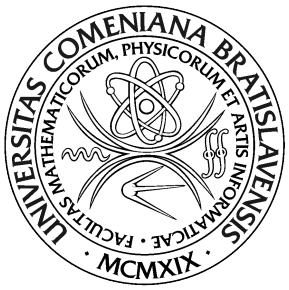
UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY

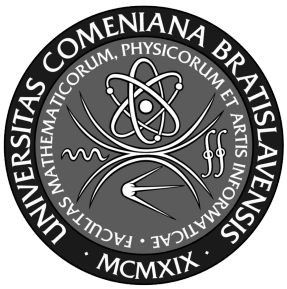



# URČOVANIE POZÍCIE PRÍZVUKU SLOV VO ZVUKOVEJ NAHRÁVKE

Diplomová práca

2022 Bc. Tatiana Gyurcsovicsová

UNIVERZITA KOMENSKÉHO V BRATISLAVE
FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY




# URČOVANIE POZÍCIE PRÍZVUKU SLOV VO ZVUKOVEJ NAHRÁVKE

Diplomová práca

| | |
|---|---|
| Študijný program: | Aplikovaná informatika |
| Študijný odbor: | 2511 Aplikovaná informatika |
| Školiace pracovisko: | Katedra aplikovanej informatiky |
| Školiteľ: | RNDr. Marek Nagy, PhD. |

Bratislava, 2022 Bc. Tatiana Gyurcsovicsová

Čestne prehlasujem, že túto diplomovú prácu som vypracovala samostatne len s použitím uvedenej literatúry a za pomoci konzultácií u môjho školiteľa.

. . . . . .

Bratislava, 2022 Bc. Tatiana Gyurcsovicsová

# Poďakovanie

# Abstrakt

Cieľom práce je vytvoriť algoritmus pomocou octave(matlab) aplikácie. Na vstupe je nahrávka reči (ideálne v slovenčine), ktorá bude na výstupe anotovaná. Vyznačené budú jadrá slabík s príznakom prízvuku.
Kľúčové slová:

# Abstract

The goal is to create an algorithm in octave(matlab) aplication. The algorithm gets recording of speech (in slovak language) and returns it anotated. The syllable core will be anotated.
Keywords:

# Obsah

# Kapitola 1

# Úvod

Prízvuk je kvalita reči, ktorá odlišuje niektoré slabiky prúdu reči od iných slabík. V slovenskom jazyku je prízvuk typicky na prvej slabike slova. Okrem hlavného prízvuku poznáme v slovenčine aj vedľajší prízvuk, ktorý sa objavuje v zložených slovách. Prízvuk spoznáme ako výraznejšie vyslovenie slabiky.

Cieľom práce je vytvorenie algoritmu v aplikácií octave. Na vstupe dostane algoritmus nahrávku reči v slovenskom jazyku. Na výstupe bude nahrávka anotovaná. Algoritmus vyznačí jadrá slabík s príznakom prízvuku.

# Kapitola 2

# Motivácia

V tejto kapitole popíšeme

# Kapitola 3

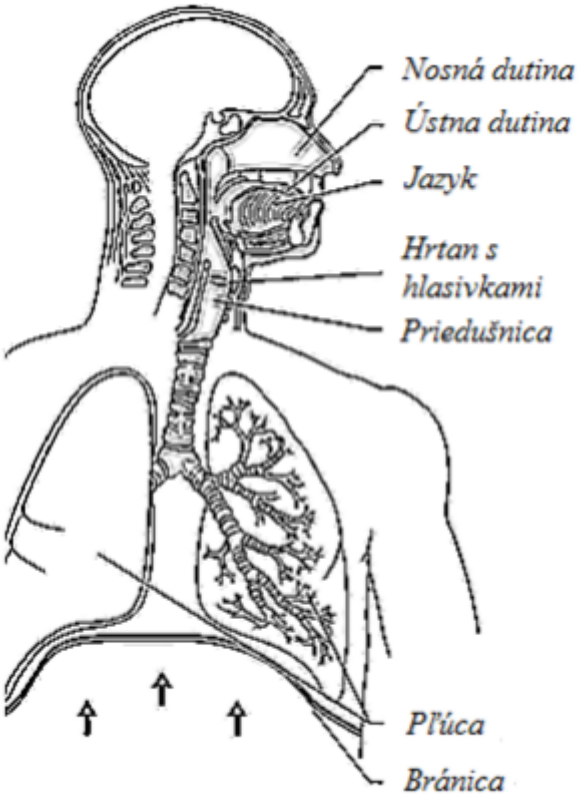
# Teoretické pozadie

V tejto kapitole si popíšeme teoretické pozadie pre našu práce. Vysvetlíme si, ako vzniká reč. Pipíšeme si čo sú to formanty. Rovnako si popíšeme aj matematické metódy, ktoré využívame pri spracovaní nahrávok.

## 3.1 Reč

V našom tele vytvára mozog reč pomocou skupiny rečových orgánov, ktoré tiež nazývame artikulátory. Základné funkcie týchto orgánov sú rôzne, aj napriek tomu, že sa podieľajú na vytváraní reči. Ak sa na tieto orgány pozrieme z pohľadu tvorby reči, tvoria spolu hlasový trakt. Hlasový trakt delíme na tri časti:

- dychové ústrojenstvo
- hlasové ústrojenstvo
- artikulačné ústrojenstvo




Obr. 3.1: rečové orgány

Na obrázku 3.1 vidíme rečové orgány tak ako sú rozmiestnené v ľudskom tele. Tieto orgány delíme na tri časti a to na dychové ústrojenstvo, hlasové ústrojenstvo a artikulačné ústrojenstvo.

