Západočeská univerzita v Plzni

Fakulta aplikovaných věd

Katedra kybernetiky

BAKALÁŘSKÁ PRÁCE

PLZEŇ, 2012 **MILAN JAROLÍN**

P R O H L Á Š E N Í

Předkládám tímto k posouzení a obhajobě bakalářskou práci zpracovanou na závěr studia na Fakultě aplikovaných věd Západočeské univerzity v Plzni.

Prohlašuji, že jsem bakalářskou práci vypracoval samostatně a výhradně s použitím odborné literatury a pramenů, jejichž úplný seznam je její součástí.

v Plzni dne 20. srpna 2012 .....

vlastnoruční podpis

P O D Ě K O V Á N Í

Tímto bych chtěl poděkovat vedoucímu bakalářské práce Ing. Mgr. Josefu Psutkovi, Ph.D. za velmi užitečnou pomoc a cenné rady při zpracování práce.

Anotace

V této bakalářské práci se zabýváme problematikou parametrizace řečového signálu pomocí různých způsobů modifikace metody Melovských kepstrálních koeficientů(MFCC) v procesu rozpoznávání řeči s ohledem na množiny řečníků.

Naším cílem je ověření optimálního rozmístění a tvaru filtrů na frekvenční ose u metody MFCC. První modifikace metody MFCC spočívá v použití alternativních tvarů filtrů (obdélníkový, cosinusový, sinusový, lichoběžníkový). Druhá modifikace se zabývá odlišným rozmístěním filtrů v melovské bance oproti původnímu, které se snaží o kompenzaci nelinearity citlivosti sluchového ústrojí vůči frekvenci.

Ve výsledku zkoumáme vliv těchto modifikací parametrizační metody MFCC na úspěšnost rozpoznávání promluv ze zvolené množiny testovaných dat, pro jednotlivé množiny trénovacích řečníků. Pro realizaci procesů trénování a rozpoznávání používáme modul HTK. Získané poznatky poté konfrontujeme s výsledky dosaženými původním nastavením metody MFCC.

Klíčová slova

parametrizace řeči, pásmový filtr, melovské filtry, MFCC

Anotation

In this bachelor thesis we deal with parametrization of the speech signal problem using different ways of modification method Mel cepstral coefficients (MFCC) in the process of speech recognition with respect to a set of speakers.

Our aim is to verify the optimal location and the shape of the filters on the frequency axis at the method MFCC. The first modification of the method MFCC consists in using alternative shapes of filters(rectangle, cosine, sine, trapezoid). The second modification deals with a different placement of filters in the mel bank compared to the original, which tries to compensate nonlinear sensitivity of the auditory system to frequency.

As a result, we examine the effects of these modifications parameterization method MFCC for recognition success utterances from a selected set of test data for each set of training speakers. For the execution of the processes of training and recognition using HTK module. Gained knowledge is then confronted with the results achieved by setting the original method MFCC.

Key words

speech parameterization, bandpass filter, mel filter, MFCC

Obsah

Zdroje

[**1**] Psutka, J., Matoušek, J., Muller, L., Radová, V.: Mluvíme s počítačem česky, Nakladatelství Academia, 2006

[**2**] Chalupníček, K.: Diplomová práce na téma Rozpoznávání diktované řeči pro medicínské aplikace, Vysoké učení technické v Brně, 2004

[**3**] Tychtl, Z.: Skripta pro předmět Zpracování signálu pro klasifikaci, Západočeská univerzita v Plzni

[**4**] http://matematika.cuni.cz/dl/analyza/37-fou/lekce37-fou-pmin.pdf

[**5**] Young, S. et al.: The HTK Book, User's manual, 1999

1 Úvod

Lidská řeč je velmi komplexní, souvisle časově proměnný proces s výrazně nelineárními charakteristikami, proto jsme jen velmi omezeně schopni vytvořit univerzální matematický model, který by tento proces odpovídajícím způsobem popisoval.

Uvážíme-li však omezenou rychlost řečových orgánů v lidském těle, vycházející z jejich nenulové hmotnosti, pak můžeme parametry matematického modelu (modelu produkce řeči) považovat na určitém krátkém časovém intervalu (v řádech milisekund) za konstantní. Tohoto předpokladu využívají metody krátkodobé analýzy, které tak mohou řečový signál po těchto intervalech rozdělit na tzv. mikrosegmenty a zpracovávat je nezávisle na ostatních.

Jednou z možností dalšího zpracování řečového signálu po těchto mikrosegmentech je parametrizace námi zkoumanou metodou MFCC. V dnešní době je zejména v úlohách rozpoznávání řeči velmi často používaná. Smysl takové parametrizace obecně tkví ve snížení objemu přenášených a zpracovávaných dat tak, aby se při tom zachovala maximální informace. Pokud uvážíme, že máme jisté základní znalosti o procesu produkce řeči. Můžeme v signálu hledat takové příznaky, které jednoznačně reprezentují určitou polohu hlasového ústrojí a tím i vydávaný tón. V úlohách nezávislých na řečníkovi se pak snažíme použít takovou reprezentaci řeči, která se snaží o abstrahování od charakteristických rysů jednotlivých hlasů a vybírá takové příznaky, jež jistým způsobem představují stejné části různých promluv.

Záměrem této práce je ověření optimálního nastavení parametrů metody MFCC (tvar a rozmístění pásmových filtrů), vzhledem k úspěšnosti rozpoznávání. A to především s ohledem na odlišné trénovací a testovací množiny řečníků (muži, ženy). První experiment tedy spočívá v zaměňování původního (trojúhelníkového) tvaru frekvenční odezvy filtrů v melovské bance za alternativní tvary (obdélník, cosinus, sinus, lichoběžník). A druhý experiment pak zkoumá alternativní rozmístění pásmových filtrů v melovské bance, které v původní podobě, kompenzuje nelinearitu při vnímání změn frekvence u lidského ucha.

2 Teorie

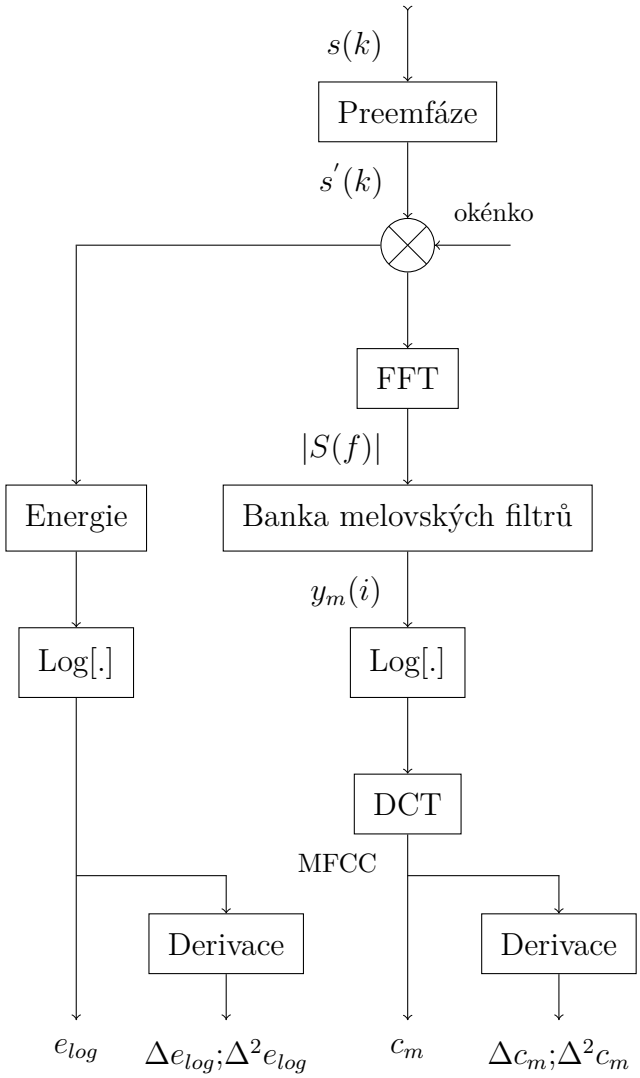
2.1 Melovské frekvenční kepstrální koeficienty MFCC

Tato metoda patří do oblasti Homomorfní analýzy, která je ze skupiny postupů nelineárního zpracování signálů využívajících principu superpozice. Tyto postupy se zaměřují na analýzu signálů, vznikajících konvolucí dvou nebo více složek. Což odpovídá řečovému signálu, jelikož jak je ukázáno například v [1], jeho vytváření se dá vnímat i jako konvoluce budící funkce a impulsní odezvy hlasového ústrojí. Kde budící funkce má podobu periodického sledu pulsů u znělých hlásek a náhodného šumu u neznělých.

Metoda MFCC využívá zpracování řečového signálu jak v časové tak i v frekvenční oblasti. Popisuje při tom spektrální vlastnosti zmíněného signálu, konkrétně krátkodobé komplexní kepstrum. Snaží se při tom o respektování poznatků o citlivosti lidského ucha při vnímání zvukového vlnění na různých frekvencích.

Rozšířením této metody o kalkulaci delta a akceleračních koeficientů, získáme aditivní informaci o dynamickém průběhu řeči. Cílem metody Melovských frekvenčních kepstrálních koeficientů je určit parametry řečového systému pomocí homomorfní filtrace. Výsledkem analýzy pro každý mikrosegment je pak vektor čísel, které popisují danou část řečového signálu.

Následující obrázek představuje schéma algoritmu výpočtu MFCC s delta a akceleračními koeficienty, jehož částem se budeme podrobněji věnovat v následujících podkapitolách.



Obrázek 1: Schéma algoritmu MFCC, [1]

2.1.1 Preemfáze

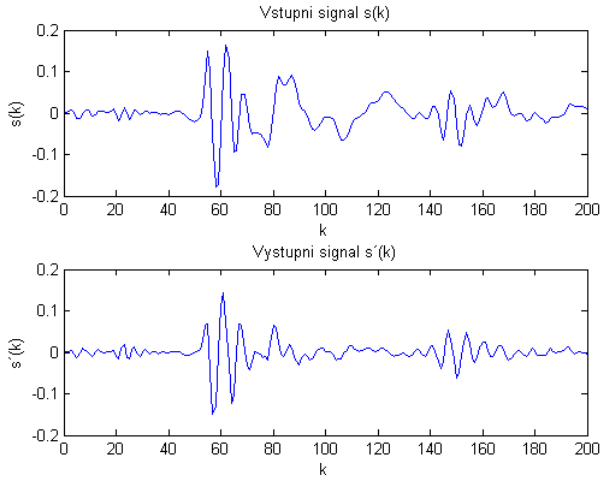
Tento proces se často používá před vlastním zpracováním řečového signálu. Jeho význam je v zesílení amplitud spektrálních složek s jejich vzrůstající frekvencí ve frekvenčním spektru signálu, ve kterém jsou právě tyto složky potlačovány přirozeným chováním řečového ústrojí. Tím se oba tyto procesy do jisté míry vzájemně kompenzují a vyrovnává se tak energetické spektrum celého přenášeného pásma. To znamená, že amplitudy frekvenčního spektra se budou pohybovat řádově na podobných úrovních, což může mít pozitivní vliv na další zpracování takto upraveného signálu. Realizaci preemfáze pro číslicové zpracování můžeme provést dvěma způsoby:

a) jako analogový filtr, použit na analogový signál ještě před jeho diskretizací vzorkováním a kvantováním, jehož frekvenční charakteristika má strmost $+20dB/dek$ od frekvence přibližně $100Hz$, od které řečové ústrojí začíná potlačovat složky frekvenčního spektra

b) jako číslicový filtr, použitý až za vzorkovačem a kvantizérem, provádějící předzpracování signálu podle vztahu

$$s^{'}(k) = s(k) - as(k-1), \tag{1}$$

kde $s(k)$ je vstupní a $s^{'}(k)$ je výstupní vzorek filtru v čase k, za parametr a volíme hodnoty z rozmezí 0.9 až 1.



