

Pécsi Tudományegyetem
Pollack Mihály Műszaki és Informatikai Kar
Mérnök Informatikus Szak MSC

TDK DOLGOZAT

Hangfelvétel, szűrés és hanganalizálás/felismerés
LabView programban

Készítette: Czimmermann Tamás
Konzulens: Shiffer Ádám

Pécs

2014

Tartalom

1. Bevezetés	3
2. A beszédfelismerésről általánosan	4
2.1. A beszédfelismerés lépcsőfokai	4
2.2. A beszédfelismerés módszerei	6
3. A magánhangzók	9
4. A beszédjel-feldolgozás matematikai háttere	12
4.1. Diszkrét idejű feldolgozási módszerek	12
4.1.1. Gyors Fourier transzformáció (FFT).....	12
4.1.2. Ablakoló függvények	14
4.1.3. Konvolúció (szűrés)	15
4.2. Folytonos idejű feldolgozási módszerek	16
4.2.1. Gördülő energia.....	16
4.2.2. Lineáris predikció.....	17
5. Magánhangzó-felismerő program	18
5.1. LabView 2012-ről röviden	18
5.2. A felvevő modul	19
5.3. Magánhangzó lokalizáló	20
5.4. Formáns kereső	22
5.5. Magánhangzó meghatározó.....	27
5.6. Kísérleti fázisban lévő szókereső	28
6. Összegzés	29
IRODALOMJEGYZÉK.....	30

1. Bevezetés

Közismert tény, hogy az ember különböző kommunikációs technikái közül, a természetes hang és beszéd képes egy adott időintervallum alatt a legtöbb információt szolgáltatni, így a kommunikációink közül a beszédnek a legnagyobb az adatátviteli sebessége. Gondoljunk csak bele, hogy azon kívül, hogy a beszédben egymás után következő képzett szavak jelentéssel bírnak, ez felett még egy mondatnak hangsúlya, egy kifejezésnek hangereje is van, mellyel még számtalan érzelmet kommunikálunk. A tudósok hosszú idők óta próbálnak e területen tökéletes hangfelismerőt, beszéd-analizátort és egyéb eszközt építeni, programozni vagy algoritmusokat kitalálni e bonyolult rendszer tökéletes vizsgálatára.

Mostanra a beszédfelismerés és hanganalizálás eljutott arra a szintre, hogy már önálló tudományágnak tekinthető. A beszédfelismerésben megjelenő problémák komplexitása miatt e tudományág jelentősen kapcsolódik más tudományokhoz, mint például anatómia, pszichológia, nyelvészet, logopédia, akusztika, matematikai statisztika, információelmélet, számítástudomány. Nálunk gazdaságilag fejlettebb országokban, vagy inkább angolul beszélő országokban már szinte napi szinten használják a beszédfelismerőket. Ennek leginkább az a fő oka, hogy a felismerők által használt hangzó prototípusok, mintaadatbázisok és nyelvi modellek erősen kötődnek az adott nyelvhez. Nálunk gazdaságilag fejlettebb országokban volt/van tökéletesen ezen technológiák megalkotására és fejlesztésére, így amíg hozzánk elérték ezen technológiák, addig a fejlettebb országokban már ipari méretekben gyártják az eszközöket. Ám hála a világháló fejlődésének, a gazdaságilag visszamaradott vagy kisebb országok is fel tudnak zárkózni, mivel nemzetközi fejlesztések keretében a cégek lépéseket tettek már, hogy minden nyelven legyen valamiféle hangfelismerő program. Ennek nagy úttörői a Google, a Mac és néhány Android operációs rendszerre fejlesztő cég. Ezen cégek az utóbbi időben saját hangfelismerő motorokat fejlesztettek ki, bár az Android operációs rendszerre fejlesztő cégek többségében a Google felismerő motorjára hagyatkoznak. Ezen programok internetkapcsolaton keresztül akárhol elérhetőek, így már nem is az hardver a fontos, hanem hogy internetkapcsolatunk legyen.

Ám nem feltétlen szeretnénk vagy tudunk mindig internethez kapcsolódni és ezen tudományágon még rengeteg kiaknázatlan terület van. Egy olyan program megalkotása volt a célom, mely nem túl bonyolult módon, de megbízhatóan, kis erőforrást használva tud szavakat felismerni. A program alapvetően feltételezi, hogy céljának megfelelően használják és beszélnek bele, egy szóval beszédet „hall”. A beszédnek és szavaknak vannak könnyen felismerhető és

őket nagymértékben jellemző alkotóelemeik. Ezek a magánhangzók. Magánhangzó észlelő, detektáló programot írni, nem túlságosan nagy kihívás, mivel egy szimpla gördülő energiaspektrummal megoldható, ám a detektált magánhangzó milyenségét meghatározni már egy jelentősebb nehezebb feladat. A szakdolgozatom és a programom nagyobb része a kiejtett szavakban való magánhangzó lokalizációt és identifikációt valósítja meg.

Minden hasznos programnak van egy célja, mely a gyakorlatban való alkalmazhatósággal reprezentálható. Terveim, hogy e programot a későbbiekben tudom implementálni egy drón vagy egy kerekesszék vezérlésébe. Ez a program egy formánsanalízisen alapuló mintaillesztéssel adja vissza a bemondott szavakat. A szavak egy adatbázisba vannak betáplálva szövegesen. Mind a formánsfrekvencia táblázat, mind a szótáblázat nem igényel nagy adatbázis, mindössze néhány kilobájt háttértárolót.

Dolgozatomban először beszélek a beszédfelismerés fejlődésének rövid történetéről, majd rátérek magára hangképző szervekre, mivel a beszédfelismerés megértéséhez elengedhetetlen, hogy megértsük a beszédképzést és a beszédet képző szervek működését. Ezek után felvázolom a keretrendszert és a programot, amiben a saját alkalmazásomat készítettem, majd részletesen bemutatom a hangfelismerő programom működési elvét és készítési fázisait. Abban a fejezetben beszélni fogok, a programom alapvető matematikai algoritmusairól, melyet a hanganalizáláshoz használtam, valamint az alkalmazás készítése során adódott problémákról, azok megoldásairól. Végezetül vázolom a programom gyakorlati alkalmazása felé tett próbálkozásaimat, és fejlesztéseimet, mely egy rendszer irányítását célozza meg.

2. A beszédfelismerésről általánosan

2.1. A beszédfelismerés lépcsőfokai

A bevezetés után, most rátérnék a beszédfelismerés elvi hátterére, történelmére és általános elgondolására.

A beszédfelismerésnek három szintjét három fő lépcsőfokát különítjük el; az akusztikai szintet, a szintaktikai szintet és a szemantikai szintet.

Az akusztikai szint a beszédjel hullámként való értelmezése. Tulajdonképpen a beszédjel jellemző, kiemelkedő fizikai paramétereit elemezzük, mint például idejét, időtartamát, spektrális eloszlását, amplitúdóját, hangerejét ... stb. Ezzel a módszerrel a kapott értékeket, paramétereket elemezve rálátásunk nyílik a beszédjel lényeges jellemzőinek

kiemelésére. Ez módszer azonban nem elég a teljes beszédfelismerés megvalósítására, ezzel a szinttel csak becsléseket lehet tenni, mivel fizikai értékeként kezeljük a beszédjelet. Fizikai értékként nézve viszont a beszédjelünkben nem jelenik meg a nyelv, a nyelvtan. Nincsenek közvetlen mondat vagy szó határaink. Ha gyorsan, hadarva elemzünk akusztikailag egy szót, lehet számunkra értelmetlen „zajt” kapunk, míg egy felsőbb szintű elemzésnél, ahol szövegkörnyezetben értelmezzük a beszédjelet, már is értelmes jelentéssel bíró jelet kapunk. Egy akusztikai elemzőprogram leginkább személyenkénti szabályrendszert és kalibrálást igényel, mivel nincs két ugyanúgy beszélő ember a világon.

A fonémák a beszédet alkotó alapszimbólumok. Ezen szimbólumok közti kapcsolatot és kapcsolódási szabályokat foglalja egybe a szintaktikai szint. Szabályok alatt kell érteni a kiejthetőséget, mi szerint egy szintaktikai szinten már nem engedhetünk meg egymás után kiejthetetlen hangzókat. Érezhető hogy ezen a szinten már megjelenik a beszédben használatos nyelvtani szabályok alkalmazása (grammatikai elemzés), szótagok illeszkedése és a szavak és toldalékok viszonya. Ezen szinten találhatjuk még az adott nyelvek szótárkészletét, azaz a nyelvben szavakként, toldalékokként megengedett szimbólumsorozatokat. Mivel a nyelveknek nincs egy mindenki számára egységesen értelmezett szabályrendszere és szótárkészlete, ezért szintaktikai szabályrendszer megalkotásakor figyelembe kell venni az szabályt megfogalmazó személy szubjektivitását.

A szemantikai szinten, a szintaktikai szinten mondhatni elfogadott szimbólumok jelentését vizsgáljuk a valóság tükrében, azaz a közlés értelmét. A szemantikai és a szintaktikai szintet olykor könnyű összekeverni, de valójában egy éles határ húzódik a kettő szint között. Gyakran a művészetben, az irodalomban találkozunk olyan írásokkal, ahol ezt az éles határt egyértelműen megállapíthatjuk. „Futtam, mint a szarvasok, lány bánat a szememben. Famardosó farkasok üznek vala szívemben.” (*József A. – Bánat, 1930 jún.*). Ezen idézetben megfigyelhetjük, hogy bár szintaktikailag teljesen rendben van, szemantikailag nem biztos hogy az olvasó teljesen tisztában van a jelentésével. Az hogy nem tudjuk feltétlen értelmezni az idézetet, nem azt jelenti, hogy rosszak a szókapcsolatok benne, hanem egy szemantikai elemzés szintjén egy elemzőnek nem kötelező tulajdonsága a műértés. A szemantikai szabályrendszer megalkotójának szubjektuma igen jelentősen megjelenik a szabályaiban. Összevetve a három szint közül, mindnél feljebb megyünk az elemzési szinten, annál bonyolultabb, komplexebb elemzési problémákkal szembesülünk, melyeknek megoldásában egyre több az önkényesség.

2.2. A beszédfelismerés módszerei

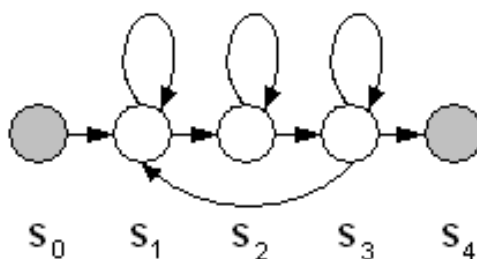
Az első beszédfelismerők egyértelműen a beszédfelismerés lépcsőjének legelső lépcsőfokán kezdtek. Akusztikai szinten vizsgálták a beszédet, kizárólag fizikai modellekből próbálták meg megállapítani a beszéd jelentését. Kváziperiodikus hangok formánsait vizsgálva rábírták a kezdeti felismerőket néhány alapvetően egyszerű zöngés mássalhangzó és a magánhangzók felismerésére. Erről kezdetekben azt hitték, hogy elég, de hamar ráébredtek, hogy sajnos ezen a szinten még nem lehet biztonsággal minden hangot felismerni. Nem tudtak megbirkózni a rövid ideig ejtett magánhangzókkal, azaz az átmeneti magánhangzókkal és emellett nem tudták megoldani a beszédfüggetlenséget.