### 3.1.1 Dychové ústrojenstvo

Dychové ústrojenstvo je umiestnené v hrudnom koši a tvoria ho nasledujúce orgány: dýchacie cesty, pľúca, bránica

Táto časť rečového trakru je základným zdrojom energie pre reč. Vzduch ktorý vydychujeme sa odvádza cex priedušnicu a hrtan, následne sa modifikuje a z pier je do okolia vypustený ako rečový signál.

### 3.1.2 Hlasové ústrojenstvo

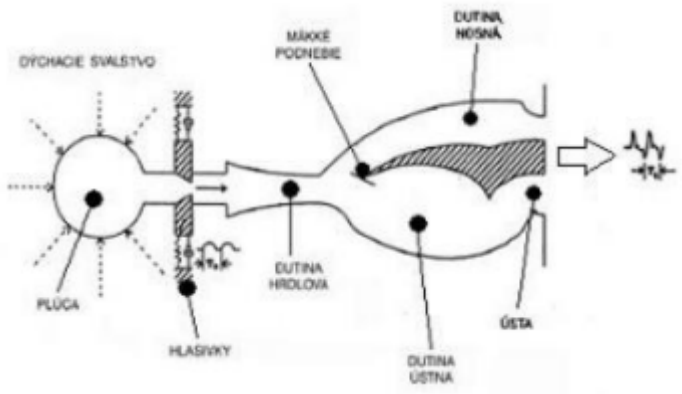
Hlasové ústrojenstvo sa nachádza v hrtane a tvoria ho nasledujúce orgýny: hrtan, hlasivky.

Najdôležitejšou časťou hlasového ústrojenstvo sú hlasivky nachádzajúce sa v hrtanovej dutine. Vydychovaný vzduch sa z pľúc cez priedušnicu dostáva do hrtanu. V hrtane sa tomuto vzduchu do cesty stavajú hlasivky a úplne mu uzavrú cestu. Hlasivky sa pod tlakom vzduchu rozkmitajú a začnú sa striedavo otvárať a prudko uzatvárať. Týmto procesom prichádza k striedanu hustejšieho vzduchu s redším vzduchom, pričom vzniká zvuková vlna, ktorú vnímame ako zvuk. Tieto periodické vzukocé impulzy považujeme za základ zvuku reči.

### 3.1.3 Artikulačné ústrojenstvo

Artikulačné ústrojenstvo tvoria nadhrtanové dutiny a artikulačné orgány: dutiny nosné, dutiny hrdelné, dutiny ústne, pery, tvrdé podnebie, mäkké podnebie, pery, jazyk, zuby, čeluste.

Toto ústrojenstvo nám umožňuje vytvárať rôzne zvuky. Vyššie spomenuté dutiny napomáhajú pasívne pri vytváraní reči. To je spôsobené tým, že sa nepohybujú. Zvyšné orgány sa podielajú na vytváraní reči aktívne.




Obr. 3.2: Zjednodušený model vzniku reči

Na obrázku 3.2 môžeme vidieť zjedodušený model vzniku reči. Kde zľava do prava prúdi vzduch z pľúc cez jednotlivé rečové orgány a postupne sa formuje reč, zrozumiteľná pre ľudí.

## 3.2 Hlasivkový tón

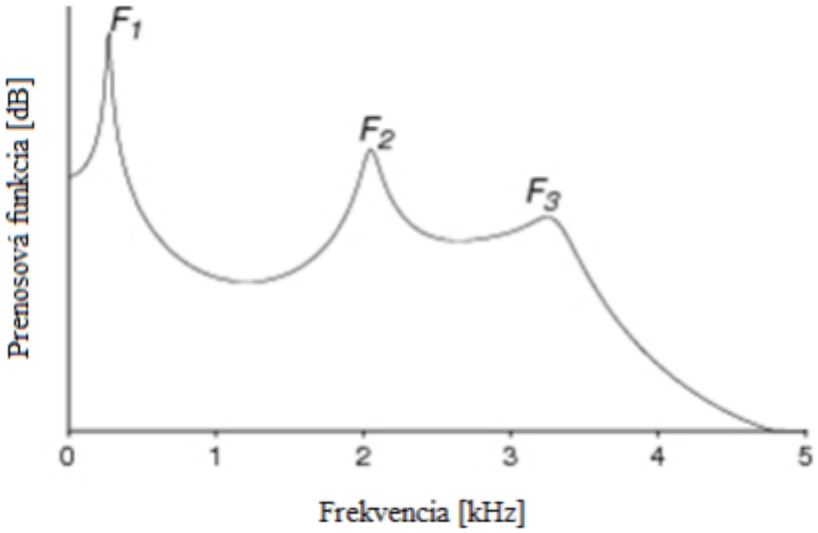
toto je asi $F_0$ skús googliť

## 3.3 Formant

Formanty boli definové prvýkrát Gunnarom Fantomom v roku 1960. Jeho definícia znela nasledovne: „Spektrálne vrcholy zvukového spektra sa nazývajú formanty.“ Formanty predstavujú rezonančné frekvencie ľudského vokálneho traktu. Formanty sa nachádzajú v znelých rečových jednotkách. Znelými rečovými jednotkami sú samohlásky a znelé spoluhlásky.
Ako nultý formant $F_0$ označujeme fundamentálnu frekvenciu. Tá predstavuje frekvenciu kmitania hlasiviek a je to fyzikálna charakteristika rečového signálu. $F_0$ odpovedá výške hlasu tak, ako ju vníma poslucháč. Môže nadobúdať

hodnoty medzi 80 - 160 Hz u mužou, medzi 150 - 300 Hz u žien a nakoniec medzi 200-600 Hz u detí.