Obrázek 2: Ukázka vlivu preemfáze na průběh signálu

2.1.2 Aplikace okénka

Další částí předzpracování signálu v časové oblasti je aplikace tzv. okénka. Na aplikaci se dá nahlížet ze dvou pohledů. Jako na diskrétní konvoluci dvou signálů(první představující vstupní řečový signál a druhý v podobě okénka). Tak i na průchod vstupního řečového signálu filtrem ztvárňující zvolené okénko. Z konvolučnímu teorému vysvětlenému například v [3] vyplývá, že jsou oba tyto přístupy ekvivalentní ale můžou nám pomoci dojít k novým důsledkům. Smyslem použití okénka je výběr zvoleného rozsahu vzorků a přiřazení každému vzorku určitou váhu.

$$Q_n = \sum_{k=-\infty}^{\infty} s^{'}(k) \cdot w(n-k), \tag{2}$$

kde Q_n je výsledný převážený signál, $s^{'}(k)$ značí vzorek akustického signálu v čase k a $w(n)$ jsou vzorky okénka, kterým se váží vzorky $s^{'}(k)$.

Rozsah a podoba okénka se mění podle zaměření. V úlohách zpracování signálu pro rozpoznávání řeči se nejčastěji používají dva druhy okének:

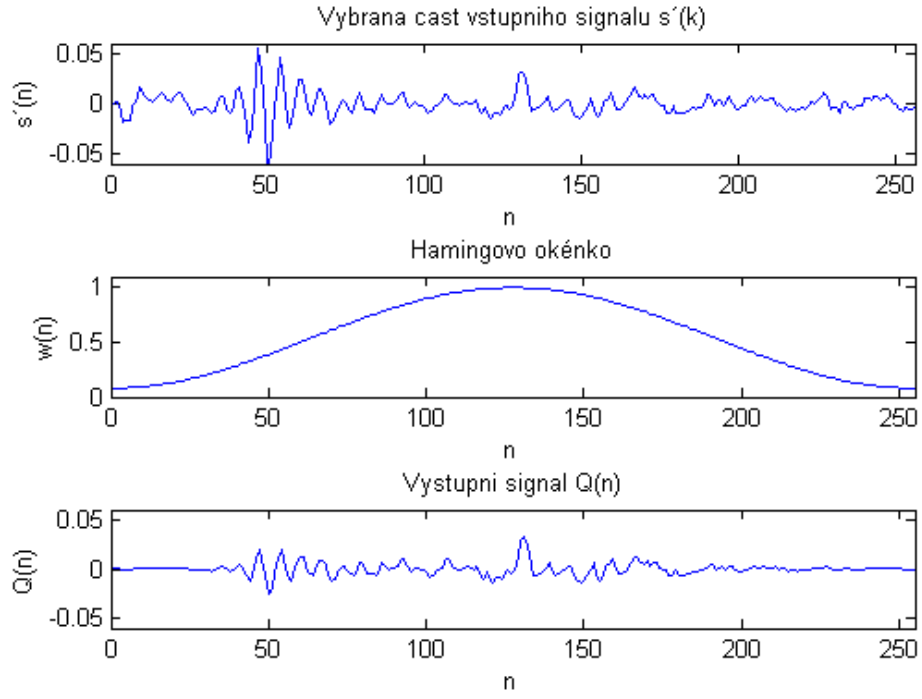
a) Pravoúhlé okénko
Přiděluje všem vzorkům vybraných okénkem stejnou váhu.

$$w(n) = \begin{cases} 1 & \text{pro } 0 \leq n \leq L-1 \\ 0 & \text{pro ostatní n} \end{cases} \tag{3}$$

b) Hammingovo okénko
Pro úlohy kde je snahou potlačit při zpracování vzorky na krajích okénka.

$$w(n) = \begin{cases} 0.54 - 0.46cos(2\pi n/L-1) & \text{pro } 0 \leq n \leq L-1 \\ 0 & \text{pro ostatní n} \end{cases} \tag{4}$$

Pro vysvětlení volby délky a posunu okénka si zopakujeme pojem mikrosegment, který jsme si definovali jako krátký časový interval o délce přibližně $10ms$ obsahující vzorky řečového signálu. V podstatě celý signál je rozdělený na posloupnost těchto mikrosegmentů, kde pro každý mikrosegment se budou jednotlivé kroky algoritmu provádět odděleně.



Obrázek 3: Konvoluce Hammingova okénka s vybranou oblastí řečového signálu

Posun okénka tedy bude odpovídat délce jednoho mikrosegmentu, tedy počtu vzorků v něm obsažených. Jestliže $T[ms]$ je vzorkovací perioda a je-li délka mikrosegmentu již zmíněných $10ms$, pak pro N platí $N = 10/T$. Například pro frekvenci vzorkování v telefonním pásmu $F_v = 8kHz$ je $T = 0.125ms$ a počet vzorků v mikrosegmentu je tudíž $N = 80$. Samotný index n, uvedený ve vztahu (2) a představující počáteční pozici oblasti na kterou se bude okénko aplikovat v diskrétním vstupním signálu, se volí ve tvaru $n = Ni - 1$, kde i můžeme vnímat jako pořadové číslo mikrosegmentu.

Délka okénka L by logicky, měla být také rovná délce mikrosegmentu. Obecně je ale výhodnější odvodit délku okénka od navazujících kroků algoritmu pro výpočet MFCC. Například pro krátkodobou Furierovu transformaci je potřeba délku okénka stanovit přibližně na $L \approx 3N$, abychom byli schopni zachytit periodické vlastnosti znělých úseků řeči. To znamená, že okénko bude převažovat vzorky přibližně o délce $30ms$. Pokud pro výpočet frekvenčního spektra navíc používáme algoritmus FFT(Fast Furier Transformation), je vhodné, aby délka okénka byla rovna nejbližší mocnině dvou. Hammingovu okénku je v tomto případě dána přednost před pravoúhlým, pro-

tože z hlediska frekvenčních vlastností popsaných v [1], které jsou podstatné při následující Fourierově analýze, poskytuje Hammingovo okénko mnohem větší útlum pro vyšší frekvence.

Shrnutím těchto faktů a uvážením důsledků působení okénka na signál, tj. že všechny vzorky vně okénka jsou váženy nulou, takže je při vyčíslování vztahu (2) nemusíme uvažovat, můžeme uvedený vztah pro výpočet izolovaného mikrosegmentu s pořadovým číslem například $i = 1$, tj. $n = N - 1$, přepsat do tvaru

$$Q_{N-1} = \sum_{k=0}^{N-1} s'(k) \cdot w(N - 1 - k) \tag{5}$$

2.1.3 Výpočet energie

Funkci krátkodobé energie signálu lze definovat vztahem

$$E_n = \sum_{k=-\infty}^{\infty} [s'(k) \cdot w(n - k)]^2, \tag{6}$$

kde $s'(k)$ je vzorek signálu v čase k a $w(n)$ reprezentuje příslušný typ okénka z předešlého kroku. Hodnoty funkce krátkodobé energie poskytují pro každý mikrosegment informaci o průměrné hodnotě energie v mirkosegmentu. Jedním z nedostatků této charakteristiky je její značná citlivost na velké změny úrovně signálu. Již tak vysoká dynamika řečového signálu je vlivem kvadrátu v rovnici (6) ještě zvýšena. Hodnoty krátkodobé energie mohou být využity například při automatickém oddělování segmentů ticha od segmentů řeči, lze jich též využít při oddělování znělých a neznělých částí promluvy. Hodnoty funkce krátkodobé energie se kdysi využívali též jako příznaky v jednoduchých klasifikátorech slov. Dochází zde ale ke znehodnocení původního informačního obsahu, takže z těchto krátkodobých charakteristik nelze rekonstruovat původní signál.

2.1.4 Diskrétní Fourierova transformace DFT

Doposud jsme se pohybovali v časové oblasti zpracování signálu a nyní se můžeme přesunout do oblasti frekvenční. Za předpokladu stacionarity signálu(podobně jako v časové oblasti) se dostáváme ke krátkodobé spektrální analýze. Pro tento účel se všeobecně jako jeden z kroků využívá právě DFT. Obecně je Fourierova transformace odvozena z Fourierových řad, které jsou založeny na myšlence, že jakýkoliv průběh signálu se dá rozložit na nekonečnou řadu harmonických (sinusových) průběhů o různé frekvenci, fázi a amplitudě. Vyjádření tohoto předpokladu vede k rovnici

$$\begin{aligned} x(t) &= \frac{a_0}{2} + \sum_{k=1}^{\infty}[a_k \cos(k\omega t) + b_k \sin(k\omega t)], \qquad (7) \\ a_0 &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)dx, \\ a_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(kx)dx, \\ b_k &= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(kx)dx. \end{aligned}$$

Z toho vyplývá, že DFT vlastně vypočítává jistou váhovou funkci, která nám poskytuje informaci o tom jaké zastoupení frekvencí a v jakém poměru jednotlivé mikrosegmenty obsahují, tedy jejich frekvenční spektra. Aplikací tohoto postupu na naši úlohu a dosazením $e^{j\omega k}$ za $\cos(k\omega t) + \sin(k\omega t)$ se dostaneme k předpisu

$$S(\omega, n) = \sum_{k=0}^{N} s^{'}(k)w(n-k)e^{-j2\pi nk/N}, \qquad (8)$$

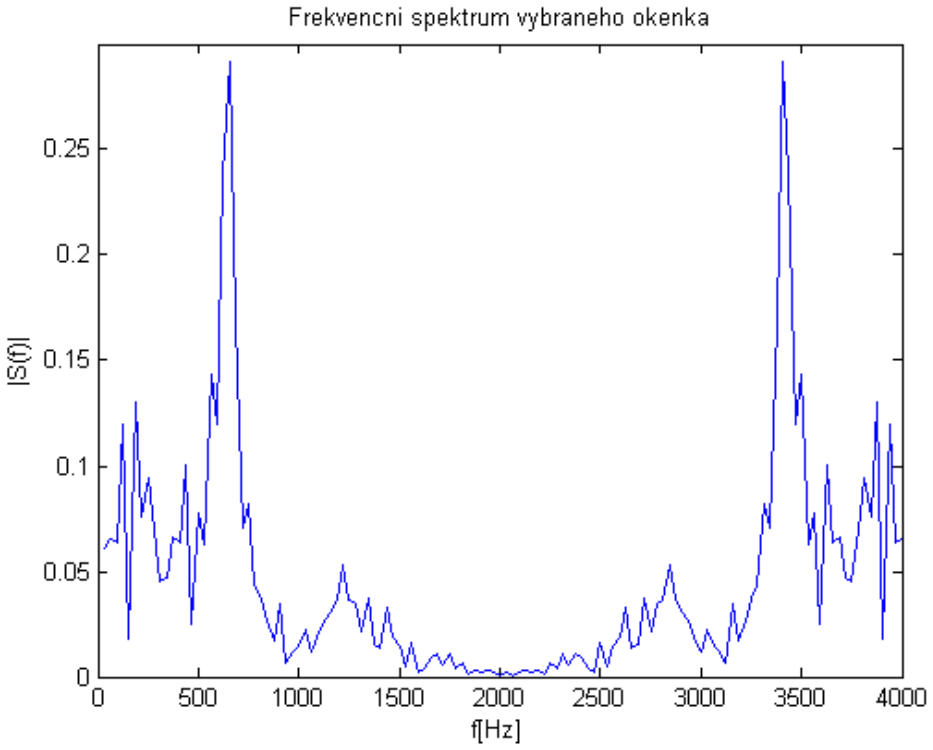
kde $s^{'}(k)$ jsou vzorky předzpracovaného řečového signálu, $w(n)$ je funkce Hammingova okénka a $S(\omega, n)$ je výsledná obecně komplexní váhová funkce.