Erre a problémára a későbbiekben találtak is megoldást. Megfigyelhetjük, hogy az ember alapvetően több hangot képes meghallani és értelmezni, mint amennyit kiejteni. Egy szóval fókuszáljunk arra, hogy milyen hangot képeztünk, nem pedig hogy milyen hangot hallottunk. Ezzel az állásponttal jutottak el, statisztikai beszédfelismerő módszerekig. Ezen felismerők alapvető tulajdonsága, hogy gigantikus adatbázisokban tárolják az adott nyelvek szótárkészletét, mégpedig egy új kiejtési modellben, ezen felül még tárolják a nyelvtani szabályokat. Ezen felismerők működési elvének alapja a valószínűségi rátákról szól, azaz a beérkezett beszédjel mely modellsorozatból épülhet fel a legnagyobb valószínűséggel. Ezen beszédfelismerők fejlődéséhez nagymértékben hozzájárult a fonetikus szimbólumok megalkotása. Ezen szimbólumok egy olyan jelölésrendszer alkotóelemei, melynek célja hogy nemzetközi szakemberek is képesek legyenek egy adott nyelv ismerete nélkül is reprodukálni vagy vizsgálni egy hangot. Az első ilyen jelölésrendszer az IPhA (International Phonetic Alphabet), mely az XX. században keletkezett és mellyel több száz nyelv, szinte össze hangját tökéletesen írja le. Egyértelmű, hogy ha szimbólumokról beszélünk, azoknak létezik valami írásos formája is. Az átlag billentyűzetek ezen szimbólumok bevitelére és kezelésére alkalmatlanok, ezért kifejlesztettek egy újabb módszer, mely sokkal inkább alkalmas írásba foglalni a fonetikus szimbólumokat. Ez a rendszer a SAMPA, melyet a 90-es években fejlesztettek ki. Az európai adatbázisok SAMPA karaktereket használnak [1.]. Ám a SAMPA igencsak nyelvfüggő, ezért fejlesztették tovább és alkották meg az X-SAMPA jelölésrendszert, melynek célja, hogy az IPA fonetikai latin betűkkel (pontosabban ASCII-karakterekkel) reprezentálja [2.]. Erről láthatunk pár példát az 2.1-es táblázatban.

X-SAMPA	IPA	IPA	Leírás	Példák (szögletes jelölésben az X-SAMPA)
a	a	a	nyílt, elől képzett, ajakkerekítés nélküli magánhangzó	francia dame [dam], spanyol padre [padre]
b	b	b	bilabiális, zöngés zárhang	angol bed [bEd], francia bon [bO~]
b_<	ɸ	ɸ	bilabiális, zöngés implozív (belső glottális hang)	
c	c	C	palatális, zöngétlen zárhang	magyar latyak [lAcAk]

2.1. táblázat Hangzók szimbólumtáblázata [2.]

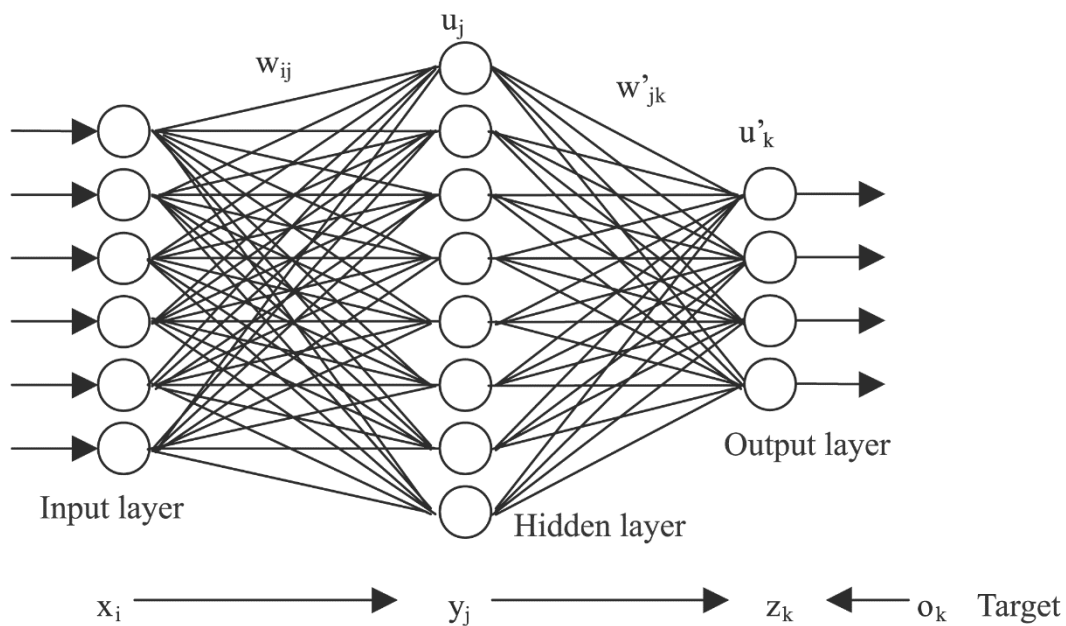
A statisztikai beszédfelismerésnél kétféle statisztikai alapú feldolgozási elvet ismerünk, a Rejtett Markov Modellt (Hidden Markov Model - HMM) és a Neurális Hálózatokat (Neural Network - NN). A HMM lényege, hogy a mintaadatbázisban eltárolt mondatokat részekre bontjuk és diszkrét véges automaták sorozataiként eltároljuk. Ez hatalmas adatbázist eredményez. Minden állapotmenetnek van egy valószínűségi rátája és ezen állapotok a fonémák prototípusai. Egyszerűsíthetjük a problémát, ha láncoljuk a listát olyan módon, hogy egy bizonyos fonéma ejtésének valószínűségét az előtte lévő N-1. fonéma határozza meg. Gyakorlatban a háromelemű trigram automaták a használatosak [4.]. Ez az automata látható az 2.2.-es ábrán.



2.2. ábra Trigram automata [4.]

Egy beszédvizsgálat folytán az algoritmus a vizsgálni kívánt jelsorozatot szakaszokra vágja és elsőlépcsős (akusztikai) vizsgálatot végez rajtuk. Az így nyert fonémaprototípusokat összehasonlítja az adatbázisban tárolt trigramokkal és eközben egy nyelvtani szabályokra támaszkodva felépít egy hasonló trigramokból álló beszédátmeneti gráfot. Ha a gráf elkészült, megkeresi a beszédszakasz elejét és végét, és meghatározza a két gráf-pont közötti legvalószínűbb utat, és a beszédfelismerés eredményeképp azt a fonémasorozatot adja vissza [4.]. Ez a módszer rendkívül hatásos egy bizonyos szótárméretig (2000-3000), ám nagyobb szótáraknál kezelhetetlen nagy referencia-adatbázisokat generáltak az algoritmusok [5.]. Összevetve az agyunk működési elvével, ez az algoritmus sorosan működik, láncolva. A beérkezett hangmintán először teljes akusztikai elemzést végez, majd felépít a terjedelmes valószínűségi állapotgráfot, melyet utána még végig kell vizsgálnia. Ez egyértelműen lassabb az agy működésénél, még elvben is. Az agy asszociatív módon dolgozza fel a beérkezett hangmintát és párhuzamosan működik több szintű elemző. Nem végez teljes akusztikai elemzést, csak egy-két kulcshangról asszociál a mintacsoportokra, amikben keresnie kell. Egyszerűen sémák alapján szűkíti nagymértékben a keresési csoportokat, majd az akusztikai elemző tovább szűkíti keresés végkimenetelét, hogy mégis mit kellene még hallani a szövegben. Ezen elemzők közben folyamatos visszacsatolásban (feedback) vannak egymással, ez adja az agy gyors, párhuzamos működési elvét. Az agy kis adatbázissal dolgozik, ám rendkívül jól struktúrált és az elemzési szintek párhuzamosan működnek, szorosan kapcsolódva egymáshoz.

Ezek után a kutatók figyelme a Neurális Hálózatok [2.3. ábra] felé irányult, mivel azok sokkalta közelebb állnak az emberi agy működési elvéhez, ám rendkívül nehéz betanítása és matematikai leírások hiánya végett hamar rájöttek, hogy ez nem lesz egyszerű megoldás. Végül a HMM és az NN hibridje adta a megoldást [1.].

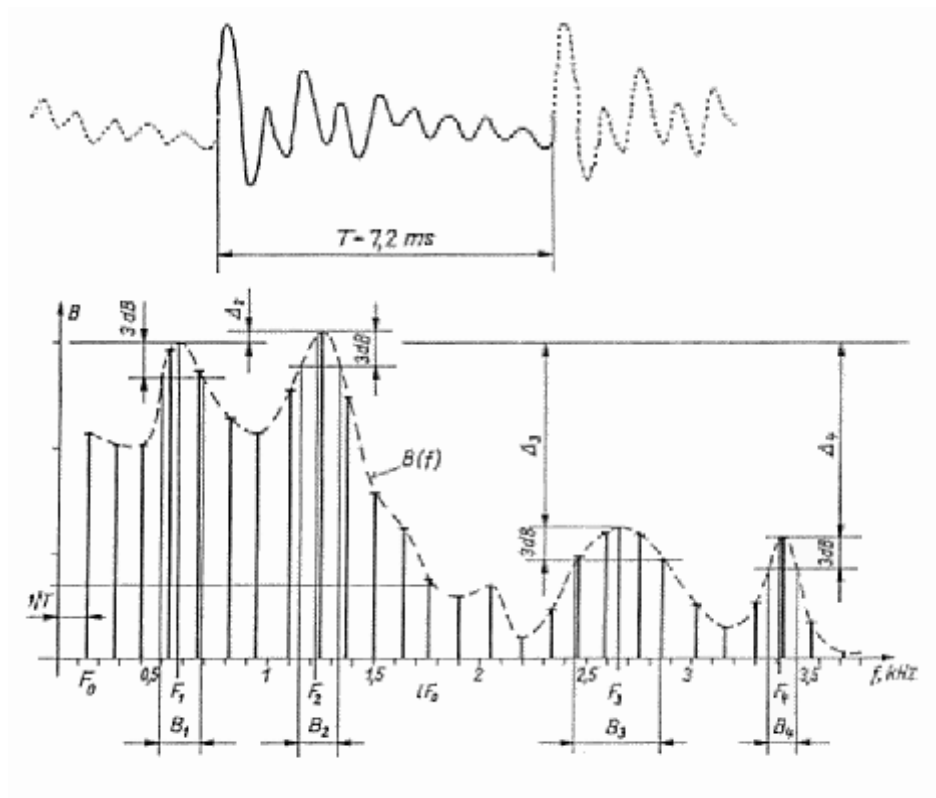


2.3. ábra Neurális hálózat felépítése [6.]