Pre analýzu reči sú však dôležité prvé tri formanty $F_1$, $F_2$, $F_3$. Prvý formant $F_1$ sa považuje za rezonančnú frekvenciu hrdelnej dutiny. Druhý formant $F_2$ sa považuje za rezonančnú frekvenciu ústnej dutiny. Tretí formant $F_3$ sa považuje za rezonančnú frekvenciu nosnej dutiny.




Obr. 3.3: Prvé tri formanty

Na obrázku 3.3 vidíme prvé tri formanty, ktoré zohrávajú dôležitú úlohu pri analýze zvuku.

Pri vyslovení spoluhlások a nosových samohlások, v písanom jazykoch sú označené tildou alebo nožičkou, prichádza kvôli vplyvu nosnej dutiny k vzniku antirezonančných frekvecií, vznikajú antiformanty. V slovenčine sa takéto samohlásky nenachádzajú.

## 3.4 Korelácia

Korelácia sa používa pri spracovaní signálov. Korelácia vyjadruje vzájomný vzťah, súvislosť alebo inými slovami povedaní väzbu medzi signálmi. Korelácia medzi dvomi signálmi vyjadruje mieru podobnosti týchto dvoch signálov.

## 3.5 Časovo-frekvenčná analýza

Časovo-frekvenčná analýza sa používa na zachytenie vývoja krátkodobého spektra v čase, ktorý vymedzíme pomocou okna so zvolenou dĺžkou. Pri analýze reči sa používa okno s veľkousťou približne 25ms. Toto okienko môžme posúvať rôznymi spôsobmi, napríklad o celú dĺžku okna, o polovicu dĺžky okna alebo o tretinu dĺžky okna.

# Kapitola 4

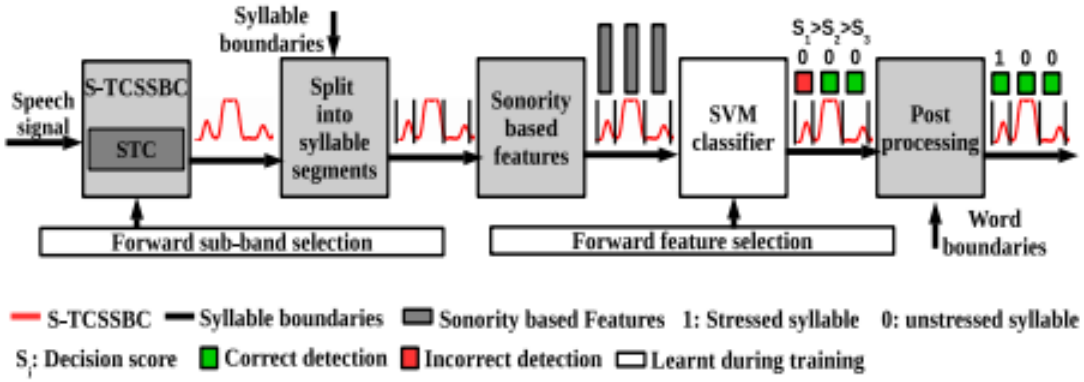
# Predchádzajúce riešenia

Automatické určenie prízvuku v slovách je užitočné pri určení kvality výslovnosti v cudzom jazyku, ktorý sa učíme pomocou automatu.

## 4.1 Automatic Detection of syllable stress using sonority based prominence features for pronunciation evaluation [YDG17]

Hlavným cieľom tohto článku je predstaviť riešenie, ktoré na rozdiel od predchádzajúcich riešení zapracováva znamenia motivované zvučnosťou v slabikách. Autori predstavili nový obrys funkcie zvučnosti pomocou spektro-časovej korelácie (STM) kontúr krátkodobej energie vo vybraných čiastkových pásmach. Navrhované riešenie má 5 krokov. V prvom kroku sa vypočíta S-TCSSBC pomocou STC na podmnožine čiastkových pásiem, ktoré boli nauťené počas tréningovej fázy. V druhom kroku, S-TCSSBC na úrovni viet je rozdelené na N segmentov slabík. N reprezentuhje počet slabík vo vete. V treťom kroku sa vypočíta množina znakov založených na zvučnosti pre

každú slabiku. Vo štvrtom kroku, sa každá slabika zaradí ako s prízvukom alebo bez prízvuku pomocou SVM klasifikátora používajúceho podmnožinu znakov vybraných prístupom forward feature selection. V poslednom kroku sa spracujú predpokladané označenia prízvuku, aby sa zabezpečilo, že každé viacslabičné slovo má označenú len jednu slabiku s prízvukom.