DFT je v současné době široce využívaná, pro svoji univerzálnost a praktické vlastnosti. Velkou výhodou DFT je úplná čili bezeztrátová reprezentace vzorkovaného signálu, tzn. že z ní můžeme signál opět zpětně rekonstruovat, k tomu účelu slouží IDFT(Inverse Discrete Fourier Transformation). Pro naše účely potřebujeme pouze reálnou část z komplexního prostoru. Za tím důvodem si z výsledného komplexního spektra vypočteme spektrum amplitudové

$$|S_{(\omega,n)}| = \sqrt{Re\{S_{(\omega,n)}\}^2 + Im\{S_{(\omega,n)}\}^2} \qquad (9)$$

kde $Re\{S_{(\omega,n)}\}$ logicky označuje reálnou složku a $Im\{S_{(\omega,n)}\}$ imaginární složku komplexního spektra. V důsledku tohoto přepočtu se nám počet vzorků snížil na polovinu avšak bez ztráty užitečné informace. Protože námi zkoumaný

signál je diskrétní a periodický s periodou odpovídající $T_v = 1/f_v$ o délce N vzorků odpovídajícím velikosti okénka L, výsledná váhová funkce $|S(\omega, n)|$ je také diskrétní a periodický signál o délce $N/2$ vzorků odpovídajícím velikosti zkoumaného frekvenčního rozsahu $f_v/2$ s intervalem rovnajícím se f_v/N. Navíc bude i symetrická, tzn. $S(N-k) = S(k)$ což ale vyplývá z vlastností samotné DFT blíže vysvětlení ve [3].



Obrázek 4: Ukázka frekvenčního spektra vypočteného metodou krátkodobé DFT

Jelikož klasická DFT je velice výpočetně náročná, používá se, v situacích které to dovolují, algoritmus FFT pro značné zjednodušení výpočetního algoritmu. Pro představu výpočet N bodové DFT vyžaduje N^2 operací, čili její složitost kvadraticky stoupá s počtem zpracovávaných bodů. Pokud bychom zpracovávaný interval bodů rozdělili na sudé $y(n)$ a liché $z(n)$ body, můžeme pro obě množiny o velikosti $\frac{N}{2}$ provést výpočet podle upraveného předpisu pro DFT

$$x(k) = \sum_{n=0}^{\frac{N}{2}-1} [y(n)e^{-j2\pi 2nk/N} + z(n)e^{-j2\pi(2n+1)k/N}]. \tag{10}$$

Výsledná výpočetní náročnost celého procesu tedy je

$$2\left[\frac{N}{2}\right]^2 = \frac{N^2}{2} \quad (< N^2)$$

Tohoto postupu právě využívá rekurzivní algoritmus FFT, který postupně rozděluje interval pro výpočet DFT až na dvoubodové části. Za omezující podmínky, že původní N musí být mocninou dvou, se dostáváme k výpočetní náročnosti odpovídající

$$\frac{N}{2}\log_2 N.$$

2.1.5 Melovská banka filtrů

Klíčová část procesu kalkulace MFCC, je pásmová filtrace frekvenčního spektra signálu vypočteného v minulém kroku. Tato filtrace je reprezentovaná bankou filtrů, nejčastěji trojúhelníkového tvaru, lineárně rozmístěných na tzv. melovské frekvenční škále. Použití této škály vychází z experimentů s vnímáním jednotlivých frekvencí při níž bylo zjištěno, že citlivost vnímání lidského ucha na změnu frekvence není lineární a s vzrůstající frekvencí klesá.Vztah aproximující citlivostní křivky ucha vůči příjmané frekvenci vlnění, byl ustanoven jako

$$f_m = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{11}$$

kde $f_m[mel]$ je frekvence v nelineární melovské škále a $f[Hz]$ je frekvence v původní lineární škále.

Počet těchto filtrů M^* závisí na počtu kritických pásem obsažených ve zkoumaném rozsahu. Kritické pásmo lze chápat jako frekvenční pásmo, kterém dochází k výrazným změnám při subjektivním vnímání zvuku. Bližší vysvětlení tohoto pojmu a způsobu určování jeho hodnoty lze najít v [1].Nejčastěji volené hodnoty počtu filtrů v závislosti na vzorkovací frekvenci f_v či zkoumaném pásmu B_w jsou uvedeny v tabulce (1).

Frekvence vzorkování F_v $[Hz]$	8000	11000	16000	22000	44000
Přenášené pásmo $(0; B_w)$ $[Hz]$	(0;4000)	(0;5500)	(0;8000)	(0;11000)	(0;22000)
Přenášené pásmo $(0; B_{mw})$ $[mel]$	(0;2146)	(0;2458)	(0;2840)	(0;3174)	(0;3921)
Počet pásem $M*$	15	17	20	22	27

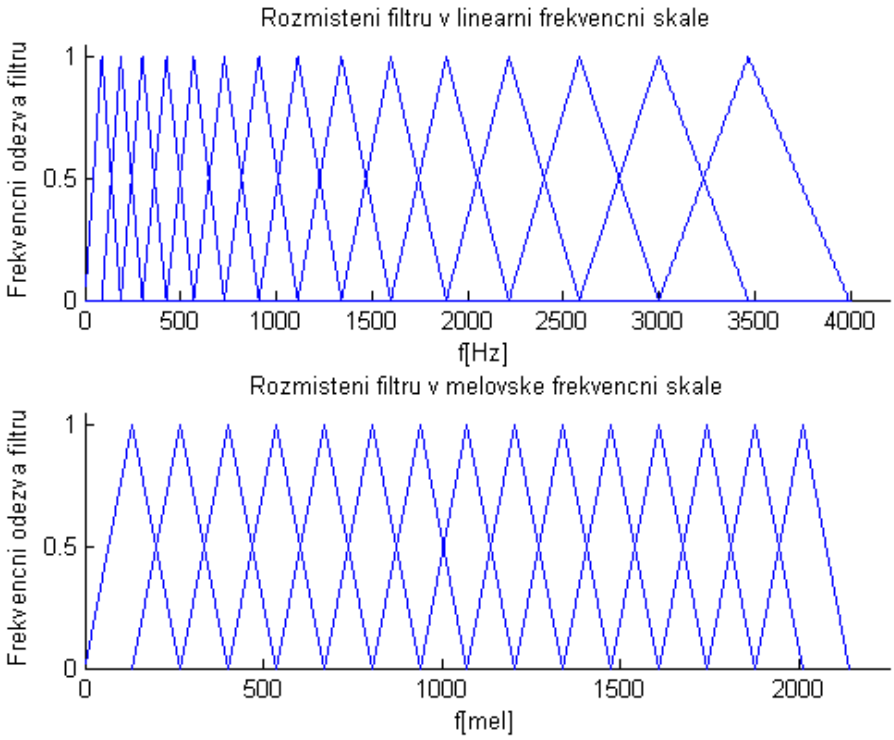
Tabulka 1: Typické hodnoty počtu filtrů $M*$ pro dané přenášené pásmo$(0; B_w)$,[1]

Jak je patrné z obrázku(5), jednotlivé filtry můžeme definovat pomocí jejich střední frekvencí $b_{m,i}$, které se posouvají o b_Δ pro jednotlivé filtry, a jejich šířkou. Předpis pro výpočet středních frekvencí lze zapsat jako

$$b_{m,i} = b_{m,i-1} + b_\Delta, \tag{12}$$

kde i označuje index filtru a pohybuje se v rozmezí $i = 1, 2, \cdot, M^*$. Hodnota intervalu mezi filtry se pohybuje přibližně okolo $b_\Delta = 140mel$ a šířka jednoho filtru je definována jako její dvojnásobek.

V reálné implementaci banky melovských filtrů existují dvě možnosti jak naložit s daty. První možností je využití lineárního rozmístění filtrů v melovské škále a přepočíst si celé frekvenční spektrum pro každé okénko taktéž do této škály. Druhá možnost tkví naopak v ponechání všech dat získaných DFT v původní podobě a přepočíst lineární rozmístění filtrů v melovské frekvenční škále, pomocí inverzní funkce k rovnici (11) na klasickou lineární škálu udávanou v hertzích. Ve většině případů se využívá druhé možnosti, pro menší výpočetní náročnost Za drobnou nevýhodu této možnosti lze považovat aproximaci při přepočtu, ve kterém dochází k deformaci stran jednotlivých trojúhelníků čímž by se z přímek měli stávat lehce nelineární křivky. Toto zjednodušení má za následek rozdíly mezi oběma možnostmi.



Obrázek 5: Ukázka rozmístění filtrů v obou škálách pro $M^* = 15$ a vzorkovací frekvenci $f_v = 8000Hz$

Nehledě na předcházející volbu se průchod signálu filtry dá zapsat následující rovnicí

$$y_m(i) = \sum_{f=b_{m,i-1}}^{b_{m,i+1}} |S(f)|u(f,i), \tag{13}$$

kde $y_m(i)$ je odezva i-tého filtru, $|S(f)|$ amplitudové spektrum vstupního signálu a $u(f,i)$ znázorňuje trojúhelníkovou funkci filtru, která je zapsána jako

$$u(f,i) = \begin{cases} \frac{f-b_{i-1}}{b_i-b_{i-1}} & \text{pro } b_{i-1} \leq f < b_i \\ \frac{f-b_{i+1}}{b_i-b_{i+1}} & \text{pro } b_i \leq f < b_i + 1 \\ 0 & \text{pro ostatní.} \end{cases} \tag{14}$$

Samotná pásmová filtrace má v závěru podobu násobení každého bodu FFT s odpovídajícím ziskem filtru na určité frekvenci a ve výsledku jsou tyto hodnoty pro každý filtr se sčítávány.

2.1.6 Melovské kepstrální koeficienty

Následujícím krokem je aplikace logaritmu na naakumulované hodnoty na výstupu jednotlivých filtrů, čímž se konečně dostáváme do kepstrální oblasti. V té můžeme využít několika, pro nás velice zajímavých vlastností. Jednou z nich je, podobně jako to dělá lidské ucho, omezení dynamiky signálu. Další a nejspíše nejdůležitější z nich je možnost po zlogaritmování oddělit jednotlivé složky signálu, který vznikl konvolucí dvou či více složek v časové oblasti či součinem obrazů ve frekvenční oblasti. Toho lze využít například u filtrace šumu, kde se nepotřebná část signálu jednoduše potlačí. Podmínkou pro účelné použití této vlastnosti je nutnost dostatečné frekvenčně rozdílnosti potlačovaných složek od zbytku signálu. Další informace o této nebo dalších vlastnostech kepstrální oblasti můžeme najít v [1].

Poslední úpravou pro výpočet melovských kepstrálních koeficientů je zpětná diskrétní Fourierova transformace (IDFT). Protože amplitudové spektrum je reálné a symetrické, můžeme výpočet zjednodušit na inverzní diskrétní kosinovu transformaci (IDCT)

$$c_m(j) = \sum_{i=1}^{M^*} \log y_m(i) cos\left(\frac{\pi j}{M^*}(i - 0.5)\right), \tag{15}$$

kde M^* je počet filtrů obsažených v bance melovských filtrů, $j = 0, 1, \cdots, M$ je index melovského kepstrálního koeficientu $c_m(j)$ a M označuje počet melovských kepstrálních koeficientů. Obvykle se počet koeficientů volí v rozmezí $M = 10$ až 13.

Nulový koeficient $c_m(0)$ odpovídá logaritmu energie signálu, spíše se ale nahrazuje logaritmem krátkodobé energie E_n, vypočtené přímo z časových vzorků signálu jak ukazuje rovnice (6)

$$c_m(0) = e_{log} = \log E_n = \log \sum_{k=-\infty}^{\infty} [s^{'}(k) \cdot w(n - k)]^2. \tag{16}$$

Výraznou výhodou MFCC je jejich vysoký informační obsah a zároveň nízká vzájemná korelovanost, která se velice příznivě projevuje například při jejich použití v metodách rozpoznávání řeči založených na skrytých Markovových modelech (HMM).