3. A magánhangzók

Az előző fejezetben betekintést nyerhettünk a beszédfelismerés általánosan kialakult képéről, történelméről és a használatban lévő módszerekről, ebben a fejezetben röviden bemutatom a számunkra fontos magánhangzók tulajdonságait, melyeket felhasználok a programomban.

A magánhangzók képzésekor a leginkább lényeges tulajdonság, hogy a levegő akadálytalanul áramlik keresztül, így egy folyamatosan ejthető, egyetlen kváziperiodikus zöngé gerjesztésű, formáns szerkezetű hangot kapunk. A hangszalagok felhangokban dús gerjesztőjeléhez még hozzáadódnak még az orr- és szájüreg valamint a garatüreg rezonátorrendszere felerősített felhangok. Ezen felhangcsoportok tipikusan a kiejtett hangzókat formálják és kiemelkednek a környezetükből, ahogyan a 3.4-es ábrán is látszik.



3.4. ábra A magánhangzók formánsstruktúrája [5.]

Ahogy a 3.4-es ábrán is látszik ezen 3dB-es lokális maximumokat nevezzük formánsoknak. Ezen formánsoknál mért frekvenciák a formánsfrekvenciák. Értelmezhetőek még a formáns sávszélességek, és a relatív formáns amplitúdók is. Ezek összességében alkotják egy magánhangzó rájellemező formánstruktúráját. Egy magánhangzó kiválóan felismerhető formánsainak kombinációjáról. Egy bizonyos magánhangzóban rengeteg formáns jelenik meg, ám az átlag felhasználásban mi az első 6 formánssal foglalkozunk. Az első három formáns kifejezetten a magánhangzóra jellemző, a következő három pedig magánhangzót kiejtett személyre, ám ezen formánsfrekvenciák nyelvekként különböznek. A 3.5-as táblázat tartalmazza a magyar formánsfrekvenciák minimum, maximum és középértékét, férfire és nőre értve egyaránt. Ezeket a méréseket Molnár József állította össze.

A magánhangzóknak egy számunkra rettentően fontos tulajdonsága, hogy a szabad légáramlásból és erős zöngéből kifolyólag a legtöbb energiát hordázzák az általunk kiejtett beszédhangok közül, ez teszi őket könnyen lokalizálhatóvá.

Magánhangzó	Formáns	Frekvencia-középérték Hz				Frekvencia-határérték Hz			
		nők		Férfiak		nők		férfiak	
		hang-súlyos	hang-súlytalan	hang-súlyos	hang-súlytalan	hangsúlyos	hangsúlytalan	hangsúlyos	hangsúlytalan
a	1. 2.	800	700	595	510	700...850	650...750	550...650	500...550
	3.	1320	1205	1115	1000	1200...1400	1200...1250	1050...1250	900...1050
		2700	2600	2350	2200	2600...2800	2500...2700	2300...2450	2100...2300
á	1. 2.	1130	1055	850	790	1000...1200	950...1100	800...900	700...850
	3.	1630	1540	1470	1430	1550...1700	1500...1600	1450...1500	1300...1500
		2660	2530	2300	2150	2550...2700	2450...2550	2200...2400	2000...2300
e	1. 2.	700	630	510	460	650...750	600...700	450...570	450...550
	3.	2100	1920	1750	1650	2000...2200	1850...2000	1700...1800	1600...1700
		3000	2850	2550	2510	2900...3100	2800...2900	2500...2600	2500...2600
é	1. 2.	490	410	425	380	450...550	350...500	300...500	280...450
	3.	2480	2390	2050	1975	2350...2550	2300...2450	1900...2150	1900...2050
		3050	2925	2600	2530	3000...3150	2800...3000	2550...2700	2500...2650
i	1. 2.	415	390	350	290	400...450	350...430	300...400	250...320
	3.	2490	2300	2100	2030	2450...2550	2250...2350	2000...2200	1950...2050
		3180	2950	2700	2650	3050...3250	2900...3000	2600...3000	2600...2750
í	1. 2.	310	275	270	250	250...350	250...300	200...300	200...300
	3.	2700	2525	2350	2300	2600...2800	2450...2550	2250...2450	2250...2400
		3450	3300	3150	2850	3100...3500	3000...3450	2950...3250	2700...2950
o	1. 2.	650	530	460	420	600...700	500...600	400...520	400...450
	3.	1225	920	1000	830	1050...1300	900...950	800...1150	750...910
		2750	2650	2400	2300	2650...2850	2550...2750	2300...2500	2200...2400
ó	1. 2.	570	430	415	335	500...650	400...500	350...450	300...400
	3.	970	750	830	700	900...1050	700...800	700...950	600...750
		2660	2600	2450	2400	2600...2750	2500...2700	2400...2500	2300...2500
ö	1. 2.	490	415	450	360	450...520	400...450	400...500	300...400
	3.	1520	1525	1360	1300	1550...1650	1500...1550	1300...1400	1250...1350
		2610	2525	2400	2300	2600...2700	2500...2550	2350...2450	2250...2400
ő	1. 2.	415	350	400	305	400...450	300...400	300...450	300...400
	3.	1630	1550	1500	1500	1600...1650	1500...1600	1400...1600	1400...1550
		2700	2600	2450	2375	2650...2750	2550...2650	2400...2500	2300...2400
u	1. 2.	450	425	350	315	400...500	400...450	300...400	300...370
	3.	1035	960	920	900	1000...1100	900...1020	800...960	800...1000
		2360	2280	2150	1950	2300...2400	2250...2350	2000...2300	1850...2050
ú	1. 2.	390	300	310	235	350...460	250...320	250...370	200...270
	3.	910	708	740	610	850...970	750...820	700...830	600...730
		2500	2450	2100	2050	2400...2600	2350...2500	2000...2250	2000...2150
ü	1. 2.	330	312	290	220	300...350	300...350	250...400	200...250
	3.	2025	1770	1870	1620	1950...2050	1750...1800	1700...1900	1550...1650
		2750	2700	2550	2430	2700...2800	2600...2750	2500...2600	2360...2500
ű	1. 2.	300	260	265	215	250...320	230...320	200...300	200...250
	3.	2180	1970	1910	1790	2100...2200	1950...2000	1850...1950	1700...1850
		2800	2725	2600	2525	2700...2850	2700...2750	2550...2650	2500...2600

3.5 táblázat A magyar magánhangzók formánsfrekvenciái [7]

4. .A beszédjel-feldolgozás matematikai háttere

Miután megismerkedtünk a beszédünkben használatos számunkra fontos magánhangzók tulajdonságaival az előző fejezetben, most bemutatom a programomban használt matematikai módszereket és elveket, melyeket sikeresen alkalmaztam a beszédfelismerő programban.

4.1. Diszkrét idejű feldolgozási módszerek

4.1.1. Gyors Fourier transzformáció (FFT)

„Összességében a három fontos összefüggés a Diszkrét Fourier Transzformációval (DFT) kapcsolatban:

$$F(h[n]) = H[n] = \Delta t \sum_{k=0}^{N-1} h[k] e^{-i \frac{2\pi}{N} kn}$$

$$H[n + N] = H[n]$$

$$H[n + k] = H[n] * e^{-i \frac{2\pi}{N} kn}$$

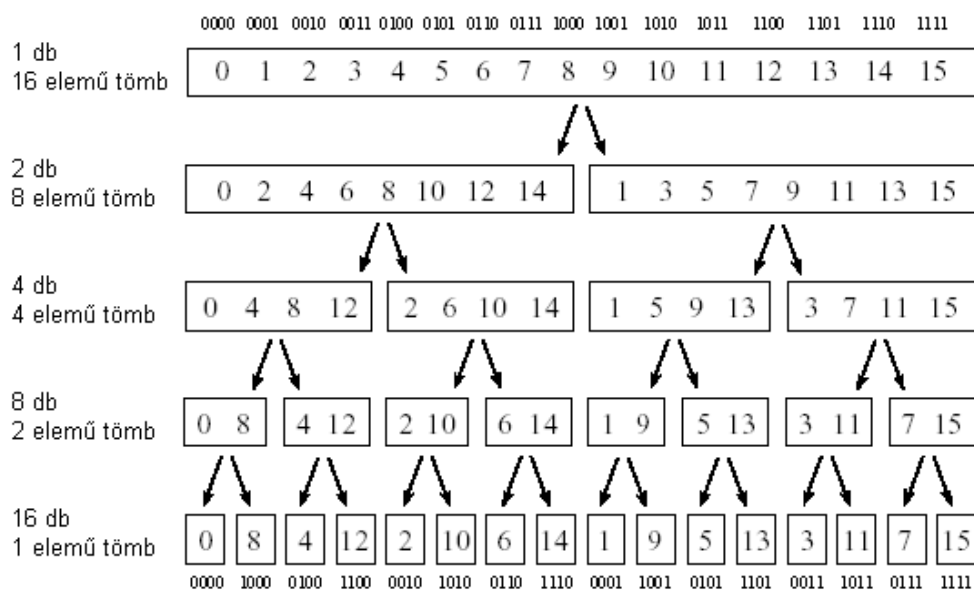
(4.1)

$H[n]$ összes elemének kiszámításához $O(N^2)$ művelet szükséges. Ezek szerint az alappontok számának növelésével az időigény négyzetesen növekszik. Ez egy 2 mp-es audio, 44100 Hz-en mintavételezett zenei részletnél $88200^2 = 7779240000$ műveletet jelent. Egy $1024 * 768$ pontból álló képnél $(1024 * 768)^2 = 618,47 * 10^9$ műveletet jelent! Nagyobb zenei részletnél, illetve több képkockánál a számítási időigény a jelfeldolgozást szinte lehetetlenné teszi a mai PC-k sebességéhez viszonyítva. A megoldást a Gyors Fourier Transzformáció adja (Fast Fourier Transform, FFT). Az FFT N db időtartománybeli diszkrét értékből N db frekvenciatartománybeli értéket állít elő a DFT-hez hasonlóan, ellenben műveletigénye $O(N * \log_2 N)$, amely logaritmikus volta miatt jelentős számításcsökkenést eredményez.

