai vizsgálatok, avagy a ritmus mérése.

6.2 Statisztikai alapú simító eljárás

A statisztikai alapú simító eljárás esetén az SVM kimenete más lesz, mint szabálybázisú esetében. Az, hogy egy adott állapotból (hangszakaszból) helyben maradunk vagy a következő állapotba (felismert fonémacsoportba) lépünk, a következő, minden 10 ms-ban kiértékelt egyenlőtlenség határozza meg:

$$P(o_t \mid f_{i\max}) \cdot b_{j,i_{\max}} > P_i(o_t \mid f_i) \cdot a_j .$$

Ha az egyenlőség fennáll, akkor az adott időpillanatban a felismerésünk a legnagyobb valószínűséggel bekövetkező fonémacsoport lesz, ellenkező esetben az előző időpillanatban felismert fonémacsoportot tartjuk meg. Az egyenlőtlenség elemei:

- $P(o_t \mid f_{i\max})$ az aktuális keret *i*-edik sorszámú osztály feltételes valószínűségi értéke, az SVM kimenetéből adódik. o_t az aktuális megfigyelés (ami az aktuális jellemzővektorunk). A $f_{i\max}$ fonémacsoport értékét

  $$\arg\max_i [P(o_t \mid f_i) \cdot b_{j,i}]$$

  szerint kapjuk meg, ami meghatározza a legvalószínűbb fonémaosztályt az adott keretre.
- $b_{j,i_{\max}}$ az adatbázisból előre kiszámított statisztikai érték, ami megadja, hogy az adatbázisban az *i*-edik fonémacsoport milyen valószínűséggel következik j-edik után. *j* a *t-1* időpillanatban felismert fonéma csoport, *i* pedig az aktuális időpillanatbeli.
- $P_i(o_t \mid f_i)$ megadja, hogy az előző időpillanatban felismert fonémacsoport az aktuális időpillanatban milyen valószínűséggel következik be. Az SVM kimenete adja.
- a_j ugyancsak a priori tudás az adatbázisból arra vonatkozóan, hogy az aktuális keret milyen valószínűséggel marad helyben, időben előrehaladva egyre csökken.

A következőkben bemutatom a statisztikai simítás eredményét a két előfeldolgozási eljárásra. A 17-es táblázat mutatja első esetben, a magyar beszédadatbázison kapott eredményeket.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.67**	0.04	0.01	0.01	0.18	0.01	0.03	0.01	0.02	0.01
hv	0.10	**0.62**	0.01	0.01	0.18	0.02	0.02	0.02	0.01	0.01
s+	0.08	0.06	**0.43**	0.07	0.25	0.05	0.02	0.01	0.02	0.02
s-	0.04	0.01	0.07	**0.66**	0.06	0.01	0.04	0.00	0.05	0.05
na	0.15	0.10	0.01	0.01	**0.60**	0.05	0.03	0.01	0.01	0.02
b+	0.02	0.04	0.02	0.01	0.41	**0.41**	0.05	0.02	0.00	0.02
b-	0.01	0.00	0.02	0.04	0.11	0.09	**0.65**	0.01	0.04	0.03
vo	0.06	0.08	0.03	0.00	0.29	0.09	0.04	**0.33**	0.03	0.04
uv	0.02	0.01	0.03	0.12	0.03	0.01	0.09	0.06	**0.49**	0.130
si	0.01	0.00	0.00	0.03	0.02	0.03	0.15	0.00	0.04	**0.71**

17. táblázat: Statisztikai simító eljárás eredménye MRBA adatbázison szűrősávokban mért energiával végzett előfeldolgozással

Az összesített felismerés 64%, ami 1%-ban tér el a szabálybázisú simítás eredményétől, ugyanezzel az előfeldolgozással. A szegmentálási pontosság 72,39%. A magánhangzók detektálásában is a szabálybázisú simítás a jobb. A második előfeldolgozási eljárással viszont nagyobb találati arányt produkált: 67%. A részletes eredmények a következő táblázatban láthatóak (18. táblázat).