Obr. 4.1: päť krokov riešenia [**?** ]

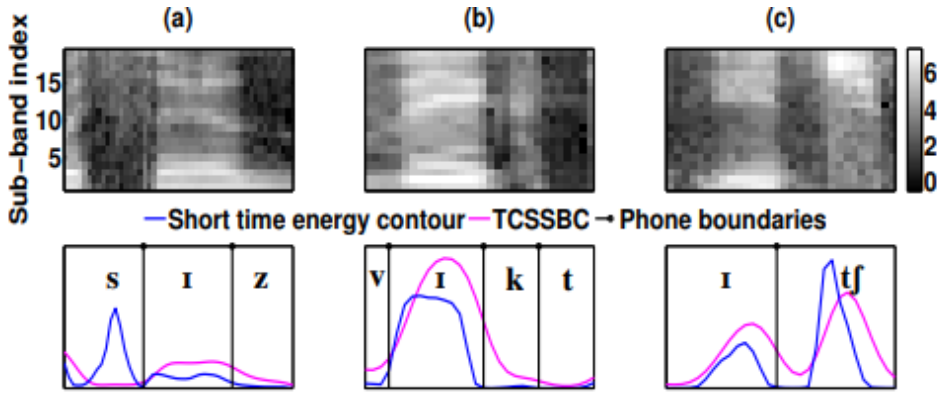
### 4.1.1 Databáza

Autori používajú ISLE corpus vo všetkých experimentoch vo svojej práci. Z databázy používajú všetkých 7834 nahrávok od štyridsiatichšiestich ľudí, ktorých natívny jazyk nie je angličtina. Každá nahrávka v databáze bola foneticky zaradená pomocou núteného zaraďovacieho procesu a následne ručne opravená tímom piatich jazykovedcov. V každom slove je označená len jedna slabika s prízvukom.

### 4.1.2 S-TCSSBC

TCSSBC je získané použitím STC na energie 19 čiastkových pásiem. Na obrázku môžme vidieť TCSSBC na troch príkladoch spolu s 19 čiastkovými

energiami.




Obr. 4.2: TCSSBC [**?** ]

Aj keď TCSSBC je robustnejšie ako krátkodobá energia, autori pozorovali, že ukazuje vrcholy keď sa v čiastkových pásmach nachádzajú konzistentné vzory. Toto sa môže stať iba pri regiónoch s vysokou zvučnosťou, ako jadrá slabík. Z obrázku môžme vidieť, že TCSSBC má vrchol vo foném "ts". je to spôsobené tým, že 17-19 čiastkových pásmach energie pre túto fonému sa nachádza silný pravidelný vzor. Práve toto naznačuje, že vrcholy ukazujúce sa v regiónoch s vysokou zvučnosťou môžu byť vylepšené odstránením irelevantných čiastkových pásiem. We refer to these as non-sonorous Autori ich identifikujú pomocou prístupu forward sub-band vybratia aby maximalizovali presnosť detekcie prízvuku. Na zvyšné čiastkové pásma sa aplikuje STC na vypočítanie S-TCSSBC kontúry $(X(m))$, kde $m$ predstavuje frame index.

### 4.1.3 Vypočet znakov na základe zvučnosti

Autori predstavujú 20 znakov používajúcich S-TCSSBC v dvoch množinách po 10 znakov.

- znaky na úrovni slabík
- znaky na úrovni jadier slabík

Všetky tieto znaky znaky sú rozdelené do troch kategórií.

- 10-dim znaky založené na sile
- 6-dim znaky založené na temporal variability
- 4-dim znaky založené na oblasti a trvaní

### Znaky založené na sile

Intenzita slabík s prízvukom je typicky vyššia ako slabík, na ktorých prízvuk nie je, keď sú zvyšné dva znaky rovnaké.

# Kapitola 5

# Návrh

V tejto kapitole popíšeme návrh nášho riešenia

# Kapitola 6

# Experiment

## 6.1 Spočítanie slabík na nahrávke

Nami navrhnutý algoritmus dostane na vstupe nahrávku typu .wav, ktorú si načítame a rozdelíme na:

- audio dáta uložené v matici
- vzorkovaciu rýchlosť vyjadrenú vo fs

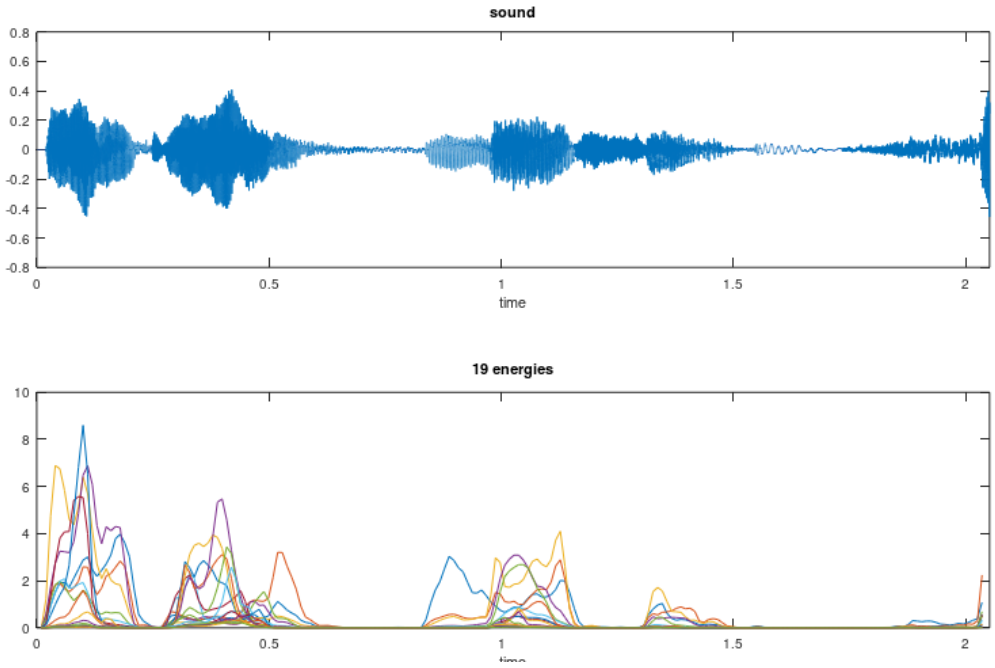
Nahrávku spracujeme pomocou troch metód na odhad počtu slabík:

- Časové váženie a korelácia
- Selekcia podpásiem a korelácia
- Hlasivkový tón

Kombináciou výsledkov z týchto metód určíme počet slabík v nahrávke. Nahrávka na ktorej si jednotlivé metódy priblížime obsahuje vetu "Long time, no see."nahovorenú ženským hlasom. Veta na tejto nahrávke má 4 slabiky.

### 6.1.1 19-kanálová sada filtrov

Po načítaní súboru typu .wav, ktorý obsahuje nahrávku reči, jeho audio dáta prejdú sadou 19-kanálových filtrov, pomocou ktorých dostaneme sériu vektorov energií. Táto sada filtrov používa butterworth filtre rozložené na pásma s centrami: 240, 360, 480, 600, 720, 840, 1000, 1150, 1300, 1450, 1600, 1800, 2000, 2200, 2400, 2700, 3000, 3300, 3750. Šírka jedného pásma siaha od jedného centra k tomu ďalšiemu. Pri okrajových centrách sme si zvolili šírku 60. Výslednú sériu vektorov uložíme do premennej $\mathbf{x}$, kde $\mathbf{x}(1)$ zodpovedá výslednému vektoru z prvého pásma, až $\mathbf{x}(19)$ zodpovedá výslednému vektoru z posledného pásma.




Obr. 6.1: 19 vektorov energií pre vetu "Long time, no see."

Na obrázku 6.1 vidíme zvukovú nahrávku potom ako prešla cez 19-kanálovú sadu filtrov, ktorá je tvorená butterworth filtrami, rozloženú na 19 vektorov

energií. Vektory predstavujú jednotlivé pásma, ktorých stredy sme si spomenuly vyššie. Každý vektor je označený inou farbou.

### 6.1.2 Časové váženie a korelácia

V nasledujúcom kroku týchto 19 vektorov energií rozdelíme na okná. Každé okno má veľkosť 20 milisekúnd. Na týchto vzorkách urobíme vzájomnú koreláciu podľa nasledovnej rovnice. [NW05]

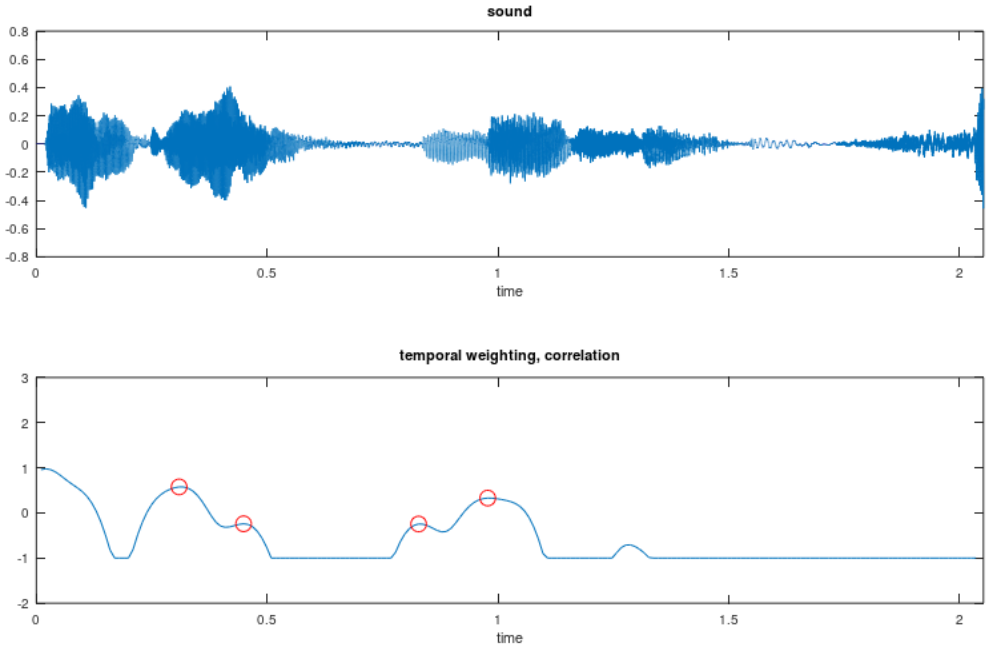
$$y_t = \frac{1}{K(K-1)} \sum_{j=0}^{K-2} \sum_{p=j+1}^{K-1} \mathbf{x}_{t+j} \cdot \mathbf{x}_{t+p}^T \tag{6.1}$$

Kde $K$ je konštanta, ktorú nastavíme počas testovania, $\mathbf{x}_t$, $\mathbf{x}_{t+1}$ až $\mathbf{x}_{t+K-1}$ predstavujú vekktory energií podpásiem vo vzostupnom poradí vzhľadom na čas.