... Az FFT algoritmus alapelve, hogy $H[n]$ DFT értéksor felírható két DFT értéksorral, amelyek elemszáma $N/2$. Egyik DFT értéksor $h[n]$ páros, míg a másik értéksor $h[n]$ páratlan elemeinek DFT transzformáltjából jön létre. Azonban az előbbi gondolatsor továbbgondolható, ugyanis az $N/2$ elemű DFT értéksor felírható két $N/4$ elemű DFT értéksorral, és így tovább. A rekurzív

algorithmus végén az N db 1 elemű DFT transzformáltból állítjuk elő a 2 elemű transzformáltakat, majd ezekből a 4 eleműeket, ... stb. Nyilvánvaló, a fenti algoritmus csak akkor használható, ha az alappontok száma kettő hatványa, amely természetesen gyakorlatban nagyon ritkán fordul elő. Ezen könnyen lehet segíteni, csak ki kell egészíteni az alappontokat 0-al, egészen a következő 2^n számú alappontig. ...

A következő fontos kérdés, hogyan lehet eldönteni, hogy az algoritmus valamely szakaszában mely elemeket mely elemekkel párosítsam. Erre ad választ ..., a bit-fordított sorrend. ... Ezek szerint a bitfordított sorrend megoldást ad a párosítási problémára, tudjuk, melyik diszkrét elemet melyikkel párosítsuk.



4.1 Az $a(n)$ sorozat felbontása [13]

Azonban kérdés maradt, hogy áll elő egy N elemű DFT transzformált, ha ismert $N/2$ páros, és $N/2$ páratlan eleme. Nyilvánvaló, hogy külön-külön az N elemű DFT-ak összességében kevesebb információt hordoznak, mint az N elemű értéksor Fourier transzformáltja. Azonban észrevehető, hogy az egyik $N/2$ elemszámú értéksor időben el van tolva a másik $N/2$ értéksorhoz képest, méghozzá $1/N$ -el. Ez azt jelenti, hogy ha ismerjük a páros, ill. páratlan alappontokra támaszkodó $N/2$ elemű transzformáltakat, azokból az N elemű transzformált egyszerűen számítható.”[8.]

(4.2)

$$H_N[n] = H_{\frac{N}{2}PÁROS} + H_{\frac{N}{2}PÁRATLAN} * e^{-i\frac{2\pi}{N}n}$$

4.1.2. Ablakoló függvények

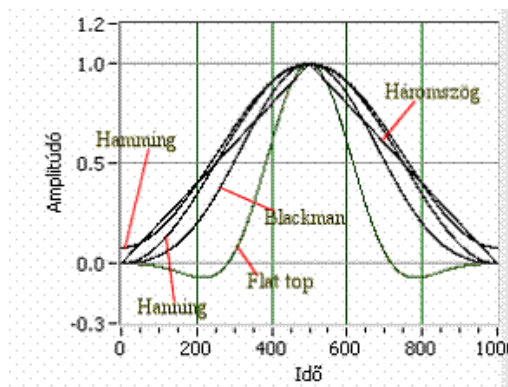
A DFT transzformációjánál az aperiodikus jelünket periodikusnak tekintjük, úgy hogy az N mintát periódusokban ismétljük, mint ahogy az előzőben is ez megjelent már összefüggésként:

$$H[n + N] = H[n]$$

(4.3)

Ez a fajta periodicitás azonban diszkrét jeleknél torzítást is előidézhethet, mert az N hosszúságú diszkrét minta ideje nem egész számú többszöröse a vizsgált frekvenciájú összetevő periódus idejének. Ezen torzulások, nem kívánt mellékvonalak megjelenésének oka egyszerű; mi a jelnek csak egy részét, időszelétet transzformáltuk úgy hogy periodikusnak tekintettük. Ám nem feltétlen biztos hogy ennek az időszelnek az elején és végén lévő jel egymás után téve egy folytonos jelet fog kiadni, sokkal inkább egy szakadással fog folytatódni. Ez egyértelműen spektrális torzuláshoz fog vezetni.[9.]

Ezen problémára a megoldást az jelentené, ha a két jel nem összeérő végét „össze tudnánk mosni” úgy, hogy a jel végét egy függvénnyel megszorozva az közel 0-t adjon. Erre jöttek létre az ablakfüggvények, melyekről a 4.1. ábrán láthatunk példákat. Ilyen függvények a Hanning, Hamming, Blackman, Háromszög, Flat top. Leggyakrabban ezeket használjuk.



4.2 ábra Ablakfüggvények [9.]

Minden ablakfüggvénynek más és más a karakterisztikája, ebből kifolyólag más és más feladatokra használjuk őket, „pl. közeli frekvencia összetevők elkülönítéséhez Hanning) , zajos jel esetén (nincs ablakfüggvény!).”[9.]

Néhány egyszerűbb ablakfüggvény, melyet én is használtam a feladat során:

$$\text{Hanning:} \quad w[n] = 0.5 - 0.5 \cos \frac{2\pi n}{N}$$

$$\text{Hamming:} \quad w[n] = 0.54 - 0.46 \cos \frac{2\pi n}{N}$$

(4.4)

4.1.3. Konvolúció (szűrés)

Jelfeldolgozás során akadnak olyan esetek, hogy az transzformált jelben megjelennek számunkra felesleges „adatok”, melyeket mi feldolgozni nem akarunk, de a további munkát nehezíti. Ilyen például a DFT során történő fodorozódás. Ennek oka, az hogy a mi a transzformálni kívánt függvényünket egy teljes valós számsíkon értelmezett és végtelen hosszúnak tekintett valamint egy véges hosszú ablakfüggvény szorzataként állítjuk elő. Az időtartományban végzett szorzás a Fourier tartományban a tagok Fourier transzformáltjai konvolúciójának felel meg. Ezt szem előtt tartva az ablakfüggvényünk Fourier transzformáltja, mely egy $\sin(x)/x$ típusú mennyiség, rásuperpolálódik a jelünkre, mely okozza a felesleges adatokat és fodorozódást. Ezeket a fodorokat valahogyan le kell szűrni a jelről. Ez Fourier tartományban konvolúciót jelent. Az $x[n]$ és $h[n]$ értéksorozat konvolúcióját a következőképpen írjuk le:

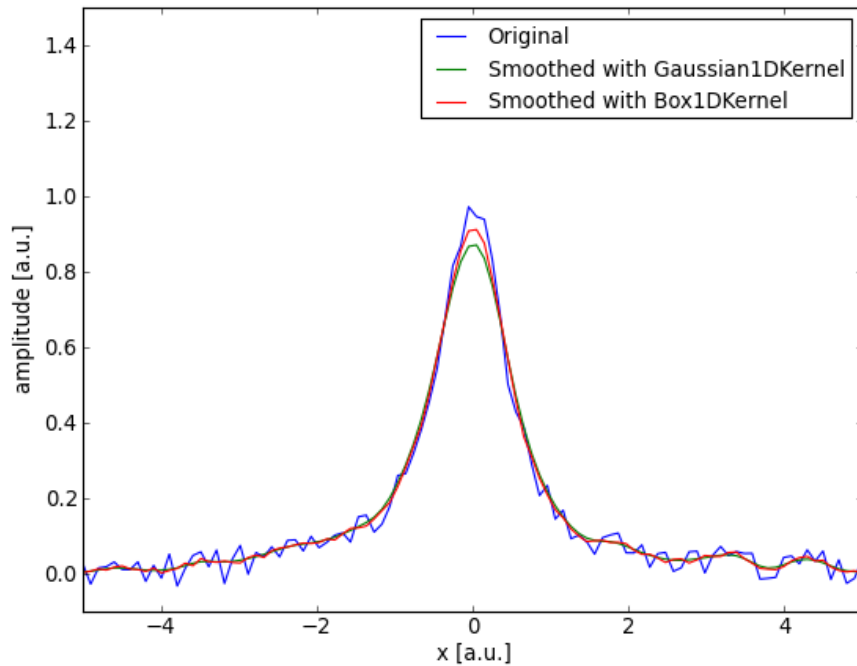
$$y[n] = x[n] * h[n]$$

(4.5)

$$y[n] = \sum_{i=0}^n x[i] * h[n-i]$$

Konvolúciós szűrésnél rettentő fontos a jól kiválasztott függvény, mellyel konvolálunk. A Gauss-függvényel vett konvolúció, időtartományban egy Gauss-függvénnyel vett „ablakolást” eredményez, mivel a Gauss-függvény Fourier transzformáltja is Gauss-függvény, mely előállítható az eredeti függvényből egy lineáris transzformációval. Ez eredményül elmossa a függvényt, ám nem keletkeznek rossz maximumok, melyek nekünk később rettentő fontosak

leszek [4.2. ábra]. A maximum hely lokalizációnál viszont számolni kell azzal, hogy ezzel az eljárással a maximum helyek esetleg eltolódhatnak.



4.3 ábra Az eredeti jel és a Gauss-függvénnyel valamint Boks-függvénnyel vett konvolúciója [10.]

4.2. Folytonos idejű feldolgozási módszerek

4.2.1. Gördülő energia

Tisztában vagyunk azzal, hogy minden analóg jelnek vannak fizikai, mérhető paraméterei. Az egyik ilyen paramétere egy hullámnak a gördülő energiája, melyet úgy képzelhetünk el, hogy a kiadott hanghullám adott időpillanatban hordozott abszolút energiamennyisége.

Ezt a gördülő energiát úgy tudjuk kiszámolni egy adott időpillanatra, hogy attól az időpillanattól visszafele egy meghatározott időablakban vesszük a jel négyzetes integrálját.

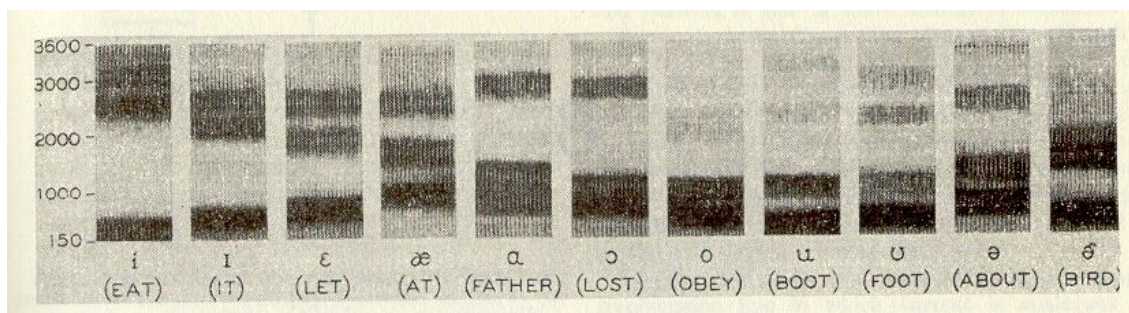
$$E_n = \sum_{m=n-N+1}^n S^2(m)$$

(4.6)

Ez diszkrét jeleknél persze ez csak az aktuális értékek négyzetének az összegzését jelenti.