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.78**	0.02	0.00	0.01	0.11	0.00	0.01	0.01	0.03	0.02
hv	0.18	**0.54**	0.00	0.01	0.19	0.01	0.01	0.01	0.03	0.02
s+	0.15	0.06	**0.32**	0.09	0.29	0.02	0.03	0.00	0.02	0.02
s-	0.06	0.02	0.01	**0.67**	0.06	0.00	0.06	0.00	0.07	0.06
na	0.30	0.07	0.00	0.01	**0.54**	0.01	0.02	0.00	0.01	0.04
b+	0.11	0.04	0.01	0.01	0.57	**0.18**	0.06	0.00	0.00	0.03
b-	0.07	0.02	0.00	0.02	0.13	0.00	**0.67**	0.00	0.00	0.09
vo	0.04	0.03	0.01	0.00	0.48	0.18	0.08	**0.09**	0.03	0.05
uv	0.01	0.00	0.00	0.09	0.04	0.00	0.27	0.00	**0.44**	0.15
si	0.02	0.00	0.00	0.01	0.03	0.00	0.09	0.00	0.02	**0.84**

18. táblázat: Statisztikai simító eljárás eredménye MRBA adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

Bár a magánhangzók felismerése magasabb, mint az első előfeldolgozással, a szabálybázisú simítás részeredményei is jobbak. Az összesített szegmentálási pontosság is kevesebb (mint előző esetnél), 70.54%-ra adódik.

Az angol adatbázisra a szűrősávokban mért energiával végzett előfeldolgozással kapott eredményeket a 19. táblázat tartalmazza.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.79**	0.05	0.01	0.01	0.09	0.00	0.01	0.02	0.02	0.00
hv	0.35	**0.31**	0.02	0.02	0.20	0.01	0.01	0.05	0.04	0.00
s+	0.20	0.02	**0.14**	0.27	0.13	0.06	0.04	0.06	0.07	0.02
s-	0.03	0.01	0.02	**0.64**	0.03	0.01	0.03	0.01	0.21	0.02
na	0.45	0.02	0.03	0.02	**0.32**	0.06	0.02	0.04	0.04	0.01
b+	0.03	0.01	0.01	0.02	0.17	**0.41**	0.19	0.05	0.03	0.08
b-	0.02	0.01	0.00	0.06	0.06	0.06	**0.49**	0.01	0.09	0.19
vo	0.08	0.03	0.03	0.03	0.15	0.05	0.05	**0.22**	0.33	0.04
uv	0.11	0.02	0.00	0.11	0.03	0.01	0.08	0.02	**0.57**	0.05
si	0.04	0.01	0.00	0.05	0.02	0.02	0.09	0.01	0.11	**0.64**

19. táblázat: Statisztikai simító eljárás eredménye TIMIT adatbázison szűrősávokban mért energiával végzett előfeldolgozással

Az összesített felsimerés 54%, 73,58% szegmentálási pontosság mellett. Ugyancsak ezen az adatbázison a frekvencia és időbeli deriválással végzett előfeldolgozással 55%-os összesített találati arányt eredményezett (20. táblázat).

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.84**	0.04	0.00	0.02	0.05	0.00	0.00	0.01	0.03	0.00
hv	0.47	**0.28**	0.00	0.02	0.15	0.00	0.01	0.02	0.03	0.01
s+	0.28	0.02	**0.06**	0.32	0.16	0.01	0.07	0.01	0.05	0.03
s-	0.05	0.01	0.00	**0.72**	0.03	0.00	0.05	0.00	0.10	0.03
na	0.56	0.02	0.00	0.02	**0.29**	0.01	0.03	0.01	0.04	0.02
b+	0.18	0.03	0.00	0.02	0.20	**0.16**	0.31	0.00	0.01	0.08
b-	0.13	0.02	0.00	0.07	0.04	0.00	**0.61**	0.00	0.03	0.10
vo	0.13	0.02	0.01	0.04	0.16	0.07	0.14	**0.07**	0.31	0.06
uv	0.12	0.01	0.00	0.11	0.03	0.00	0.16	0.00	**0.50**	0.06
si	0.08	0.02	0.00	0.05	0.03	0.00	0.08	0.00	0.07	**0.67**

20. táblázat: Statisztikai simító eljárás eredménye TIMIT adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

Majd végül a német KIEL adatbázisra. Az első esetben az összesített találati ráta csak 47%, míg a szegmentálási pontosság 73,25% (21. táblázat).