2.1.7 Dynamické koeficienty

Výpočet MFCC je, jak je výše uvedeno, založen na předpokladu stacionarity řečového ústrojí a tedy i parametrů jej charakterizujících, na určitém krátkém časovém úseku, tzv. mikrosegmentu potažmo okénku. Tento přístup nám přináší pouze statickou informaci o signálu. Účelem dynamických „delta“ Δc_m a akceleračních, tzv. „delta-delta“ $\Delta^2 c_m$ koeficientů je k příslušnému vektoru statických příznaků c_m připojit další koeficienty obsahující dynamickou informaci o signálu. Ve skutečnosti vektor příznaků Δc_m není ničím jiným než derivací vektorů příznaků pro každý mikrosegment, vypočítávané z $2L_1 + 1$ navazujících mikrosegmentů. Stejně tak vektor příznaků $\Delta^2 c_m$ je roven druhé derivaci původních statických příznaků, vypočítávané z $2L_2 + 1$ navazujících mikrosegmentů.

$$[\Delta c_m(j)]_n = \frac{\sum_{\kappa=-L_1}^{L_1} \kappa [c_m(j)]_{n+\kappa}}{\sum_{\kappa=-L_1}^{L_1} \kappa^2}, \tag{17}$$

$$[\Delta^2 c_m(j)]_n = \frac{\sum_{\kappa=-L_2}^{L_2} \kappa [\Delta c_m(j)]_{n+\kappa}}{\sum_{\kappa=-L_2}^{L_2} \kappa^2}, \tag{18}$$

kde $L_1 = L_2 = 1$ představuje rozsah okének, pro výpočet derivací. Výsledný vektor všech příznaků je poté složený, z těchto tří vektorů, každý o velikosti $M + 1$ příznaků

$$c_m^{all} = [c_m(j), \Delta c_m(j), \Delta^2 c_m(j)]. \tag{19}$$

2.2 Proces trénování a rozpoznávání pomocí HTK

V této kapitole stručně popíšeme jakým způsobem probíhá proces rozpoznávání řeči pomocí modulu HTK. Podrobnější informace o tomto procesu je možné nalézt v dokumentu [5]. Samotný modul HTK(Hidden Markov Model Toolkit) je rozsáhlý univerzální nástroj pro vytváření a práci s HMM(Hidden Markov Model). Pro naši úlohu se zaměříme pouze na specifickou část potřebnou k předzpracování dat, procesu trénování HMM a procesu rozpoznávání.

2.2.1 Příprava k trénování

- Prerekvizity k práci s HTK Před samotným započetím procesu rozpoznávání pomocí modulu HTK potřebujeme následující výčet prerekvizit.
  - WAV soubory
  - seznam promluv pro parametrizaci `param.scp`
  - seznam trénovacích promluv `train.scp`
  - seznam testovaných promluv `test.scp`
  - soubor s transkripcí všech promluv na úrovni slov `words.mlf`
  - slovník výslovnosti `dict`
  - soubor obsahující seznam symbolů fonetické abecedy `monophones`
- Vytváření souborů s přepisem na úrovni fonémů
  Stejně jako při psaní jsou jednotlivá slova složená z jednotlivých písmen, jsou při vyslovování slova složená z fónů(monofónů). Náš systém pro rozpoznávání bude založen právě na modelování těchto monofónů. Pro každý monofón ze souboru `monophones` vytvoříme jeden skrytý Markovův model, který jej bude reprezentovat. Proto kromě souboru s transkripcí na úrovni slov `words.mlf`, potřebujeme i soubor s transkripcí těchto slov na úrovni fónů. Ten vytvoříme pomocí programu `HLEd`, obsaženém v balíku HTK, který slouží pro manipulaci se soubory `MLF`.

  ```
  HLEd -l * -d dict.txt -i phones.mlf mkphones.led words.mlf
  ```

`-l *`	Znak * se přidá do jmen souborů v MLF souboru místo skutečné cesty
`-d dict.txt`	Načte slovník výslovnosti.
`-i phones.mlf`	Výstupní soubor s transkripcí na úrovni fónů.
`mkphones.led`	Soubor s příkazy, které se mají provést.
`words.mlf`	Soubor s transkripcí promluv na úrovni slov.

- Parametrizace řečových dat

  Pro parametrizaci promluv, kterou se v této práci převážně zabýváme, slouží v modulu HTK program `HCopy`. Tento program podle zadaných vstupních parametrů dokáže vyčíslit různé druhy posloupností koeficientů (např.: MFCC, PLP, LPC). Význam parametrizace a popis průběhu výpočtu Melovských kepstrálních koeficientů je podrobně popsán v předchozí kapitole (2.1).

  ```
  HCopy -T 1 -C CF_param.mfc -S param.scp
  ```

`-T 1`	Hodnota představuje detailnost výpisů.
`-C CF_param.mfc`	Konfigurační soubor.
`-S param.scp`	Seznam promluv pro parametrizaci.

  V našich experimentech ovšem program `HCopy` nahrazujeme vlastní aplikací `MFCC_calculation`, která umožňuje provést experimentální změny v nastavení metody MFCC.

  ```
  MFCC_calculation param.scp 1 1.0 8000 0.97 13
  ```

`param.scp`	Seznam promluv pro parametrizaci.
`1`	Hodnota odpovídající tvaru frekvenční odezvy filtrů v melovské bance.
`1.0`	Hodnota parametru κ.
`8000`	Vzorkovací frekvence.
`0.97`	Hodnota parametru preemfáze.
`13`	Počet Melovských kepstrálních koeficientů.

2.2.2 Tvorba monofónních modelů a jejich trénování

- Definice topologie HMM a jejich inicializace

  Předpokládá se, že jsme již seznámeny s teorií HMM. Zvolená podoba Markovských modelů je taková, že má pouze tři stavy. Přičemž dva z nich slouží jen ke spojování s dalšími HMM, tudíž zbývá jediný emitující stav. Specifická podoba prototypu HMM, se střeními hodnotami a kovarianční maticí určujících pravděpodobnost jednotlivých vektorů parametrů, je uvedena v souboru `proto`. Stejně tak je v tomto souboru uvedena matice přechodů určující pravděpodobnost přechodu z jednoho stavu do dalšího. Tento prototyp bude rozkopírováním použit jako výchozí nastavení jednotlivých monofónů před samotným trénováním.

  Vektor střední hodnoty a matice kovariance je vypočtena z množiny všech trénovacích dat pomocí programu `HCompV`.

  ```
  HCompV -C CF.mfc -f 0.01 -m -S train.scp -M hmm0 proto
  ```

`-C CF.mfc`	Konfigurační soubor.
`-f 0.01`	Vytvoří makro vFloor1.
`-m`	Signalizuje vypočtení nejen kovariance ale i střední hodnoty.
`-S train.scp`	Seznam trénovacích promluv.
`-M hmm0`	Adresář pro uložení výsledného souboru.
`proto`	Definovaný prototyp.

  Rozkopírování modelu `proto` a jejich přejmenování podle seznamu monofónů do souhrného souboru MMF obstarává příkaz `MakeMMF`.

  ```
  MakeMMF proto monophones vFloors models
  ```

`proto`	Vstupní prototyp modelu.
`monophones`	Seznam monofónů.
`vFloors`	Obsahuje dolní mez kovariance.
`models`	Výstupní soubor s modely monofónů.

- Trénování modelů monofónů

  Dále přistoupíme k samotnému trénování jednotlivých modelů, představujících monofóny. Tuto úlohu má na starost program `HERest`, který je založen na „forward-backward“ algoritmu (viz.[5]). Program vyhledává v `hmm0/MODELS` modely podle názvu, uvedených v `monophones`.

Při nalezení se model aktivuje, reestimuje použitím trénovacích dat v `train.scp` a opět se uloží tentokrát do složky `hmm1`.

```
HERest -T 1 -C CF.mfc -I phones.mlf -t 250.0 150.0 1000.0
       -S train.scp -H hmm0/MODELS -M hmm1 monophones
```

`-C CF.mfc`	Konfigurační soubor.
`-I phones.mlf`	Soubor s monofónní transkripcí trénovacích promluv.
`-t 250.0 150.0 1000.0`	Nastavuje práh prořezávání ve forward-backward algoritmu.
`-S train.scp`	Seznam trénovacích promluv.
`-H hmm0/MODELS`	Vstupní soubor MMF.
`-M hmm1`	Adresář pro uložení výsledného souboru.
`monophones`	Seznam monofónů čili seznam jmen modelů.

Pro kvalitnější natrénování modulů HMM je vhodné trénování provést několikrát. Při dalších průbězích programu se spouštěcí příkaz liší pouze ve vstupních a výstupních adresářích.

- Přerovnání trénovacích dat

  Smysl přerovnání trénovacích dat zjednodušeně spočívá v překladu transkripce na úrovni slov na transkripci na úrovni monofónů. Pokud ve slovníku existuje více možností překladu daného slova, vybere se varianta, která nejvíce odpovídá trénovacím datům a natrénovaným modelům. K tomuto procesu se využívá program `HVite`, založený na Viterbiho algoritmu.

  ```
  HVite -T 1 -l * -y lab -o SWT -b _SIL_ -C CF.mfc -m -a -H
        hmm4/MODELS -i aligned.mlf -t 250.0 -I words.mlf -S
        train.scp dict.txt monophones
  ```

`-l *`	Znak * se přidá do jmen souborů v MLF souboru místo skutečné cesty
`-y lab`	Koncovka jmen souborů v MLF souboru.
`-o SWT`	Formát výstupního MLF souboru.
`-b _SIL_`	Vkládá slovo *_SIL_* na začátek a konec každé promluvy.
`-C CF.mfc`	Konfigurační soubor.
`-m`	Nastavuje výstup na úrovni fonémů.
`-a`	Zajištuje provedení přerovnání trénovacích dat.
`-H hmm4/MODELS`	Vstupní soubor MMF.

`-i aligned.mlf`	Výstupní přerovnaný soubor MLF.
`-t 250.0`	Nastavuje práh prořezávání pro „beam search“.
`-I words.mlf`	Vstupní soubor MMF.
`-S train.scp`	Seznam trénovacích promluv.
`dict.txt`	Rozpoznávací slovník.
`monophones`	Seznam monofónů čili seznam jmen modelů.

Promluvy, které se nepodařilo zarovnat se do `aligned.mlf` neukládají a musí se tedy vyhledat a odstranit i ze seznamu trénovacích promluv `train.scp`. O to se stará program `CreateAligned`.

```
CreateAligned.exe aligned.mlf train.scp aligned.scp ne.scp
```

`aligned.mlf`	Vstupní soubor s přerovnanou transkripcí. trénovacích promluv
`train.scp`	Vstupní seznam trénovacích promluv. promluv.
`aligned.scp`	Výstupní soubor se seznamem dobře přerovnaných promluv.
`ne.scp`	Výstupní seznam nepřerovnaných promluv.

- Trénování přerovnaných modelů monofónů

  Modely se po přerovnání opět několikrát trénují, tentokrát se ale použije nový soubor s monofónním přepisem `aligned.mlf` a místo seznamu trénovacích promluv `train.scp` použijeme přerovnaný seznam `aligned.scp`.

```
HERest -T 1 -C CF.mfc -I aligned.mlf -t 250.0 150.0 1000.0
       -S aligned.scp -H hmm4/MODELS -M hmm5 monophones
```

`-C CF.mfc`	Konfigurační soubor.
`-I aligned.mlf`	Soubor s přerovnanou monofónní transkripcí trénovacích promluv.
`-t 250.0 150.0 1000.0`	Nastavuje práh prořezávání ve forward-backward algoritmu.
`-S aligned.scp`	Seznam trénovacích promluv.
`-H hmm4/MODELS`	Vstupní soubor MMF.
`-M hmm5`	Adresář pro uložení výsledného souboru.
`monophones`	Seznam monofónů čili seznam jmen modelů.