A választott időablak mértéke nagyban függ attól, hogy pontosan mit is akarunk lokalizálni a beszédjelben. 5-10 ms-os időablakkal tökéletesen lehet nyomon követni egy hangzón belüli energiaváltozást, a 20-30 ms-os időablak a zöngés/zöngétlen detektálására elegendő, ennél nagyobb, 50-100 ms-os időablakkal pedig a beszéd kezdetét végét, beszédszünetet vagy szavak közötti időrés lokalizálását végezhetjük el.[5] Ez persze már egy olyan nagyságú változó, ami nyelvtől és embertől függ. Mindnél gyorsabb beszédet elemzünk, ezek az időablakok annál kevésbé működnek és annál kisebbre kell őket állítani. Ezen felül csak a gördülő energiával nem lehet zöngés/zöngétlen hangot felismerni, viszont tökéletes a magánhangzók lokalizálására, mivel ahogy már előzőleg említettem, a magánhangzók hordozzák a legtöbb energiát a beszédben, így egy gördülő energiafüggvényben élesen kicsúcsosodik a függvény ott ahol magánhangzót ejtünk.

Az előzőnél kicsit számításigényesebb módszer a gördülő energiaspektrum, mely szinte ugyanaz, mint a gördülő energiafüggvény, csak itt eredményül nem egy kétdimenziós függvényt, hanem egy háromdimenziós spektrumot kapunk. Ennél a módszernél is ablakban számolunk, csak itt Fourier transzformáltat, ebből kifolyólag az ablak méretét érdemes 2 hatványaként választani. Ezen módszer eredményében már közelebb kerülünk a hangfelismeréshez, mivel itt már szépen kirajzolódnak a magánhangzók formánsai, ám a spektrumot sem árt szűrni. A 4.3. ábrán látszik néhány angolul ejtett magánhangzó spektrogramja, melyeken látszik, hogy milyen hosszú ideig, milyen erősen ejtették őket.



4.4 ábra Angolul ejtett magánhangzók spektrogramja [12.]

4.2.2. Lineáris predikció

„Mint már korábban tárgyaltuk, a formáns a beszédspektrum helyi maximuma. A spektrum számítására alkalmazhatunk gyors Fourier-transzformációt (FFT), de ez egy számos maximumot és minimumot tartalmazó függvényt állít elő. Egy ilyen spektrum csúcsait nehéz

megbízhatóan meghatározni. Formáns-meghatározáshoz a lineáris predikción alapuló spektrumszámítás az elterjedtebb, mert számos előnnyel rendelkezik: A predikció fokszámanak (azaz az átviteli függvényben szereplő pólusok számának) változtatásával határozhatjuk meg a spektrum „felbontását”. A lineáris predikció a (formáns-detekció szempontjából érdekes) csúcsoknál közelíti legpontosabban a spektrumot. Még rövid szakaszok elemzésekor is elfogadható eredményeket ad. A lineáris predikció együtthatóiból becsült átviteli függvény:

$$H(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}} \quad z = r * e^{j2\pi \frac{f}{f_0}}$$

(4.6)

Az eredmény egy olyan spektrumközelítés, amely jóval simább, mint az FFT, de a formánsokat nem torzítja el.” [11; 11.-12.o]

5. Magánhangzó-felismerő program

Az előző fejezetben a beszédjelfeldolgozás matematikai hátterével foglalkoztam, most bemutatom az általam készített program algoritmusait, működését, elvi hátterét.

Mielőtt bele kezdenék a program ismertetésébe, figyelembe kell venni, hogy a program még mindig fejlesztés alatt áll és gyakorlati alkalmazása, mely valós idejű meghatározásra képes. A program teljes egésze saját fejlesztés, valamint az egész kalibrációt is én végeztem.

A jelenleg szemléltetett program három fő részből áll. Az első része a felvevő rész, mely gyakorlatilag elkülönítve működik a felismerőtől. A második és harmadik része már szorosan kapcsolódnak egymáshoz, először a magánhangzó lokalizálás hajtódik végre, majd átadja a vezérlés a magánhangzó felismerésnek. Ezt a három feladatot egy LabView-ban írt programmal valósítottam meg.

A LabView rövid ismertetése után a következő fejezetek mutatják be a programot és annak készítését.

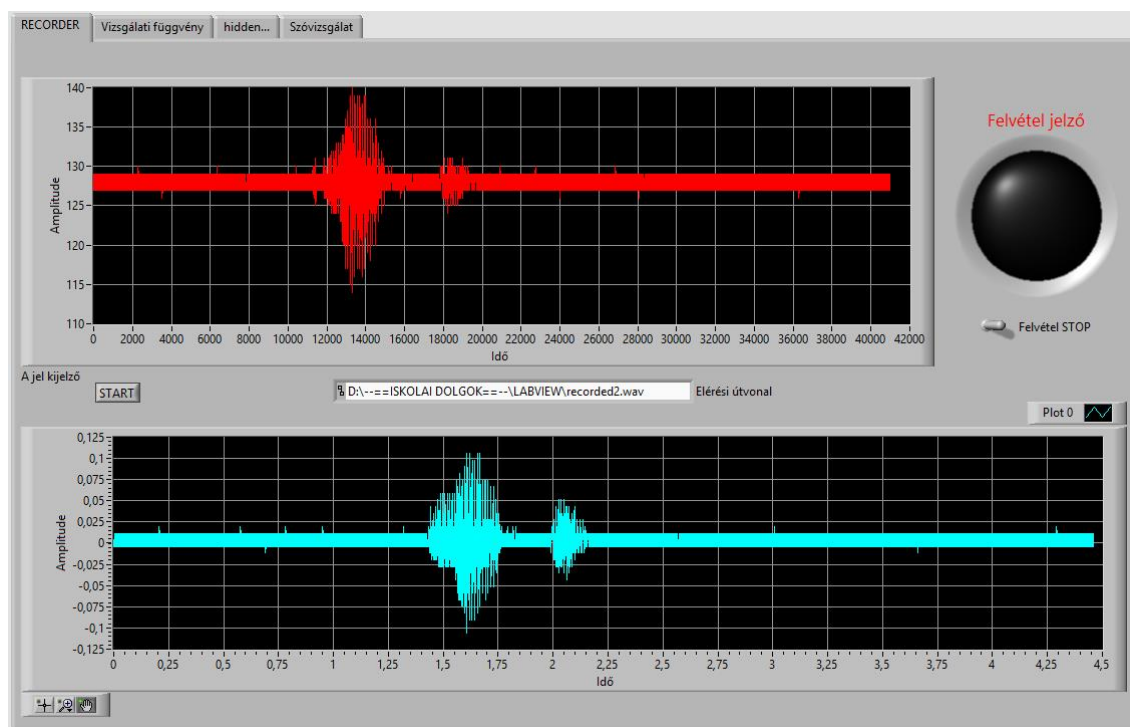
5.1. LabView 2012-ről röviden

„A National Instruments LabVIEW fejlesztői környezet már 27 éve forradalmasítja a tesztelési, mérési és vezérlő alkalmazások fejlesztését. A mérnökök és fejlesztők tapasztalatától függetlenül

rövid idő alatt, költséghatékonyan hozhatnak létre illesztő-felületeket a mérési és vezérlő hardverekhez, elemezhetik a mért adatokat, megoszthatják az eredményeket, és terjeszthetik a rendszereket.” [14; termékismertető]

5.2. A felvevő modul

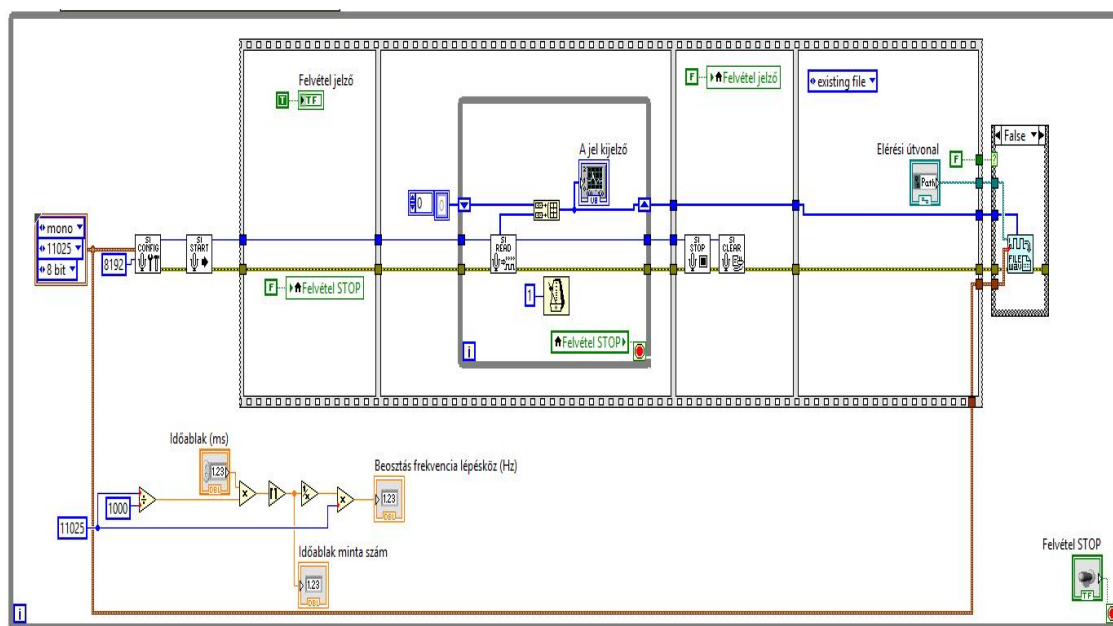
A hang analizálása előtt a legelső teendőnk, hogy digitalizálva rögzítjük az analóg hangunkat. Én a hangot egy GX Gaming Cavimanus Headset-tel rögzítettem, és a program ezen fejhallgató mikrofonjára van kalibrálva. A mikrofon specifikációjáról röviden: érzékenység: -39 ± 3 dB, frekvencia átvitel: 100Hz - 10kHz. A felvevő akármilyen hosszú hanganyagot képes rögzíteni, majd az előre beállított helyre menti. A felvételt a START gomb lenyomása után indul és egy Felvétel STOP billenő-gomb váltására áll le. Amint az az 5.1. ábrán is látszik a piros függvény jelzi az éppen felvétel alatt álló hanghullámot, a kék pedig a STOP lenyomása után, az elmentett és visszaolvasott hanghullám.