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.53**	0.11	0.02	0.01	0.19	0.05	0.01	0.06	0.01	0.01
hv	0.15	**0.52**	0.01	0.00	0.20	0.06	0.00	0.05	0.01	0.00
s+	0.05	0.04	**0.25**	0.05	0.09	0.24	0.03	0.18	0.03	0.04
s-	0.05	0.04	0.13	**0.44**	0.07	0.07	0.02	0.10	0.05	0.03
na	0.04	0.09	0.03	0.03	**0.46**	0.24	0.01	0.06	0.01	0.03
b+	0.01	0.02	0.04	0.02	0.13	**0.44**	0.07	0.10	0.04	0.14
b-	0.01	0.01	0.04	0.05	0.07	0.47	**0.17**	0.11	0.03	0.05
vo	0.04	0.06	0.07	0.01	0.05	0.02	0.00	**0.41**	0.26	0.07
uv	0.12	0.06	0.02	0.05	0.02	0.01	0.03	0.37	**0.19**	0.11
si	0.02	0.01	0.01	0.01	0.01	0.11	0.06	0.12	0.04	**0.61**

21. táblázat: Statisztikai simító eljárás eredménye KIEL adatbázison szűrősávokban mért energiával végzett előfeldolgozással

A másik előfeldolgozással ugyan jobb eredmények jöttek ki, viszont ez sem mérhető össze a szabálybázisú simító eljárás eredményeivel (22. táblázat). Az összesített találati arány 58%, a szegmentálási pontosság pedig 72,15%.

	mv	hv	s+	s-	na	b+	b-	vo	uv	si
mv	**0.67**	0.05	0.01	0.01	0.18	0.02	0.01	0.01	0.03	0.02
hv	0.29	**0.39**	0.01	0.01	0.23	0.03	0.00	0.01	0.02	0.01
s+	0.07	0.04	**0.21**	0.13	0.19	0.16	0.06	0.01	0.04	0.08
s-	0.07	0.04	0.03	**0.65**	0.08	0.01	0.06	0.00	0.05	0.03
na	0.11	0.08	0.01	0.04	**0.57**	0.07	0.03	0.01	0.02	0.06
b+	0.07	0.05	0.01	0.04	0.25	**0.22**	0.21	0.00	0.02	0.15
b-	0.12	0.04	0.00	0.08	0.15	0.08	**0.44**	0.00	0.01	0.07
vo	0.03	0.01	0.03	0.04	0.10	0.16	0.06	**0.08**	0.42	0.06
uv	0.03	0.01	0.01	0.07	0.04	0.03	0.10	0.08	**0.47**	0.17
si	0.02	0.00	0.00	0.01	0.02	0.01	0.15	0.00	0.03	**0.76**

22. táblázat: Statisztikai simító eljárás eredménye KIEL adatbázison frekvencia és időbeli deriválással végzett előfeldolgozással

Itt is készítettem egy, a simítóeljárás összegzésére vonatkozó táblázatot: 23. táblázat.

	Adatbázis	Előfeldolgozás szűrősávokban mért energiával	Előfeldolgozás frekvencia és időbeli deriválással
Összesített találati arány	MRBA	64%	67%
	TIMIT	54%	55%
	KIEL	47%	58%
Magánhangzó találati arány	MRBA	71.5%	76%
	TIMIT	75%	81.5%
	KIEL	65.5%	70%
Összesített szegmentálási pontosság	MRBA	72.39%	70.54%
	TIMIT	73.58%	70.79%
	KIEL	73.25%	72.05%
Magánhangzó szegmentálási pontosság	MRBA	73.32%	71.38%
	TIMIT	72.34%	68.97%
	KIEL	75.64%	72.36%