V niektorých prípadoch, kde je rýchlosť nahovorenej reči na nahrávke vysoká, sa môže stať, že príde k rozmazaniu susednej energie. Stáva sa to pretože jedno okno v takomto prípade môže zachytývať viacero slabík. Toto nám môže stažiť rozoznávanie jednotlivých slabík. Preto pred použítím predchádzajúcej korelácie urobíme na vektore energie $x$ operáciu váženia. Na váženie používame Gaussove okno, ktoré je vycentrované na stred.[NW05]

$$\hat{\mathbf{x}}_{t+j} = w_j \mathbf{x}_{t+j} \tag{6.2}$$

Kde $w_1$ až $w_{K-1}$ sú koeficientami pre jednotlivé okná, ktoré volíme ako Gaussove okno vycentrované na stred každého okna.




Obr. 6.2: Veta "Long time, no see."po spracovaní časovým vážením a korelácií

Na obrázku 6.2 vidíme výsledok spracovania 19 vektorov energií, pomocou časového váženia a korelácie. 19-vektorov energrií sme získali spracovaním nahrávky pomocou 19-kanálovej sady filtrov. Následne sme nahrávku rozdelili na okná s 20 milisekundovou dĺžkou. Tieto okná sme pomocou Gaussoveho okna vycentrovaného na sted ováhovali. Medzi takto ováhovanými vektormi sme urobili vzájomnú koreláciu. Výsledná krivka je vyhladená a pomocou thresholdu na nej hľadáme vrcholy, ktoré predstavujú jednotlivé slabiky. Tieto miesta sú na krivke vyznačené červenou kružnicou.

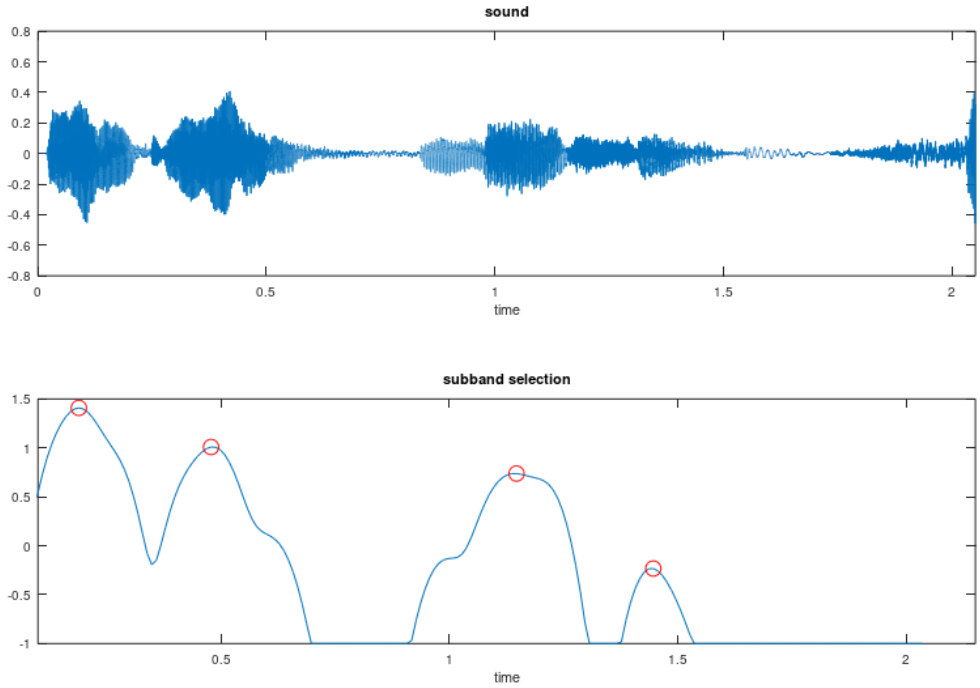
### 6.1.3 Selekcia podpásiem a korelácia

Pri vyslovení dlhšej slabiky sa môže hlas jemne zakolísať, čo na nahrávke môže spôsobiť jemné posunutie formantu. Tákéto zakolísanie môže spôsobiť to, že formant sa posunie do iného pásma, čo môže vo vyššie spomenutej me-

tóde viesť k tomu, že jednu slabiku zaregistruje ako dve. Preto zároveň s predchádzajúcou metódou používame aj metódu selekcie subpásiem a následnú koreláciu. V tejto metóde vypočítame podľa modulu podpásiem trajektóriu, ktorá je priemorom všetkých párov z kompresovaných vektorov energií podpásiem, podľa nasledovného vzorca. [NW05]

$$y_t = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathbf{x}_t(i)\mathbf{x}_t(j) \tag{6.3}$$

Kde $N$ je počet podpásiem a $M = N(N-1)/2$ vyjadruje číslo jedinečných párov, ktoré sa nám vytvoria pri použití tohto vzorca.