5.1 ábra Hangfelvevő modul

Felmerülhet a kérdés, hogy miért lényeges ez a megoldás, hogy kétszer jelenítsük meg akár ugyanazt a jelet. A függvények tengelyein látszó feliratozás adja meg a választ a kérdésre,

mi szerint a wav-ban való tárolás nem feltétel ugyanazokkal a számadatokkal történik, mint amilyen nagyságú értékeket mi rögzítettünk. Mivel szerettem volna, hogy a beszédfelismerésnek ne legyen feltétele a hangrögzítés, így célszerű az eltárolt hangot vizsgálni. Az rögzítendő hangnál a kijelző milliszekundumonként egy 8 bites egész számot reprezentál nekünk 0-255ig, míg a visszaolvasottban lebegőpontos értékeket látunk már rendes időegységben. A felvétel formátumának 11,025kHz-t választottam, mono csatornát és 8 bites mintavételt [5.2. ábra]. Ez persze tetszőlegesen változtatható, de tapasztalatom szerint ez a minőség már tökéletesen megfelel a célra és kisebb a feldolgozás idő- és háttértárigénye.

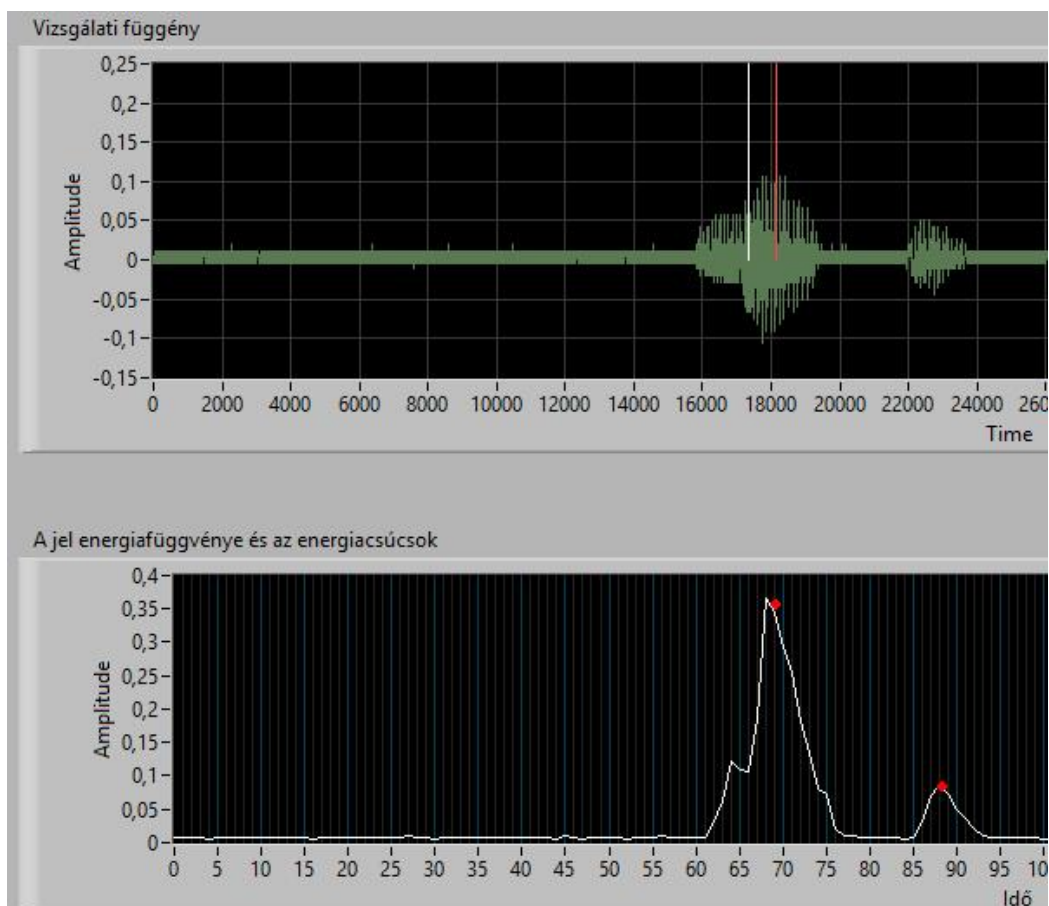


5.2 ábra Hangfelvevő blokk-diagrammja

5.3. Magánhangzó lokalizáló

A 4.2.1-es fejezetben tárgyalt módon, a gördülő energiafüggvény vizsgálati módszerét továbbgondolva alkottam meg egy saját, magánhangzó lokalizálásra alkalmas módszert. Én 23 ms-os időablakkal dolgoztam, úgy véltem az tökéletes lesz a csúcsok keresésére. Először ki kell számolni, hogy a felvételi tömbünk 23ms alatt hány értéket tárol el. Ezt megkapjuk a mintavételi frekvencia kHz-ben (11,025) és az időablak ms-ban való szorzata, ezzel megkapjuk hogy 23ms alatt a tömbben kerekítve 254 elemet tárolunk. Az energiafüggvény leképzését úgy végezzük, hogy ezt a 254 mintaszámos „ablakot” léptetjük a tömbben és minden ablakban lévő elemet kivágunk egy melléktömbbe. Ezt a LabViewban az Array Subset funkcióval lehet megtenni. Egyszóval az eredeti tömbünket szétvágjuk 254 darabonként, majd minden egyes al-

tömb - melyet a programban „Integrálási Időablak Tömb”-nek hívjuk - négyzetét integráljuk. Eredményül minden egyes Integrálási Időablak Tömb egyetlen pozitív számként jelenik meg. Ha ezeket a számokat sorba állítjuk egy tömbbe, majd ábrázoljuk, megkapjuk a felvett hangunk gördülő energiafüggvényét. Az 5.3. ábrán láthatjuk a „lámpa” szó időfüggvényét és alatta a gördülő energiafüggvényét.



5.3 ábra A lámpa szó időfüggvénye vizsgálati időablakkal, alatta a gördülő energiafüggvénnyel, a lokalizált magánhangzókkal

A használt numerikusintegrálás a LabViewban a Numeric Integration, melyben a $dt=1$ -el és trapéz szabállyal integráltam.

Az energiafüggvényt felrajzolva gyönyörűen látszanak a magánhangzók a csúcsokban. Az algoritmus a következőkben megkeresi ezeket a magánhangzókat és megállapítja, hogy milyen magánhangzók ejtettünk ki. A keresési eljárásnál a LabView csúcskeresőjét, azaz Peak Detector (obs2010).vi-t alkalmazzuk. Ez a subVi gyakorlatilag egy bizonyos szélességen belül megkeresi a csúcsot vagy csúcsokat és visszaadja annak X és Y tengelybeli értékét. Bölcs dolog

beállítani egy bizonyos Threshold-ot, határértéket, mely alatt a Peak Detector nem értelmez, ez nálam 0,01.

A magánhangzó detektálónk jelzi is, hogy a talált magánhangzóink száma mennyi, jelen esetben a „lámpa” szónál 2. Az 5.3. ábrán látszik, hogy a program piros ponttal jelöli a talált magánhangzókat. A jól kalibrált Peak Detector, ahogy az ábrán is látszik, figyelmen kívül hagyja a kevésbé érvényes csúcsokat. Az első piros pont előtt van egy csúcs, melyet a Peak Detector nem jelölt ki. Arról persze mi tudjuk hogy a „lámpa” szavunkban az „L” hangzó felpendülése .

Eredményül a Peak Detector két tömböt adott vissza, a „Magánhangzók időablak indexei” tömböt, melyben a gördülő energiafüggvényben talált csúcsok X tengely értékeit találjuk, valamint az „Energia” tömböt, melyben ezen csúcsok X értékeit találjuk. X tengelyi értékeknél figyelembe kell venni, hogy ezek csak index értékek; hogy viszonyíthassuk az eredeti jelben vett helyüket, vissza kell szorozni őket, az integrálási időablak mintaszámával. Ez a művelet megadja, hogy a felvételen a talált csúcsok pontosan hány ms-nál hordozták a legnagyobb energiát. A példában említett első magánhangzó X értéke, indexe 69.2, ezt visszaszorozva 254-el a mintaszámmal, - láthatjuk az 5.3. ábrán, hogy az „á” hangunk a „lámpa” szóban - körülbelül 17.6 másodpercnél hordozza a legnagyobb energiát.

Ez a program ebben a formában egy könnyen kalibrálható vizsgálati módszert ad számunkra, és ami talán a legfontosabb, hogy akármilyen hosszan felvett szövegre alkalmazható, hiszen az algoritmus végigiterál minden egyes csúcson és megvizsgálja az aktuális magánhangzót; majd ha végzett az összes csúccsal, leáll.

5.4. Formáns kereső

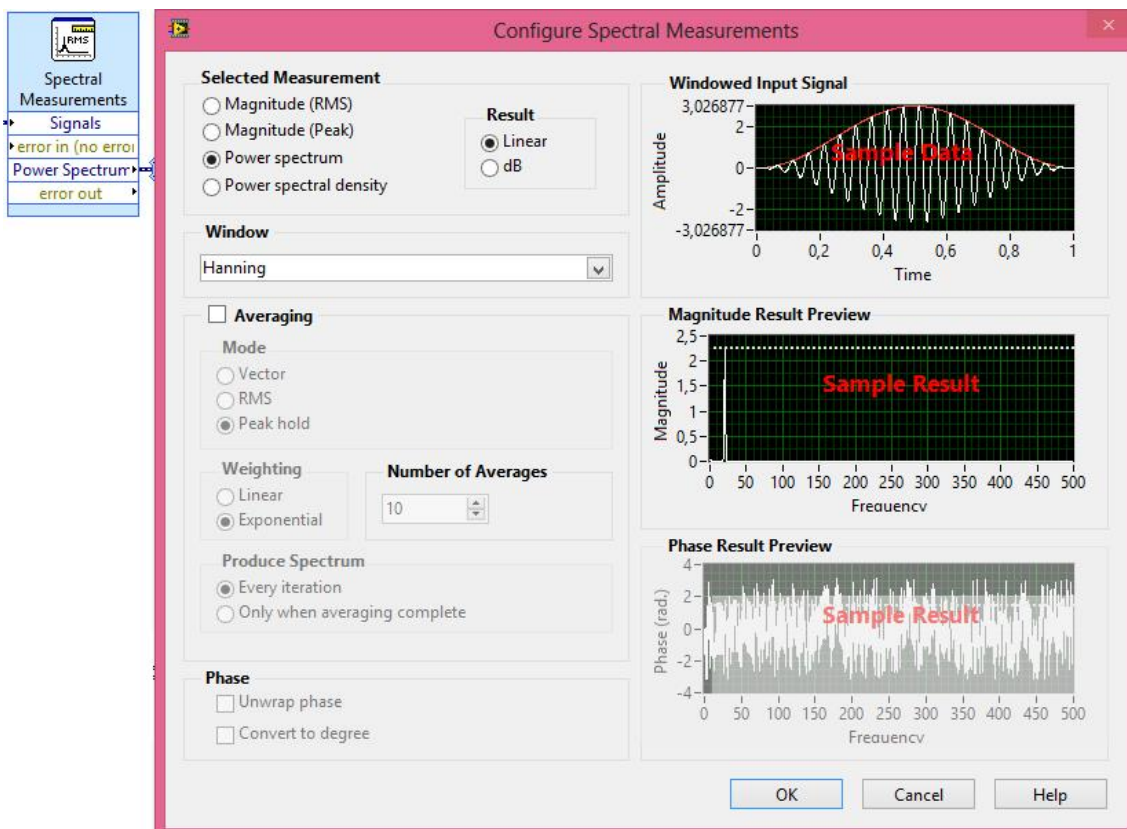
A gördülő energiafüggvény vizsgálatával szinte milliszekundum pontossággal meghatározta az algoritmus a rögzített beszédhangban kiejtett magánhangzókat.