23. táblázat: Statisztikai simító eljárás összesített eredményei

Általánosan csökkentek az értékek az előző simító eljáráshoz képest, nemcsak összesített találatban, hanem magánhangzó detektálásban és szegmentálási pontosságban is. Összességben a frekvencia és időbeli deriválással végzett előfeldolgozási eljárással jobb eredmények születtek, azon belül is az összesített felismerésben, a magyar MRBA adatbázison 72% nagyon jónak minősül. A 89,5%-os magánhangzó detektálással pedig további céloknak megfelelő. Az összesített eredményeket tartalmazó táblázat alapján jól látszik, hogy az eljárás nyelvfüggetlennek mondható. A különböző nyelvű adatbázisokon végzett kiértékelés eredménye összevethető egymással. Így elmondható, hogy felhasználhatóak olyan alkalmazásokban, amelyek több (európai) nyelvre készülnek.

7 Összegzés, megjegyzés, további feladatok

A feladatom céljául 9 akusztikai osztály szeparálását tűztem ki, azon belül is a magánhangzók jó elkülönítését, mindezt nyelvfüggetlenül.

A dolgozatom folyamán egy új előfeldolgozási eljárást hasonlítottam össze egy, a beszédfelismerés területén sokszor használt előfeldolgozási eljárással. Először összemértem az előfeldolgozások hatására keltett SVM kimenetet. Ehhez három adatbázist használtam fel: a magyar MRBA, az angol TIMIT és a német KIEL adatbázisokat. Eredményül azt kaptam, hogy az új eljárás, amely idő és frekvenciabeli deriváltak segítségével építi fel a bemenő vektorokat, jobb összesített felismerési aránnyal rendelkezik, valamint a magánhangzókat is nagyobb hatékonysággal detektálja. Ezután az SVM kimenetére két fajta simító eljárást alkalmaztam (szabálybázisú és statisztikai), amelyeknek célja, hogy valósítsák meg a fonémacsoport szintű végső szegmentálást. A simító eljárások vizsgálata közben több paraméter által vetettem össze az két típust: összesített találati arányt, magánhangzó találati arányt, összesített szegmentálási pontosságot és magánhangzó szegmentálási pontosságot is néztem. Arra a végkövetkeztetésre jutottam, hogy a legjobb eredmény akkor született, amikor a frekvencia és időbeli deriválással végzett előfeldolgozással építjük fel a tulajdonságvektorokat, az SVM kimenetén pedig szabálybázisú simítást végzünk. Ilyenkor a magánhangzók találati aránya nagy, a nyelvfüggetlenség jelentős, a magánhangzó szegmentálási pontossága is elfogadható.

A sikeres előfeldolgozási eljárást alkalmazásba implementálva lehetőségünk nyílik a magánhangzók nyelvfüggetlen detektálására. A munka folytatásaként kivalló lehetőség nyílik az eljárás felhasználására a beszéd akusztikai paramétereinek megjelenítésében, audió-vizuális kiejtésoktató rendszerekben.

8 Irodalomjegyzék

[1] Vicsi K, Sztahó D: Recognition of emotions on the basis of different levels of speech segments. Journal of advanced computational intelligence and intelligent informatics 16:(2) pp. 335-340. (2012)

[2] Vicsi Klára, Sztahó Dávid, Kiss Gábor: Examination of the sensitivity of acoustic-phonetic parameters of speech to depression. In: 3rd IEEE International Conference on Cognitive Infocommunications (CogInfoCom 2012). Kassa, Szlovákia, 2012.12.02-2012.12.05. pp.

[3] Klára Vicsi, Viktor Imre, Gábor Kiss: Improving the Classification of Healthy and Pathological Continuous Speech.In: Petr Sojka, Aleš Horák, Ivan Kopeček, Karel Pala (szerk.) Text, Speech and Dialogue: 15th International Conference, TSD 2012. Brno, Csehország, 2012.09.03-2012.09.07. Springer, pp. 581-588.