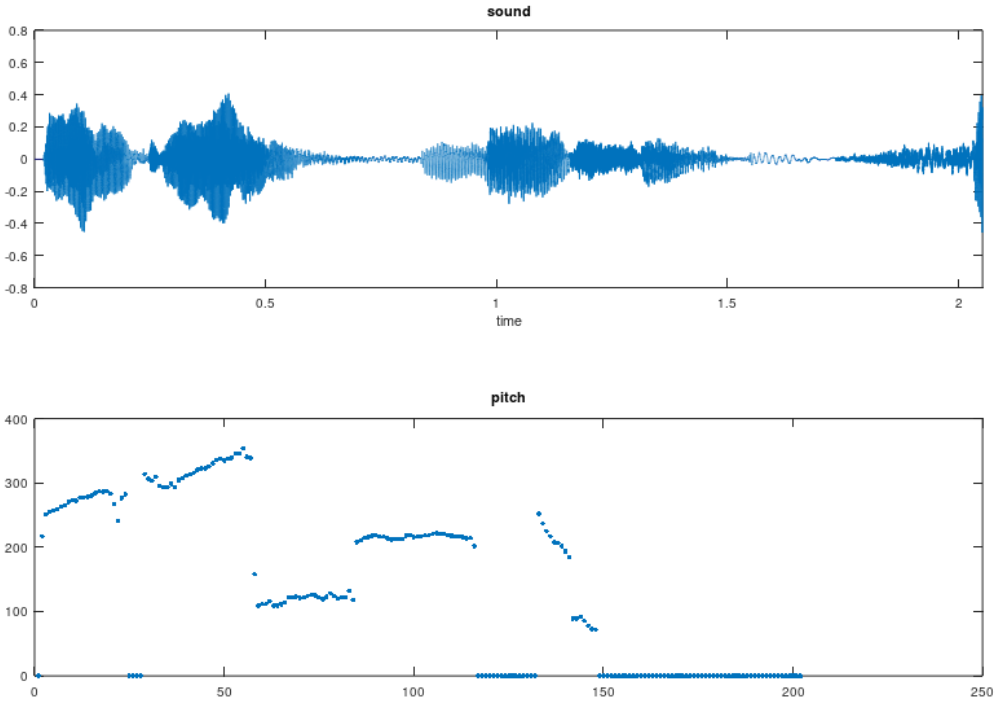

Obr. 6.3: Veta "Long time, no see."po selekcií subpásiem a korelácií

Na obrázku 6.3 vidíme krivku po selekcií subpásiem a korelácií. Pri tejto metóde si po vyššie spomenutách oknách vyberieme niekoľko najvýraznejších vektorov energií, ktoré sme získali z 19-kanálovej sady filtrov. Vytvoríme

jedinečné páry na ktorých urobíme koreláciu. Výsledná krivka je rovnako ako pri časovom vážení vyhľadená a pomocou thresholdu sú na nej nájdené vrcholy, ktoré predstavujú jednotlivé slabiky. Miesta, kde tento prístup na krivke našiel slabiku, sú vyznačené červenou kružnicou. Môžme si všimnúť, že v slove "time"(drhý vrchol zľava),ktoré je jednoslabičné, táto metóda na rozdiel od predchádzajúcej označila iba jednu slabiku.

### 6.1.4 Hlasivkový tón

Ako ďalšiu možnosť na detekovanie slabiky používame estimáciu hlasivkového tónu. Ako estimátor používame wavesurfer, ktorý nám z nahrávky poskytne odhad toho, kde sa nachádza hlasivkový tón. Estimácia hlasivkového tónu nám slúži ako pomocná metóda popri vyššie spomenutých metódach časového váženia a korelácie a selekcií podpásiem a a korelácie. Pomocou tejto metódy sa uistíme, či sa v mieste, kde obe vyššie spomenuté metódy určili slabiku, nachádza hlasivkový tón. Ak je hlasivkový tón prítomný, slabiku započítame. Ak prítomný nie je, vieme že ak by na tomto mieste program zaznamenal slabiku, nemáme ju do výsledku započítať.