Ebben a fejezetben bemutatom, hogy milyen algoritmussal alkottam meg a magánhangzókon alkalmazott formánskeresőmet.

Az 2.2. fejezetben tárgyalt beszéd felismerési metodikák túlságosan bonyolultak egy egyszerű projektben való alkalmazáshoz, ezért egy saját súlyozáson alapuló rendszert dolgoztam ki. A metodika első fázisánál, az előzőleg meghatározott csúcsok időpontjainál a rendszer a beszédjel függvényén egy időablakot képez, melyet a későbbiekben FFT időablaknak fogok nevezni. Ez az időablak látható az 5.3. ábrának a felső függvényén, ahol

megjelenik egy fehér és piros függőleges vonal. Ezek a vonalak az FFT időablak határai. Felmerül a kérdés, hogy ezen időablak milyen szélességen működjön. Gósy Mária és Beke András tanulmányában, a Magánhangzó-időtartamok a spontán beszédben [7] szó kerül, hogy az átlag magyar emberek milyen gyorsan beszélnek és mennyi ideig ejtik a magánhangzókat. Ezen mérési adatok alapján, de legfőképpen saját méréseim alapján én a saját beszédhangomhoz 70 ms-t állítottam be. Ez persze sok személynél különbözik és abszolút szubjektív, ezért a programban ez dinamikusan állítható. - Itt jegyezném meg, hogy a program ezen részét a későbbiekben fejleszteni fogom, hogy alkalmazkodjon az ember beszédgyorsaságához és ezt ne manuálisan kelljen beállítani. - Ez a 70 ms már elég arra, hogy a hosszan ejtett magánhangzók lényegi része benne van és a röviden ejtett vagy éppen csak lokalizált magánhangzók egésze pedig belefér.

Miután meghatároztuk az FFT időablakunk nagyságát a programban a Gyors Fourier Transzformációs lépés következik. Ez a 4.1.1. fejezetben leírt matematikai módon történik, a LabViewban pedig több lehetőség is van az algoritmusra. Én a Spectral Measurements funkciót alkalmaztam a GUI felületen való könnyű konfigurálhatóság miatt [5.4. ábra]. Az algoritmusban az FFT időablakkal kivágott függvényrészleten alkalmazott FFT, Hanning ablakolást használ az összemosásra és erőspektrumot ad végeredményül.

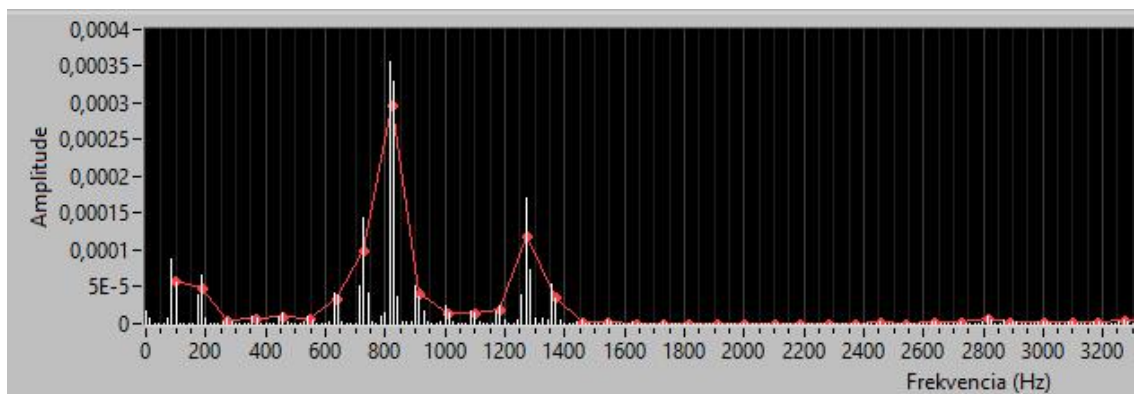


5.4 ábra Spectral Measurements funkció és konfigurálása LabViewban

Ezt az eredményhalmazt ábrázolva a Fourier Transzformáció spektrumát kapjuk meg. A pontos értelmezéséhez tisztában kell lenni a paraméterekkel. Ahogy a gördülő energiatartomány számoltuk a ms-ban megadott időablak hosszát mintaszámra, úgy ezt itt is megteszi a program. 70ms-os időablakba, 11025Hz frekvenciánál 772 darab minta fér bele. Ezt a mintamennyiséget Fourier transzformáltuk. 11,025kHz-es frekvencia és a 772 mintaszám hányadosából megkapjuk, hogy a megjelenített FFT spektrumon, az X tengelyen megjelenített indexek frekvenciában vett lépésköze 14,28Hz. Ez elég részletes felbontást eredményez, és ahogy az 5.2. fejezetben említettem, hogy a 11kHz-es mintavételi frekvenciát elégnek éreztem, itt válik lényegessé. Vizsgálni ugyanis a mintavételi frekvenciánk felét tudjuk csak, így maximum 5,5kHz-es frekvenciájú hangokat tud megjeleníteni a program, ám ez nem probléma, ugyanis a 3.5. fonématáblázatot vizsgálva látható, hogy az „í” magánhangzó harmadik formánsa is csak 3,1kHz környékén rezonál. Az 5,5kHz maximum vizsgálati frekvencia a magánhangzó felismerésnek tökéletes, ha csak az első három formánst akarjuk vizsgálni.

A paramétereket ismerve már tökéletesen ábrázolható az Fourier spektrumunk. Az 5.5. ábrán láthatjuk a „lámpa” szóból kivágott és Fourier transzformált „á” hangot. A spektrum elején megfigyelhető a Fourier spektrumra jellemző DC jel, valamint 100Hz-nél lévő csúcs.

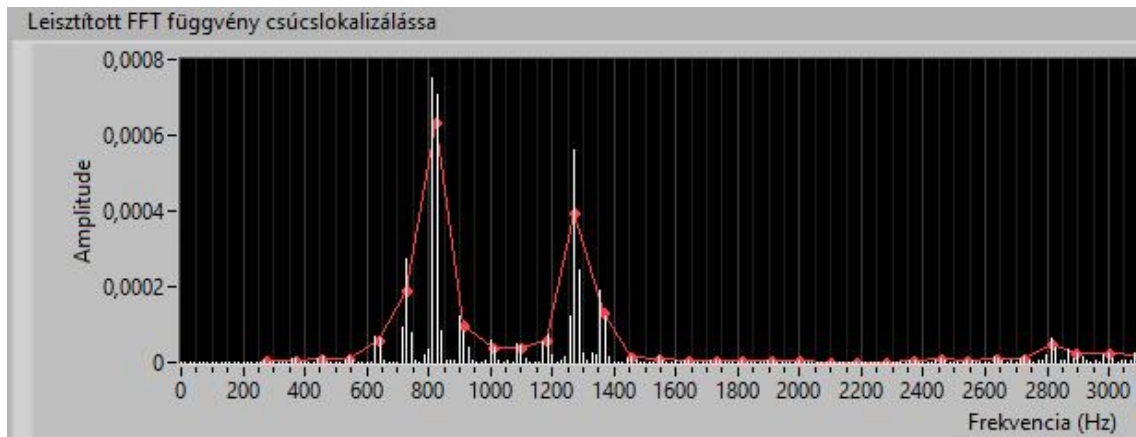
Mivel szüretlen hangot Fourier transzformál a program, ezért a fehér zajok is feldolgozásra kerülnek. A fehér zajok kiszűrése rettentő nehéz, ezért én egy számomra egyszerűbb megoldást választottam ennek megoldására, egyszerűen nem veszem figyelembe a 200Hz alatti hangokat. Ez a megoldás elsőként durva megoldásnak tűnhet, de ha átgondolva a magánhangzó felismerést nem hátráltatja, mivel 200Hz alatti formánsú magánhangzót nem ejtünk ki.



5.5 ábra Az „á” hang Fourier transzformáltja

Az ábrázolás után a program egy nagyon illeszkedő burkológörbét húz a spektrumra. Ezt a Peak Detector oldja meg, miképpen nagyon érzékeny és szűk keresztmetszetű időablakkal dolgozik így szinte minden csúcsot érzékel, majd az eredményül kapott, csúcsokat reprezentáló tömbökből egy függvényt készítve ráültetjük a spektrumra. Így jön létre az 5.5. ábrán látható piros burkológörbe. Sok tesztelés után sajnos rá kellett jönnöm, hogy a hangzóban lévő első 2 formáns szépen kirajzolódik az FFT spektrumban, de a harmadik már szinte alig. Ezért egy exponenciális erősítést állítottam be a spektrumra, mi szerint mindnél nagyobb frekvenciájú az FFT spektrumon megjelenítendő érték, annál nagyobb értékkel szorzódik meg, így a jól kivehető első és második formáns szinte erősítés nélkül jelenik meg, mint csúcs; ám a nehezen megjeleníthető harmadik formáns már szépen kicsúcsosodik. Ez azért lesz fontos, mert a burkológörbére ráültetett saját Peak Detectorom könnyebben és pontosabban tudja lokalizálni a csúcsokat, és a formánsokat.

Az 5.6. ábrán látható, már leszűrt és felerősített FFT spektrum. Ezen a függvényen szépen látszik már az „á” hangom három formánisa 800Hz, 1300Hz és 2800Hz-nél. A függvény, formánserősség meghatározására már alkalmatlan, mert torz, viszont a programnak csak a csúcsok pontos helyére van szüksége az X tengelyen.



5.6 ábra Az „á” hang leszűrt és felerősített FFT spektruma

A burkológörbét vizsgálva, a saját csúcsmeghatározóm végigmegy a burkolófüggvény értékhalmozán és megkeresi az első csúcsot, azaz az első számot ami után a következő néhány szám kisebb. Ennek a csúcsnak az X tengelyi értéke lesz a hang első formánsa. A program továbbiterál és megkeresi az első vermet vagy másképpen mondva lokális minimumot, ami után a függvény elkezd növekedni. Ettől a veremtől fogva az algoritmus szétvágja a burkolófüggvényt két részre és a még nem vizsgált részben megkeresi a legnagyobb számot. Az exponenciális-erősítés miatt, ez nem lehet más csak a második formáns csúcsa, mivel a harmadik formáns arányaiban sokkalta gyengébb csúcsban emelkedik ki. A második csúctól kezdve ismét vermet keres a program, majd a maradék burkolófüggvényt is kettévágja és megismétli az előzőekben leírt keresést. Így találja meg mind a három csúcsot, mely a három formánst reprezentálja.