2.2.3 Rozpoznávání

- Rozpoznávání testovacích promluv

  Po dostatečné natrénování monofónního akustického modelu můžeme přejít k hlavnímu kroku a tím je rozpoznávání „neznámých“ promluv. Tyto promluvy, obsažené v `param.scp`, jsou již parametrizované a jejich seznam je uveden v souboru `test.scp`. Program `HVite`, použitý již u přerovnávání dat, použijeme i k účelu rozpoznávání „neznámých“ promluv. Samozřejmě s odlišnými vstupními parametry.

```
HVite -C CF.mfc -H hmm8/models -S test.scp -l * -p -60.0
      -i vysledek_all.txt -w wdnet dict.txt monophones
```

`-C CF.mfc`	Konfigurační soubor.
`-H hmm8/MODELS`	Vstupní soubor MMF.
`-S test.scp`	Seznam testovaných promluv.
`-l *`	Znak * se přidá do jmen souborů v MLF souboru místo skutečné cesty
`-p -60.0`	Hodnota penalizace.
`-i vysledek_all.txt`	Výstupní soubor.
`-w wdnet`	Rozpoznávací síť.
`dict.txt`	Rozpoznávací slovník.
`monophones`	Seznam monofónů čili seznam jmen modelů.

- Zhodnocení úspěšnosti rozpoznávání

  Finálním krokem je zhodnocení úspěšnosti rozpoznávání „neznámých“ promluv. Program `Hresults` porovná přepis rozpoznaných promluvy se skutečnou transkripcí těchto promluv a vypočte úspěšnost a přesnost rozpoznávání. Struktura výstupního souboru, spolu s postupem výpočtu je podrobně popsány v [5].

```
Hresults -f -I words.mlf monophones vysledek_all.txt
         > vysledek.txt
```

`-I words.mlf`	Referenční MLF soubor s přepisy testovaných nahrávek.
`monophones`	Seznam monofónů čili seznam jmen modelů.
`vysledek_all.txt`	Vstupní soubor s výsledky rozpoznávání.
`vysledek.ttx`	Výstupní seznam s vyhodnocením rozpoznávání.

3 Popis dat

Zpracovávaná data jsou telefonní audio nahrávky v pásmu o rozsahu přibližně 0 až $4000Hz$, ve formátu `WAV`. Jednotlivé nahrávky představují promluvy nahrané od mužů a žen, kde každý řečník postupně počítá od 0 do 9. Celý soubor dat tak obsahuje přesně 1341 nahrávek o délce 31 minut a 55 sekund od téměř 140 řečníků, v nichž lehce převažují muži. Počet nahrávek přesně neodpovídá předpokládanému množství, dle počtu řečníků, protože z trénovací množiny byly vyřazeny poškozené či nesrozumitelné promluvy.

Při volbě trénovací množiny jsme vycházeli ze základního logického poznatku. Čím větší trénovací množina je, tím lepší výsledky bude klasifikátor poskytovat. Proto jsme ji zvolili shodnou s celým souborem dat a nazvali ji SI („speaker independent“, což v překladu znamená „na řečníku nezávislá“). Dle zadání, jsme však měli zkoumat výslednou úspěšnost v závislosti na rozpoznávaném řečníku či množině řečníků. Pro vytvoření dalších trénovacích množin jsme nechtěli použít náhodný výběr ale spíše jsme se snažili přijít s určitým obecným kritériem podle, kterého by se dali klasifikovat i jiná data. Jako nejjednodušší a zároveň velmi zajímavé se přímo nabízí rozdělení na muže a ženy. Jak je uvedeno v [1] ženy mají obecně vyšší frekvenci základního hlasivkového tónu F_0, který odpovídá výšce hlasu řečníka, s průměrnou hodnotou $F_0 = 223Hz$. Zatímco u mužů je průměrná frekvence základního hlasivkového tónu $F_0 = 132Hz$. Mohlo by tak být zajímavé zjistit, jaký bude mít tento fakt vliv na výslednou úspěšnost rozpoznávání. Množinu dat obsahující výhradně promluvy získané od žen, jsme nazvali Z. Stejně tak množinu promluv od můžu jsme nazvali M.

Z důvodu toho, že předkládaná práce se snaží ověřit optimální nastavení parametrizační metody MFCC, nikoli kvalitu samotného akustického modelu reci, musíme při volbě testovací sady eliminovat vliv neviděného řečníka. Toho docílíme použitím celé množiny řečníků nejen k trénování, ale také k testování (trénovací a testovací množina jsou schodné). Obě sady jsou tedy také rozděleny shodným způsobem na muže, ženy a řečníka.

4 Experimenty

V této kapitole se budeme věnovat popisu experimentů, prováděných na vstupních datech popsaných v kapitole (3). Těmito experimenty se snažíme ověřit optimálnost původního nastavení parametrizační metody MFCC, co se týče volby tvaru filtrů v melovské bance filtrů a způsobu jejich rozmístění, při použití této metody v systémech rozpoznávání řeči realizovaných Skrytými Markovovými modely(HMM). Programový balík obsahující všechny potřebné nástroje pro vytváření a práci s HMM se nazývá Hidden Markov Model Toolkit(HTK) a manuál k jeho použití můžeme najít na [5]. Konkrétní použití tohoto modulu pro naši úlohu je popsáno v kapitole (2.2).

Nastudováním balíku HTK jsme zjistili, jaké funkce mají jeho jednotlivé části. Parametrizaci dat obstarává aplikace `HCopy`, která umí dle nastavení vypočíst nejen námi upřednostňované koeficienty MFCC ale i další druhy koeficientů jako LPC a PLP. Velice užitečný a univerzální nástroj, který však nenabízí možnost upravení metody MFCC podle našich záměrů a tudíž je pro nás nevyhovující. Byli jsme tak nuceni nahradit tuto aplikaci vlastní, nazvanou `MFCC_calculation`, sloužící pouze ke stanovení koeficientů MFCC ovšem s požadovanou možností zmíněných specifických modifikací. Jako programovací jazyk, ve kterém jsme zmíněný program vytvořili jsme zvolili $C++$ a jako vývojové prostředí nám posloužilo Microsoft Visual Studio 2010 Ultimate.

Nejdříve si v krátkosti projdeme celou metodu MFCC pro upřesnění jejího nastavení a poté si v dalších podkapitolách představíme jednotlivé experimenty, způsob jejich provedení a následně i získané výsledky.

Výsledné hodnoty úspěšnosti rozpoznávání budou reprezentovány pomocí parametru CW uvedeném v souboru `vysledek.txt`, který je výstupem programu z modulu HTK `HResults`. Tento nástroj porovnává oštítkované soubory (výstup z rozpoznávacího nástroje `HVite`) s referenčními přepisy souborů. Parametr CW (Correct Word) představuje procentuální úspěšnost rozpoznání promluv na úrovni slov. Způsob výpočtu tohoto parametru je popsán následujícím vztahem

$$CW = \frac{N - D - S}{N} \cdot 100\%,$$

kde N je celkový počet rozpoznávaných promluv, D je celková suma chyb „vypuštěním“ a S je celková suma chyb „záměnou“.

4.1 Konkrétní postup výpočtu a volba parametrů

Jak již je patrno z názvu, v této podkapitole si v krátkosti projdeme celou metodu MFCC. Jasně si definujeme hodnoty všech potřebných parametrů spolu s popisem myšlenkových pochodů, které k nim vedli, abychom tak předešli možným nejasnostem při implementaci metody a pro možnost opakování těchto experimentů se stejnými výsledky.

Jelikož víme, že vstupní data jsou audio nahrávkami promluv ve frekvenčním pásmu od 0 do $4000Hz$ volíme, podle Shannonova vzorkovacího teorému $f_v \geq 2f_m$, vzorkovací frekvenci $f_v = 8000Hz$. Jako první krok předzpracování signálu aplikujeme na navzorkovaný vstupní signál preemfázi s hodnotou parametru $a = 0.97$. Poté signál rozdělíme na mikrosegmenty o délce $T = 10ms$ obsahující $N = 80$ vzorků a pomocí konvoluce aplikujeme Hammingovo okénko o délce L na odpovídající vzorky vstupního signálu. Počet vzorků v uvnitř okénka volíme, s ohledem na nacházející krok algoritmu (FFT), jako nejbližší mocninu dvou k trojnásobku počtu vzorků v mikrosegmentu čili

$$L \approx 3 * N = 3 * 80 = 240 \approx 2^8 = 256.$$

Interval posunu okének volíme shodný s velikostí mikrosegmentu N. Celkový počet okének P_k, odpovídající počtu kolikrát bude celý proces výpočtu koeficientů MFCC proveden pro danou promluvu, je definován jako

$$P_k = \arg\min_{j \in N} \left| \frac{k - L}{N} - j - 0.5 \right| = \arg\min_{j \in N} \left| \frac{k - 256}{80} - j - 0.5 \right|,$$

kde k je počet všech vzorků v dané promluvě. Zbývající vzorky, kterých už není dostatek aby vyplnili celé okénko jsou zapomenuty. Nejedná se ovšem o přílišnou ztrátu informace, protože tyto vzorky v naprosté většině případů odpovídají tichu a pro proces rozpoznávání tak nejsou důležité.

Po rozdělení signálu na okénka postačuje popsat postup výpočtu koeficientů MFCC jen pro jediné, protože je pro všechny okénka totožný. Dalším krokem je výpočet krátkodobé diskrétní Fourierovi transformace(DFT) pomocí algoritmu(FFT). Vstupem do FFT je pak mikrosegment s $N = 256$ vzorky v časové oblasti a výstupem je 256 vzorků komplexního signálu obsahujícího reálnou a imaginární složku popisující frekvenční spektrum okénka. Pro naše potřeby je ale ještě zapotřebí převod komplexního spektra do amplitudového pomocí výpočtu absolutní hodnoty čili vzdálenosti bodu od počátku na souřadnicích odpovídajících reálné a imaginární složce v komplexní rovině podle rovnice (9).

Ve výsledku tak dostáváme $\frac{N}{2} = 128$ vzorků, které rovnoměrně popisují frekvenční spektrum na zkoumaném rozsahu od 0 do $4000Hz$ po intervalech f_{int}. Velikost tohoto intervalu je

$$f_{int} = \frac{f_{max}}{\frac{N}{2}} = \frac{4000}{128} = 31.25Hz.$$

Pro podrobnější popis frekvenčního spektra jsou informace mezi těmito intervaly dosažitelné zvýšením vzorkovací frekvence f_v.

Následně na amplitudové spektrum okénka aplikujeme banku melovských pásmových filtrů podle rovnice (13). Počet těchto filtrů v bance pro zmíněný frekvenční rozsah je, podle tabulky (1), $M^* = 15$. Z dvou různých způsobů implementace melovské banky filtrů, popsaných v kapitole (2.1.5) volíme logicky ten výpočetně úspornější postup, tedy přepočet vzorků lineárně rozmístěných filtrů, po intervalu Δ_m, v melovské frekvenční škále do škály lineární pomocí inverzní funkce k rovnici (11). Velikost intervalu Δ_m si vypočteme pomocí rovnice (20), kde se podle tabulky (1) hodnota $f_{max} = 2146mel$.

$$\Delta_m = \arg\min_{k \in N} \left| \frac{f_{max}}{M^* + 1} - k \right| \tag{20}$$

Pro první experiment si definujeme nový parametr $t_{filtr} \in \{1, 2, 3, 4, 5\}$ kde jednotlivá čísla označují index tvaru filtrů v melovské bance (trojúhelník, obdélník, cosinus, sinus, lichoběžník), což má vliv na podobu rovnice (14).