[4] Kiss Gábor, Vicsi Klara: Akusztikai hangosztályok felismerésén alapuló, nemlineáris idővetemítés megvalósítása a mondathanglejtés és a szóhangsúlyozás oktatásához. Beszédkutatás: pp. 247-261. (2013)

[5] Bachu R.G., Kopparthi S., Adapa B., Barkana B.D.,"Voiced/Unvoiced Decision for Speech Signals Based on Zero-Crossing Rate and Energy ", IEEE International Joint Conferences on Computer Information, and Systems Sciences, and Engineering (CISSE'08).

[6] M.Malcangi: Softcomputing approach to segmentation of speech in phonetic units, International journal of computers and communications. Issue 3, Volume 3, 2009. 41-48

[7] Jalil, M. ; Butt, F.A. ; Malik, A. Short-Time Energy, Magnitude, Zero Crossing Rate And Autocorrelation Measurement For Discriminating Voiced And Unvoiced Segments Of Speech Signals , Technological Advances In Electrical, Electronics And Computer Engineering (Taeece), 2013, 208 - 212

[8] WANG Li-juan, CAO Zhi-gang. Automatic Phonetic Segmentation Using HMM Model. Journal of Data Acquisition & Processing, 2005, 20(4):381-384.

[9] Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László: Beszédadatbázis irodai számítógép-felhasználói környezetben, Second Conference on Hungarian Computational Linguistics (MSZNY 2004), Szeged, 2004. (p. 315)

[10] Benno Peters. The Kiel Corpus of Spontaneous Speech.
http://www.ipds.uni-kiel.de/kjk/pub_exx/aipuk35a/aipuk35a_1.pdf

[11] Garofolo, J.; Lamel, L.; Fisher, W.; Fiscus, J.; Pallett, D.; & Dahlgren, N. (1990). DARPA, TIMIT Acoustic-Phonetic Continuous Speech Corpus CD-ROM. National Institute of Standards and Technology, 1990.

[12] Vicsi Klára: „Adatbázisok a beszédtechnológia szolgálatában”. In: Németh G.-Olaszy G. (szerk.): A magyar beszéd, Akadémiai kiadó, 2010. p. 265.

[13] Olaszy Gábor: „Hangjelölések”. In: Németh G.-Olaszy G. (szerk.): A magyar beszéd, pp 77-79

[14] Vicsi, Klára: „SAMPA computer readable phonetic alphabet, Hungarian,” (2008)

[15] Szaszák György: Kepsztrum – A magyar beszéd, 2010 Bp., 239. – 242. oldal

[16] Beke András – Szókezdetek automatikus osztályozása spontán beszédben http://www.c3.hu/~nyelvor/period/1352/135209.pdf

[17] Mády Katalin: Beszédpercepció és pszicholingvisztika 7. – 8. oldal

[18] Vicsi Klára, Kocsor András, Teleki Csaba, Tóth László: Beszédadatbázis irodai számítógép-felhasználói környezetben. http://alpha.tmit.bme.hu/speech/docs/cikkek/mrba.pdf

[19] Vicsi Klára: Tanító adatbázisok gépi beszédfelismeréshez – A magyar beszéd, 2010 Bp., 274. – 276. old.

[20] Benno Peters – The Kiel Corpus of Spontaneous Speech. http://www.ipds.uni-kiel.de/kjk/pub_exx/aipuk35a/aipuk35a_1.pdf

[21] Szaszák György: MFCC-paraméterek - A magyar beszéd, 2010 Bp., 240. – 242. old.

[22] Vicsi Klára: A hallás folyamata- A magyar beszéd, 2010 Bp.,33.-35. old.

[23] Zwicker, E.: Subdivision of the audible frequency range into critical bands. The Journal of the Acoustical Society of America,. 1961.

[24] http://alpha.tmit.bme.hu/speech/databases.php

[25] Horváth Gábor, Altrichter Márta, Pataki Béla, Strausz György, Takács Gábor, Valyon József: Neurális hálózatok, Hungarian Edition, Budapest: Panem Könyvkiadó Kft., 2006