5.5. Magánhangzó meghatározó

Ahogy az előző fejezetben is említettem, a magánhangzó felismerésnek a metodikáját és saját magam által kitalált metodikával oldottam meg. A metodika harmadik lépcsőfoka maga a magánhangzó felismerése, mely maga után von pár szubjektív problémát.

Az akusztikai hangfelismerésből adódóan a programnak három formánsfrekvenciából kell eldöntenie, hogy milyen magánhangzó van a felvett hangban, szótár és mindenféle szemantika vagy szintaktika nélkül. Én ezt súlyozással oldottam meg, mivel a fonématablázatot nézve a legtöbb fonéma határai átlógnak egymásba. Egyszóval a mérés rendelkezésére álló 5,5kHz-t a fonémák nem osztják fel teljesen, hanem inkább kombinációkat alkotva lehet őket megkülönböztetni. Egy olyan súlyozó algoritmust építettem a programba, ami a hangzó három fonémáját együtt kezeli. Először megvizsgálja, hogy az első fonéma melyik magánhangzó első fonémás intervallumába esik bele. Ha beleesik valamelyikbe is, akkor az alapján kap pontot, hogy a meghatározott intervallum milyen széles, mivel egy 140Hz-es intervallumba könnyebben beleesik, mint egy 80Hz-esbe. Az intervallum szélessége fordítottan arányos a kapott ponttal. Második lépésként megvizsgálja, hogy a vizsgált fonéma frekvencia, milyen távol esik a fonéma intervallum frekvencia-középértékétől. Mindnél közelebb esik, hatványozottan több pontot kap érte az adott hangzó. Ez a vizsgálat minden magánhangzóra lefut, független hogy az fonéma érték, nem esik bele a magánhangzó intervallumába. Végezetül mind a két vizsgálat lefut a második és a harmadik formánsfrekvenciára. Az első formáns vizsgálatánál kapható értékek közel dupla olyan nagyok, mint a második formánsnál kapható értékek, és közel hatszorosa a harmadik formánsnál kapható értékeknél. Ennek oka, hogy az első formáns jellemző a legjobban a magánhangzóra és az intervallumok sokkal kisebbek, pontosabban vizsgálhatóak. A harmadik formáns lényegében akkor játszik döntő szerepet, ha első és második formánsban is hasonló magánhangzóról van szó, mint például az „ö” és „ü”, ahol az első kettő formánsaik határai nagyon átfedik egymást, viszont a harmadikban teljesen elkülönülnek. Itt döntő szerepe lesz a kevés pontot adó harmadik formánsvizsgálatnak.

Teszteléseim alapján többségében az első kettő formánsnál már eldől, hogy melyik lesz a kiválasztott magánhangzó, de az esetleg 10%-ban a harmadik formáns pontja döntött. Végértékelésként egy pontozótáblában összegzi a program, hogy melyik magánhangzó hány pontot kapott, majd a legnagyobb pontút kiírja „string”-ként egy indikátorba, későbbi szó keresés céljából. A „lámpa” szóban megtalált első magánhangzó pontozását és a felismert „á” hangot láthatjuk az 5.6. ábrán.

szóponthozó	Magánhangzók tömbje
4,62757	a
50,1822	á
5,05982	e
2,97952	é
12,9916	i
2,10918	í
3,45307	o
2,7363	ó
3,02198	ö
2,67145	ő
2,28435	u
2,10067	ú
2,26424	ü
2,26219	ű

Első formáns (Hz)

824,961

Második formáns (Hz)

1272,06

Harmadik formáns (Hz)

2814,4

Talált magánhangzók

á

5.7 ábra A „lámpa” szóban felismert „á” hang és szóponthozó táblázata

Az 5.4. fejezetben és ebben a fejezetben taglalt vizsgálatok futnak le, minden egyes 5.3. fejezetben leírt módon lokalizált magánhangzóra. A program a talált magánhangzókat egymásba fűzi.

5.6. Kísérleti fázisban lévő szókereső

Célom volt, hogy ezzel a magánhangzó-felismerővel válogatni tudjak előre betáplált szótáblázatok bemondott szavakat. A program ezen része még kísérleti fázisban van.

Ahogy a tanulmány elején leírtam, a beszéd felismerők arra „számítanak”, hogy beszédet fognak hallani, ebből kifolyólag egy olyan szóválasztó modult szeretnék beépíteni a programba, amely egy 10-15 szavas szótárból kiválasztja a legnagyobb valószínűséggel bemondott szót, csak is a magánhangzók alapján. A bemondható szavakat string tárolás mellett egy fajta X-SAMPA jelölésrendszerhez hasonlóan tárolnám el. Ezeket a szavakat elsősorban a magánhangzóik sorozata teszi egyedivé, vagyis alapfeltétel lenne, hogy nem lehet benne két ugyanolyan sorrendű magánhangzóból álló szó, mint pl.: lefele-felfele-befe-lefele. Feltételezve hogy a vizsgálat akár tévedhet is, a szavakhoz kapcsolva lenne egy táblázat, melyben a hozzájuk hasonló magánhangzó-kombinációk találhatók, mint pl.: előre → eőe, eoe, eűe.

Végezetül, ha nem talál semmiféle egyezést új szót kér. Ezen szókeresőnek egy pontosítására szolgálna még az a modul, mely a magánhangzó csúcsok közötti eltelt időt leméri, és statisztikai alapon megmondja, hogy a két magánhangzó közé, mennyi és milyen típusú mássalhangzók „férnek be”. Így a szókereső pontosságát növelni lehetne, és akár több szavas irányítást is lehetne csinálni.

6. Összegzés

Összességében állítható, hogy a program némi kalibrálás után, - mint például saját fonématablázat, időablakok beállítása – igen csak szép eredménnyel ismeri fel a felvett szavakban a magánhangzókat. A tesztelés során közel 90% feletti pontossággal ismerte fel a magánhangzókat, amely egy pusztán akusztikai szinten működő felismerőtől igen csak szép teljesítmény. Akusztikai szinten a fülünk is közel ilyen aránnyal működik, hiszen sokszor mi is félreértjük vagy nem értjük a szavakat, ha nem szöveggörnyezetben halljuk.

Bebizonyosodott azonban, hogy ezt a fajta hangfelismerési rendszert szinte lehetetlen beszélő-függetlennek és nyelv-függetlennek felprogramozni, mivel a kalibrálás szinte csak a beszélő szubjektív paramétereit igényli. Egy minimális automatizálással és adatbázis pontosítással valamint növeléssel talán a felismerő pontosságát is növelni lehet, ám mindig szükséges a felhasználó személy paramétereinek kalibrációja.

Végezetül úgy gondolom, hogy a kísérleti fázis befejezése után, a program képes lesz az eredetileg kitűzött célokat megvalósítani, melyeket a tanulmány bevezetés részében megfogalmaztam, mivel azoknál egy személy beszédhangjához illesztjük a programot, viszont magasabb színvonalú felismerést igénylő feladatok ellátására ilyen szintű hatékonysággal képtelen.

IRODALOMJEGYZÉK

- [1.] VICSI K.: *Beszédatadabázisok a gépi beszédfelismerés segítésére*. Beszédtechnológia és Intelligens Interakciók Laboratórium (látogatva : 2013.12.21) http://speechlab.tmit.bme.hu/beszinfo/beszedadatbazis_info.pdf
- [2.] JOHN WELLS: *Computer-coding the IPA: a proposed extension of SAMPA*. Department of Phonetics and Linguistics, University College London (látogatva : 2013.12.21) <http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm>
- [3.] RUSSELL STUART, NORVIG PETER : *Mesterséges intelligencia*. 2005 Hungarian Translation Panem Könyvkiadó, Budapest
- [4.] DESMUKH, N. and GANAPATHIRAJU, A. and PICONE, J. (1999). Hierarchical Search for Large-Vocabulary Conversational Speech Recognition. *IEEE Signal Processing Magazine* 1999. szeptember: 84-107.
- [5.] GORDOS, G. és TAKÁCS, GY. (1983). *Digitális Beszédfeldolgozás*. Műszaki Könyvkiadó, Budapest
- [6.] PAOLO PRIORE, DAVID DE LA FUENTE, RAU'L PINO, JAVIER PUENTE : *Dynamic scheduling of flexible manufacturing systems using neural networks and inductive learning*. ISSN: 0957-6061 (látogatva : 2013.12.28) <http://www.emeraldinsight.com/journals.htm?articleid=850973&show=html>
- [7.] MOLNÁR, J. (1986). *A Magyar Beszédhangok Atlasza*. Tankönyvkiadó, Budapest.
- [8.] PTE PMMIK : *Digitális kép hangfeldolgozás*. Fourier transzformáció (látogatva : 2013.12.24) http://e-oktat.pmmf.hu/kepeshang_2_fejezet
- [9.] PTE PMMIK : *Digitális kép hangfeldolgozás*. Ablakoló függvények (látogatva : 2013.12.24) http://e-oktat.pmmf.hu/kepeshang_3_fejezet
- [10.] THE ASTROPY DEVELOPERS : *Convolution Kernels*

(látogatva : 2013.12.29) <http://docs.astropy.org/en/latest/convolution/kernels.html>

[11.] BÖHM TAMÁS, NÉMETH GÉZA (2006/8). *Algoritmus formánsok követésére, módosítására és szintézisére*, BME Távközlési és Médiainformatikai Tanszék

[12.] DAVID R. HILL : *A Conceptionary for Speech & Hearing in the Context of Machines and Experimentation* 2001. (látogatva : 2013.11.27)
<http://pages.cpsc.ucalgary.ca/~hill/papers/conc/body.htm>

[13.] SMITH, S. W.: *The Scientist and Engineer's Guide to Digital Signal Processing*, California Technical Publishing, 1997.

[14.] <http://hungary.ni.com/labview>

[15.] CSÁNYI VILMOS: [*Az emberré válás etológiai rekonstrukciója*](#) (pdf). Er. kiadás: Vince kiadó, 1999.

[16.] HORVÁTH LÁSZLÓ: *Funkcionális anatómia*, Tankönyvkiadó, Budapest, 2004.

[17.] GÓSY MÁRIA, BEKE ANDRÁS : *Magánhangzó-idtartamok a spontán beszédben*.
MTA Nyelvtudományi Intézet ELTE Fonetikai Tanszék
Magyar Nyelvőr 134. évfolyam 2. szám, ISSN 1585-4515
(látogatva : 2013.12.30) <http://www.c3.hu/~nyelvor/period/1342/134202.pdf>