Pro druhý experiment se, zavedením parametru κ, modifikuje rovnice (11). Zároveň se s hodnotou κ mění maximální frekvence f_{max} v nelineární škále odpovídající horní hranici zkoumaného frekvenčního rozsahu $f_m = 4000Hz$ ve škále lineární. Tento fakt ovlivní i volbu intervalu Δ_m.

Pro zavedení příznaku kepstrální analýzy zlogaritmujeme výstupní hodnoty filtrů a provedeme zpětnou diskrétní cosinovu transformaci dle rovnice (15), čímž dostáváme vypočtené jednotlivé kepstrální koeficienty $c_m(j)$, kde j označuje index koeficientu a pohybuje se v rozmezí $j = 0, 1, \cdots, M$. Hodnota M se obvykle volí v rozmezí $M \in \langle 10, 13 \rangle$. Pro tuto implementaci volíme $M = 12$. Kde původní koeficient $c_m(0)$ nahrazujeme logaritmem krátkodobé energie signál E_n podle rovnice (16). Závěrečným krokem algoritmu je vypočtení dynamických koeficientů Δc_m a $\Delta^2 c_m$. Při jejich výpočtu je nutné definovat výpočet těchto koeficientů pro okrajové mikrosegmenty. Podle rovnic (17), (18) potřebujeme pro výpočet dynamických koeficientů, melovské kepstrální koeficienty od dvou předchozích a dvou následujících mikrosegmentů. Z těchto důvodů nastává problém vyčíslení dynamických koeficientů

pro zmíněné okrajové mikrosegmenty. V některých postupech se chybějící informace doplňují různými kombinacemi okolních vektorů koeficientů.

Po uvážení faktu, že se jedná pouze o informace z počátečních a konečných $20ms$ každé promluvy, které v naprosté většině případů představují ticho, jsme se pro zjednodušení rozhodli pro tyto mikrosegmenty dynamické koeficienty nevypočítávat a celkově je dál neuvažovat.

4.2 Experimentování s tvarem filtrů

Na původní trojúhelníkové bance melovských filtrů si můžeme všimnout dvou zajímavých vlastností (A,B) v její konstrukci. První a nejvíce nápadný rys A, je vzájemné překrývání filtrů a to takovým způsobem, že na střední frekvenci i-tého filtru, končí pásmo $i-1$-tého filtru a zároveň začíná pásmo $i+1$-tého filtru. Díky značné úspoře souřadnic nutných pro jednoznačné definování podoby banky je tento způsob téměř optimální pro implementaci banky filtrů v programovém prostředí. Druhým rysem B je, že celkový zisk banky, až na výjimky v krajních intervalech, se rovná jedné. To má za důsledek rovnoměrné vážení celého spektra vstupního signálu bez upřednostňování či potlačování jakékoliv frekvence, čímž se zachovává maximální nezkreslená informace o signálu. Obě vlastnosti A a B se vzájemně ovlivňují ale nejsou na sobě závislé, jak se můžeme přesvědčit na příkladech níže. Frekvenční odezvy alternativních filtrů, jsme volili právě tak, aby obsahovali kombinace zmíněných rysů. Pro přehled, rys A obsahují frekvenční odezvy filtrů 2(obdélník), 3(cosinus), 4(sinus) a rys B odezvy filtrů 2, 3, 5(lichoběžník). Trojúhelníkový tvar má označení 1 a proto ho ve výčtu neuvádíme. Samotný experiment je založený na tom, zda a jak tyto vlastnosti banky či samotné tvary filtrů jsou důležité pro kvalitu výsledných MFCC koeficientů.

Před přiblížením jednotlivých filtrů si nadefinujeme hodnoty b_i mající význam hraničních frekvencí všech pásmových intervalů. Tím pádem obsahují všechny střední a zároveň hraniční frekvence filtrů, přepočtené z melovské frekvenční škály do lineární. Hodnoty b_i jsou společné pro každý tvar filtru s výjimkou filtru 5 kde se používá jiné rozestavení banky filtrů pomocí souřadnic b_j. Pro rovnice (23) až (27) definující tvar filtrů nabývá index i hodnot pouze $1, 2, \cdots, M^*$ a označuje tak pořadí filtru i index hraniční frekvence.

$$b_i = 700 \cdot \left(10^{\left(\frac{\Delta_m i}{2595}\right)} - 1\right); \quad i = 0, 1, \cdots, M^* + 1 \tag{21}$$

$$b_j = 700 \cdot \left(10^{\left(\frac{\frac{\Delta_m j}{2} + \frac{\Delta_m}{4}}{2595}\right)} - 1\right); \quad j = 0, 1, \cdots, 2(M^* + 1) \tag{22}$$

Kvůli periodicitě goniometrických funkcí na intervalu $(0; 2\pi)$, využívaných u filtrů 3, 4 jsme byli nuceni, pro jejich konstrukci na různých intervalech, zavést dvě nové proměnné. Hodnota C_i označuje počet vzorků splňujících podmínku v rovnici (24) čili počet vzorků obsažených v pásmu filtru i. Hodnota c má význam pořadí vzorku splňujícího tutéž podmínku. Tak zajistíme plynulý diskrétní průběh jednotlivých filtrů po krocích o velikosti $\frac{1}{C_i}$.

4.2.1 Trojúhelníkový filtr

Originální tvar frekvenční odezvy filtrů je zde uváděn pro objektivní porovnání s alternativními odezvami. Konkrétně, aby se zajistilo, že koeficienty budou kalkulovány stejným postupem pro všechny filtry.

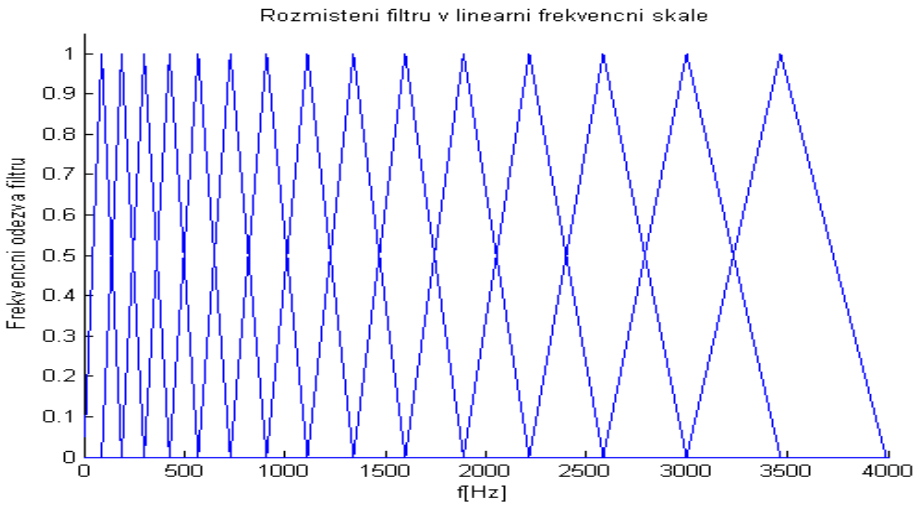
- Definice

  Trojúhelníkový filtr, má bezpochyby výhodu v jednoduchosti a linearitě provedení. Z jeho tvaru je vidět, že propouští beze změny pouze vzorky odpovídající jeho střední frekvenci $b_{m,i}$, ostatní vzorky v jeho okolí $(b_{m,i}-\Delta_m, b_{m,i}+\Delta_m)$ potlačuje přímo úměrně se vzrůstající vzdáleností od $b_{m,i}$ až k nule.

$$u(f,i) = \begin{cases} \frac{f-b_{i-1}}{b_i-b_{i-1}} & \text{pro } b_{i-1} \leq f < b_i \\ \frac{f-b_{i+1}}{b_i-b_{i+1}} & \text{pro } b_i \leq f < b_{i+1} \\ 0 & \text{pro ostatní.} \end{cases} \tag{23}$$

  kde $i = 1, 2, \cdots, M^*$ označuje pořadí filtru a zároveň index hraniční frekvence.

- Podoba banky



Obrázek 6: Banka melovských trojúhelníkových filtrů

- Tabulka výsledků rozpoznávání

	Testovací množina		
Trénovací množina	Muži[%]	Ženy[%]	Všichni[%]
Muži	59.08	44.74	52.22
Ženy	50.14	65.93	57.70
Všichni	57.49	63.11	60.18

Tabulka 2: Tabulka zobrazující úspěšnost rozpoznávání při použití trojúhelníkového tvaru filtru

- Zhodnocení

  Na výsledcích v tabulce můžeme pozorovat závislost úspěšnosti rozpoznávání na trénovací množině řečníků. Pro trénovací množinu řečníků složenou z mužů (dále pouze muži, ženy a všichni) dosáhla nejlepších výsledků opět rozpoznávaná množina mužů. Podobně tak pro množinu žen, byl proces nejúspěšnější na rozpoznávané množině žen s překvapivě vyššími hodnotami. Výjimku tvoří množina všichni, pro kterou stejná množina vykazuje průměrnou úspěšnost rozpoznávání, což je logické s ohledem na to, že množina všichni je tvořená sjednocením množin muži a ženy. Toto chování je, jak se můžeme přesvědčit v ostatních tabulkách, schodné pro všechny druhy filtrů, z toho se dá vyvodit, že to nezávisí na volbě filtru, ale pouze vychází z vlastností vstupních dat. Proto, budeme uvažovat výsledky této části experimentu jako referenční k porovnání se zbylými částmi.

4.2.2 Obdélníkový filtr

Vybrali jsme tento tvar, protože stejně jako originální trojúhelníkový filtr 1 splňuje obě vlastnosti A i B, popsané výše. Chtěli jsme tak ověřit zda tyto rysy jsou rozhodující pro výslednou úspěšnost rozpoznávání a jestli by bylo možné oba tvary zaměnit s obdobnou přesností.

- Definice

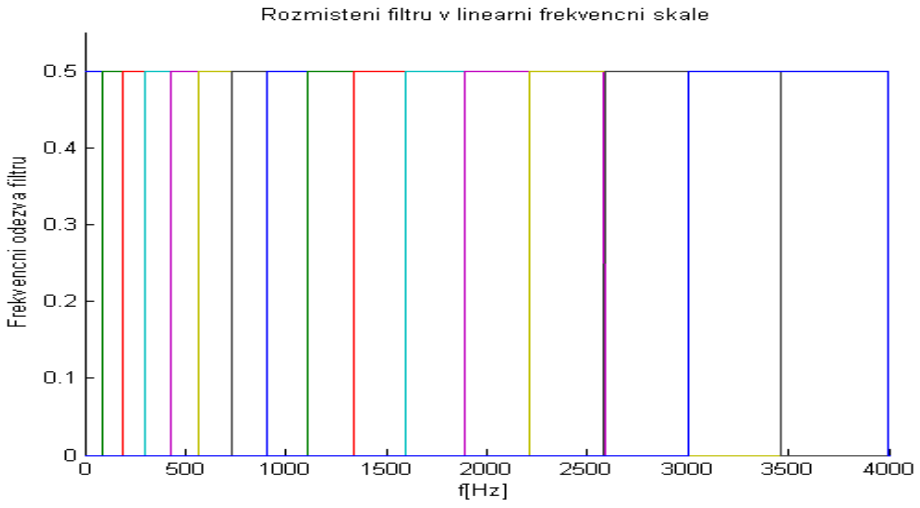
  Funkci filtru lze prakticky popsat jako vybírání vzorků spadajících do daného frekvenčního intervalu a jejich rovnoměrné potlačení na polovinu. Bez tohoto útlumu by podmínka A nebyla splněna v takové podobě v jaké jsme si ji definovali.

$$u(f,i) = \begin{cases} 0.5 & \text{pro } b_{i-1} \leq f < b_{i+1} \\ 0 & \text{pro ostatní.} \end{cases} \tag{24}$$

  kde $i = 1, 2, \cdots, M^*$ označuje pořadí filtru a zároveň index hraniční frekvence.