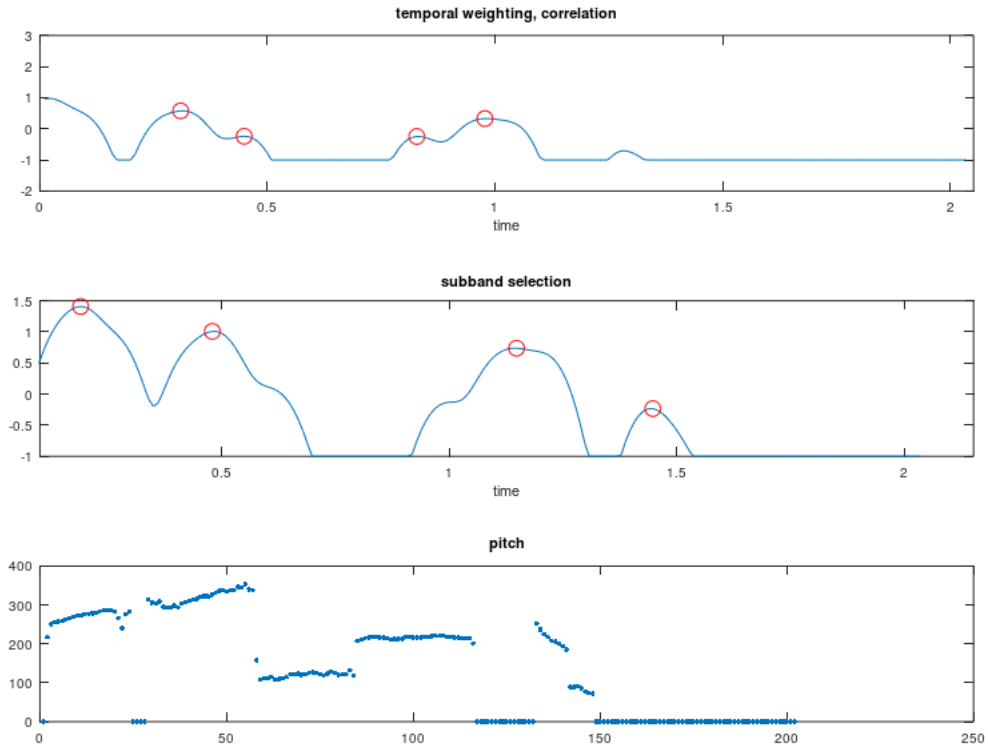

Obr. 6.4: odhad hlasivkového tónu pre vetu "Long time, no see."

Na obrázku 6.4 vidíme odhad hlasivkového tónu vygenerovaný pomocou wavesurfer. Ak je hodnota hlasivkového tónu 0, znamená to, že ho na nahrávke nedetekujeme. Ostatné hodnoty sa môžu pohybovať v rozmedzí od 60Hz do 600Hz, táto hodnota závisí od veku a pohlavia rečníka.

### 6.1.5 Spojenie výsledkov

Spojenie týchto troch metód robíme postupne. Najprv prejdeme výsledky časového váženia a korelácie, skontrolujeme, či na miestach, kde táto metóda detekuje slabiku, je prítomný hlasivkový tón. Ak prítomný je, slabiku započítame, ak nie je, slabiku ignorujeme. Tento istý postup zopakujeme aj pre selekciu podpásiem a koreláciu. Potom ako dostaneme konečný počet slabík z oboch metód spojených s metódou prítomnosti hlasivkového tónu, tieto dva

počty spočítame a spriemerujeme.




Obr. 6.5: ukážka metód pre vetu "Long time, no see."

Na obrázku 6.5 vidíme porovnanie výsledkov všetkých vyššie spomenutých metód. Červenými kružnicami je zvýraznený vrchol, ktorý daná metóda detekuje ako slabiku. Dolu je zobrazený hlasivkový tón, podľa ktorého rozhodujeme, či danú samohlásky započítame.

Náš algoritmus je citlivý na parametre a preto je ich nastavenie dôležité. Niektoré parametre od ktorých táto metóda závisí sú veľkosti okna, hodnota thresholdu a počet subpásiem na ktorých robíme v poslednom kroku koreláciu.

Jedným z parametrov od ktorého je náš algoritmus závislý je počet subpásiem na ktorých robíme koreláciou pri selekcií subpásiem. Ako hodnotu tohto

parametra sme zvolili 3, pretože tento parameter by mal mať hodnotu blízku počtu formantov.

Ďalším dôležitým parametrom je výber thresholdu, pomocou ktorého na krivkách hľadáme vrcholy, ktoré nám reprezentujú slabiky. Preto aby sme našli správnu hodnotu thresholdu, sme sa rozhodli vyskúšať nastaviť mu hodnotu v rozmedzí od -4,5 po -3,6 s krokom 0,1. Ako konečnú hodnotou pre threshold sme vybrali tú, s ktorou bola úspešnosť na testovacích dátach najvyššia.

| **hodnota thresholdu** | **úspešnosť** |
|---|---|
| -4,5 | 41,212 |
| -4,4 | 41,510 |
| -4,3 | 41,375 |
| -4,2 | 41,792 |
| -4,1 | 41,896 |
| -4,0 | 42,073 |
| -3,9 | 42,292 |
| -3,8 | 42,500 |
| -3,7 | 42,750 |
| -3,6 | 42,760 |

Tabuľka 6.1: úspešnosť pri rôznych hodnotách tresholdu

# Kapitola 7

# Implementácia

V tejto kapitole popíšeme implementáciu nášho riešenia.

# Kapitola 8

# Výsledky

V tejto kapitole sa pozrieme na výsledky, ktoré naše riešenie dosiahlo.

# Kapitola 9

# Záver

Záver diplomovej práce.

# Literatúra

[MFL98] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP '98 (Cat. No.98CH36181)*, volume 2, pages 729–732 vol.2, 1998.

[NW05] S. Narayanan and Dagen Wang. Speech rate estimation s temporal correlation and selected sub-band correlation. In *Proceedings. (ICASSP '05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, volume 1, pages I/413–I/416 Vol. 1, 2005.

[YDG17] Chiranjeevi Yarra, Om Deshmukh, and Prasanta Ghosh. Automatic detection of syllable stress using sonority based prominence features for pronunciation evaluation. pages 5845–5849, 03 2017.

# Zoznam obrázkov