- Podoba banky

  Kvůli absultnímu vzájemnému překrývání, jsme museli pro rozlišitelnost jednotlivý filtrů použít odlišnou barvu pro každý z nich.



Obrázek 7: Banka melovských obdélníkových filtrů

- Tabulka výsledků rozpoznávání

	Testovací množina		
Trénovací množina	Muži[%]	Ženy[%]	Všichni[%]
Muži	52.02	41.13	46.81
Ženy	45.97	56.99	51.24
Všichni	51.30	49.76	50.56

Tabulka 3: Tabulka zobrazující úspěšnost rozpoznávání při použití obdélníkového tvaru filtru

- Zhodnocení

  Ve srovnání s hodnotami naměřenými u trojúhelníkového filtru, dosahuje obdélníkový filtr o poznání horších výsledků. Zejména pak při rozpoznávání množin schodných s trénovacími, kde pokles dosahuje téměř k deseti procentům.

4.2.3 Cosinusový filtr

Stejně jako filtry $1, 2$ i tento filtr splňuje obě vlastnosti A, B. Jedná se tak o další porovnatelný tvar, na rozdíl od předchozích s nelineárním charakterem.

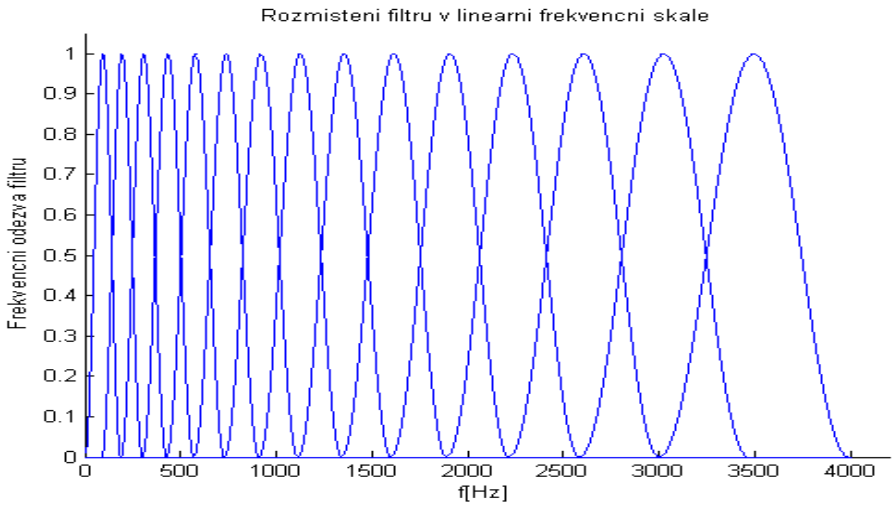
- Definice

  Tento filtr, jak je vidno z průběhu, silněji potlačuje vzorky v krajních čtvrtinách a naopak méně potlačuje vzorky uprostřed intervalu.

$$u(f,i) = \begin{cases} 0.5 - 0.5\cos(\frac{2\pi c}{C_i}) & \text{pro } b_{i-1} \leq f < b_{i+1} \\ \\ 0 & \text{pro ostatní.} \end{cases} \tag{25}$$

  kde $i = 1, 2, \cdots, M^*$ označuje pořadí filtru a zároveň index hraniční frekvence. .

- Podoba banky



Obrázek 8: Banka melovských cosinusových filtrů

- Tabulka výsledků rozpoznávání

	Testovací množina		
Trénovací množina	Muži[%]	Ženy[%]	Všichni[%]
Muži	53.31	44.27	48.99
Ženy	47.26	62.32	54.47
Všichni	52.88	56.51	54.62

Tabulka 4: Tabulka zobrazující úspěšnost rozpoznávání při použití cosinusového tvaru filtru

- Zhodnocení

  Oproti klasickému trojúhelníkovému filtru je úspěšnost rozpoznávání, stejně jako v předešlém kroků, nižší i když ne tak rapidně. Zvláště na trénovací množině žen se celkově nejvíce snížila úspěšnost.

4.2.4 Sinusový filtr

První z alternativních filtrů, který splňuje pouze vlastnost A. Dochází tak ve výsledku k nerovnoměrnému vážení vzorků frekvenčního spektra vstupního signálu. To znamená, že pokud bychom sečetli dohromady zisky všech filtrů nebyly by rovné jedné na celém frekvenčním intervalu čili nesplnění vlastnosti B.

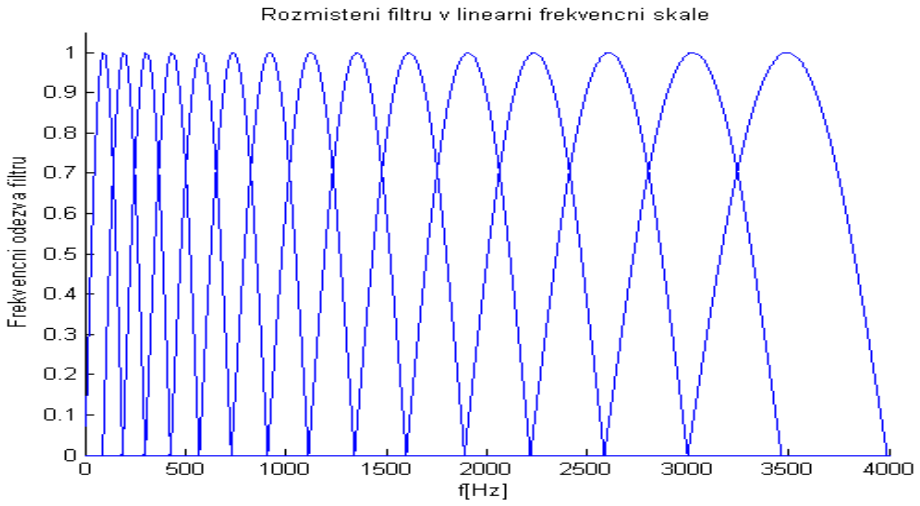
- Definice

  Filtr pozvolna nelineárně utlumuje vzorky v okolí středních frekvencí filtrů b_i, se vzrůstající vzdáleností se ale strmost útlumu zvyšuje až k maximu.

$$u(f,i) = \begin{cases} \sin(\frac{\pi c}{C_i}) & \text{pro } b_{i-1} \leq f < b_{i+1} \\ \\ 0 & \text{pro ostatní.} \end{cases} \tag{26}$$

  kde $i = 1, 2, \cdots, M^*$ označuje pořadí filtru a zároveň index hraniční frekvence.

- Podoba banky



Obrázek 9: Banka melovských sinusových filtrů

- Tabulka výsledků rozpoznávání

	Testovací množina		
Trénovací množina	Muži[%]	Ženy[%]	Všichni[%]
Muži	52.16	43.80	48.16
Ženy	45.82	60.44	52.82
Všichni	53.89	55.26	54.55

Tabulka 5: Tabulka zobrazující úspěšnost rozpoznávání při použití sinusového tvaru filtru

- Zhodnocení

  Úspěšnost rozpoznávání sinusového filtru je opět nižší oproti původnímu trojúhelníkovému, je ovšem srovnatelná s předcházejícím cosinovým filtrem.

4.2.5 Lichoběžníkový filtr

Pro tento filtr jsme navrhli nové rozmístění filtrů b_j, taktéž lineární v melovské škále jako to původní, tak aby byl splněn požadavek pouze na platnost rysu B. Můžeme tak, dle výsledků, nepřímo posoudit důležitost vlastnosti A.

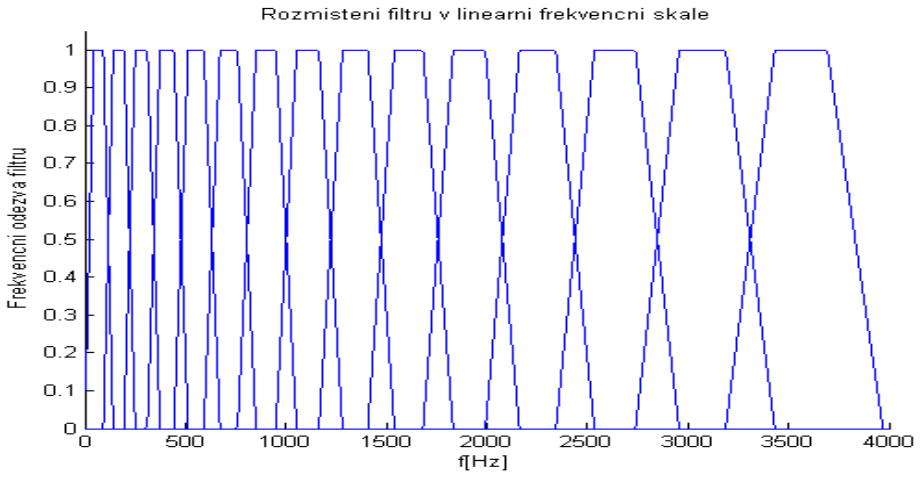
- Definice

  Průběh odezvy filtru odpovídá v krajních čtvrtinách filtru 1 ovšem s dvojnásobnou strmostí útlumu. Prostřední polovina frekvenčního intervalu ponechává vzorky nedotčené.

$$u(f,i) = \begin{cases} \frac{f-b_{j-1}}{b_j-b_{j-1}} & \text{pro } b_{j-1} \leq f < b_j \\ 1 & \text{pro } b_j \leq f < b_{j+1} \\ \frac{f-b_{j+2}}{b_{j+1}-b_{j+2}} & \text{pro } b_{j+1} \leq f < b_{i+2} \\ 0 & \text{pro ostatní.} \end{cases} \tag{27}$$

  V tomto novém rozmístění, které popisuje rovnice (22) je nutno vyčíslit dvojnásobné množství hraničních frekvencí indexovaných parametrem j. Je také nutné definovat přepočet mezi dvojím indexováním $j = 2i-1$, kde i nabývá hodnot $1, 2, ..., M^*$.

- Podoba banky



Obrázek 10: Banka melovských lichoběžníkových filtrů

- Tabulka výsledků rozpoznávání

	Testovací množina		
Trénovací množina	Muži[%]	Ženy[%]	Všichni[%]
Muži	59.37	43.49	51.77
Ženy	48.56	66.72	57.25
Všichni	55.91	60.75	58.23

Tabulka 6: Tabulka zobrazující úspěšnost rozpoznávání při použití lichoběžníkového tvaru filtru

- Zhodnocení

  Lichoběžníkový filtr jako jediný ze zkoušených filtrů podává srovnatelné výsledky s trojúhelníkovým filtrem. V některých případech dokonce byla úspěšnost rozpoznávání mírně lepší, například u rozpoznávání množin shodných s trénovacími. Průměrná hodnota, která v tabulkách,pro naše data, odpovídá hodnotě pro trénovací množinu „Všichni“ a shodnou testovací množinu řečníků, byla ovšem pro lichoběžníkový filtr o několik desetin horší.

4.2.6 Zhodnocení experimentu

V závěru experimentu vzájemně zhodnotíme kvalitu filtrů a seřadíme je podle úspěšnosti rozpoznávání. Jako nevhodnější se skutečně ukázal původní trojúhelníkový filtr 1. Hned za ním se ovšem podle výsledků umístil lichoběžníkový filtr 5, který podával srovnatelně dobré výsledky. Výrazně horších výsledků dosahoval na třetí pozici cosinusový filtr 3. Jen ještě o málo hůř si vedl sinusový filtr 4 a jako naprosto nevyhovující se ukázal obdélníkový filtr 2. Z pořadí se dá vyčíst, že při návrhu tvaru alternativního filtru, nehrají rysy A a B až tak podstatnou roli na kvalitu vypočtených koeficientů MFCC. Spíše se zdá, že v experimentu obstáli nejlépe lineární tvary filtry, které preferovali určitou střední frekvenci b_i či b_j.

	Trénovací množina / Testovací množina		
Tvar filtru	Muži/Muži	Ženy/Ženy	Všichni/Všichni
1 [%]	59.08	65.93	60.18
2 [%]	52.02	56.99	50.56
3 [%]	53.31	62.32	54.62
4 [%]	52.16	60.44	54.55
5 [%]	59.37	66.72	58.23

Tabulka 7: Tabulka zobrazující úspěšnost rozpoznávání v případech shodné množiny trénovacích a testovaných dat pro všechny tvary filtrů

4.3 Experimetování s rozmístěním filtrů

Tímto experimentem se pokusíme konfrontovat původní rozmístění filtrů na lineární škále v melovské bance, zastoupené rovnicí (21), s alternativními rozmístěními. Inverzní rovnice (11) ke zmíněnému vztahu (21) je experimentálně získaná pomocí psychoakustických pokusů prováděných na rozsáhlé množině posluchačů, popisující zjištěnou nelineárně klesající citlivost lidského sluchu s rostoucí frekvencí akustického signálu. Budeme tak ověřovat ideálnost originálního rozmístění oproti jiným možnostem, realizovaných zavedením parametru κ na různých množinách trénovacích a rozpoznávaných dat. Zároveň se přitom nabízí příležitost zkoumat možné rozdíly v úspěšnosti rozpoznávání mezi ženami a muži pro různá nastavení experimentu.

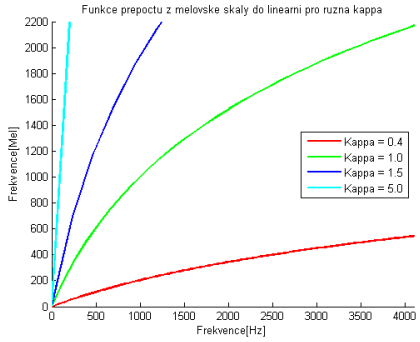
Způsob zavedení bezrozměrného parametru κ do rovnice (21) částečně kompenzující nelinearitu citlivosti lidského ucha je znázorněn v následující rovnici

$$b_i = \frac{700}{\kappa} \cdot \left(10^{\left(\frac{\Delta_m i}{2595\kappa}\right)} - 1 \right); \quad i = 0, 1, \cdots, M^* + 1. \tag{28}$$

Abychom při změnách této křivky dosáhli, při přepočtu z nelineární (již ne melovské) škály do lineární, pokrytí stejného frekvenčního pásma 0 až $4000Hz$ jsme nuceni zpětně přepočítávat tyto souřadnice. Bod představující 0 je neměnný. Pro opačný konec intervalu vyčíslujeme hodnotu f_{max}, níže uvedeným způsobem

$$f_{max} = 2595\kappa \log_{10} \left(\frac{f_v}{2} \cdot \frac{\kappa}{700} + 1 \right). \tag{29}$$

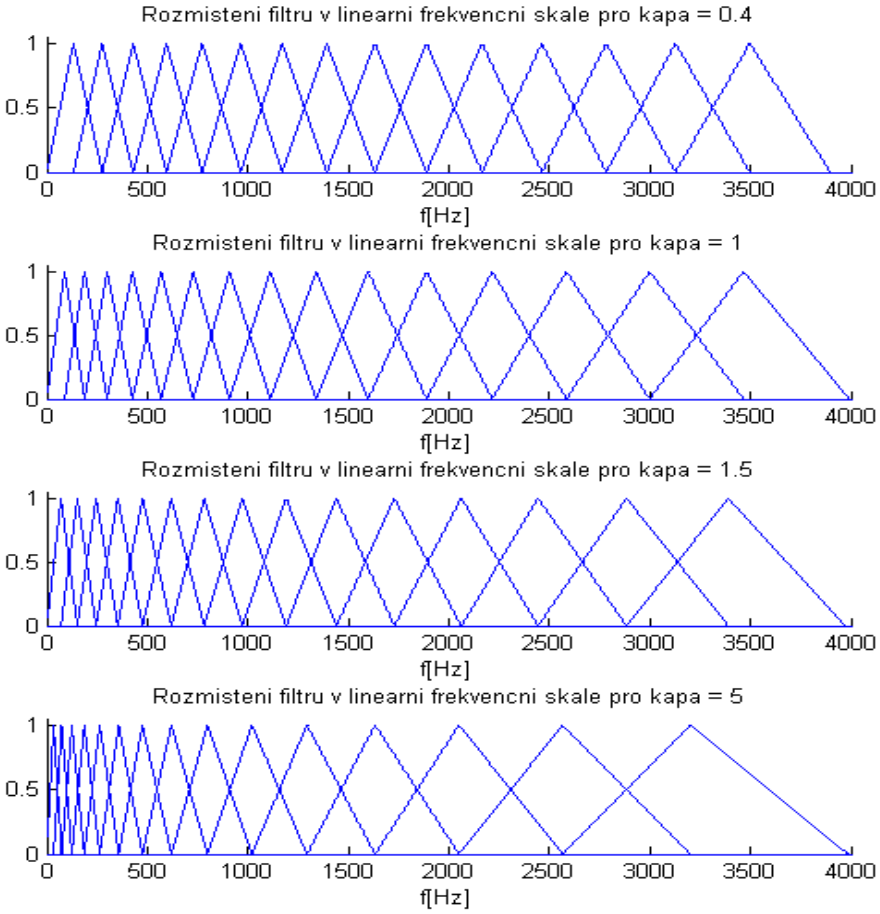
Tento vztah má samozřejmě vliv na hodnotu parametru Δ_m v rovnici (20) a nepřímo tak ovlivňuje i hraniční frekvence b_i v rovnici (28).



Obrázek 11: Funkce prepočtu z melovské škály do lineární pro různá κ

4.3.1 Ukázka rozložení filtrů v bankách

Nepřesnosti v rozložení grafů jsou způsobeny chybou zaokrouhlování na celé čísla při vypočtu Δ_m. Nejznatelněji je tento jev vidět pro nízké hodnoty κ, protože jejich maximální frekvence f_{max} je oproti vyšším hodnotám κ taktéž nízká.



Obrázek 12: Ukázka rozložení filtrů v melovských bankách pro různé hodnoty κ.

4.3.2 Výsledky experimentu

Experiment jsme rozdělili pro přehlednost do tří částí podle trénovací množiny dat. Můžeme tak přímo sledovat jak působí rozdílné nastavení parametru κ na úspěšnost rozpoznávacího procesu.

- Muži

kapa	Muži[%]	Ženy[%]	Všichni[%]
0.2	53.75	37.68	46.06
0.4	53.75	37.99	46.21
0.6	51.15	45.84	48.61
0.8	55.48	45.68	50.79
1.0	59.08	44.74	52.22
1.2	53.03	42.54	48.01
1.5	45.82	38.62	42.37
2.0	47.26	38.78	43.20
3.5	42.51	35.01	38.92
5.0	49.57	37.36	43.73

Tabulka 8: Tabulka zobrazující úspěšnost rozpoznávání pro trénovací množinu řečníků mužů v závislosti na změně κ

- Ženy

kapa	Muži[%]	Ženy[%]	Všichni[%]
0.2	46.69	58.24	52.22
0.4	47.12	56.99	51.84
0.6	48.13	61.38	54.47
0.8	50.29	63.11	56.42
1.0	50.14	65.93	57.70
1.2	46.69	67.35	56.57
1.5	43.52	63.27	52.97
2.0	43.23	54.63	48.69
3.5	41.50	51.18	46.13
5.0	46.11	58.71	52.14

Tabulka 9: Tabulka zobrazující úspěšnost rozpoznávání pro trénovací množinu řečníků žen v závislosti na změně κ

- Všichni

kapa	Muži[%]	Ženy[%]	Všichni[%]
0.2	51.15	47.72	49.51
0.4	47.69	54.16	50.79
0.6	54.90	59.50	57.10
0.8	57.49	61.22	59.28
1.0	57.49	63.11	60.18
1.2	55.19	60.60	57.78
1.5	49.71	54.16	51.84
2.0	44.96	48.19	46.51
3.5	43.23	44.74	43.95
5.0	52.02	53.69	52.82

Tabulka 10: Tabulka zobrazující úspěšnost rozpoznávání pro celou trénovací množinu v závislosti na změně κ

4.3.3 Zhodnocení experimentu

Z výsledků uvedených v tabulkách je skutečně patrný vliv parametru κ, který má značně záporný charakter u obou výsledných parametrů ať už se hodnota κ snižuje či zvyšuje. Nejoptimálnějším nastavením se tak, na trénovaných datech, opravdu ukázalo to původní $\kappa = 1$. Pouze na nejbližším okolí $\kappa = 1 \pm 0.2$ dosahuje proces rozpoznávání podobných výsledků, některých případech dokonce lehce lepších. Analýzou výsledků jsme se také přesvědčili o tom, že původní křivka představující nelineární citlivost sluchu je optimální pro obě pohlaví alespoň v závislosti na zavedeném parametru κ i přes rozdíly v procesu vytváření řeči a odlišných hodnotách základních hlasivkových tónů.

5 Závěr

V počátku práce jsme se věnovali teoretickému rozboru algoritmu výpočtu metody Melovských kepstrálních koeficientů. Poté jsme si nastínili postup vytváření a trénování skrytých Markovských modulů a rozpoznávání řečového signálu, použitím modulu HTK, v závislosti na množině řečníků, rozdělených na muže, ženy a smíšenou množinu. V dalších kapitolách jsme se zabývali popisem prováděných experimentů, ve kterých jsme ověřovali optimálnost původního nastavení metody MFCC.

V prvním experimentu jsme se zaměřili na tvar frekvenční odezvy filtrů obsažených v melovské bance filtrů, který je v originále trojúhelníkový. My jsme ho postupně nahrazovali čtyřmi odlišnými tvary (obdélníkový, cosinusový, sinusový a lichoběžníkový). Porovnáním procentuální úspěšnost rozpoznávání pro různé trénovací a testovací množiny řečníků, jsme zjistili, že všechny tvary frekvenčních odezev filtrů jsou vůči trojúhelníkovému tvaru znatelně méně úspěšné, s výjimkou lichoběžníkového tvaru, který podával srovnatelné, nicméně v průměru lehce horší, výsledky. Ověřili jsme tak optimalitu volby trojúhelníkového tvaru frekvenční odezvy ze zkoušené množiny tvarů.

V druhém experimentu jsme se pozastavili nad podobou rovnice kompenzující nelinearitu citlivosti vnímání akustického signálu lidským uchem v závislosti na frekvenci, která má zásadní vliv na rozmístění filtrů v melovské bance filtrů. Testovali jsme význam tohoto rozmístění filtrů na výslednou úspěšnost identifikace promluv na shodných množinách řečníků jako v prvním experimentu. Zavedením parametru κ do zmíněné rovnice jsme mohli lehce modifikovat podobu citlivostní křivky pro námi zvolené, potenciálně zajímavé, hodnoty. Z celé množiny výsledků uvedených v tabulkách je zřetelně vidět trajektorie monotónně stoupající s ubývající vzdáleností od maxima umístěného v okolo hodnoty 1.0, která odpovídá původnímu rozmístění filtrů v melovské bance. Dokázali jsme tak, že na všech námi testovaných množinách a pro zvolený způsob zavedení parametru κ je originální rozmístění pásmových filtrů skutečně nejvhodnější.

Zhodnocením obou provedených experimentů můžeme dojít k závěru, že pro naše podmínky je původní nastavení parametrizační metody MFCC skutečně optimální.

Seznam obrázků

Seznam tabulek