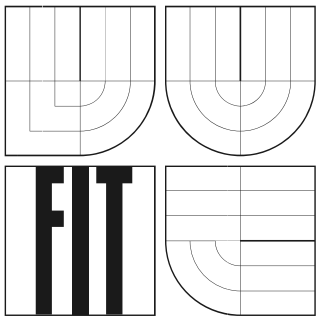
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ



Robustní parametrizace řeči ve frekvenční oblasti

Ročníková práce

2005 Jaroslav Šebesta

Robustní parametrizace řeči ve frekvenční oblasti

Odevzdáno na Fakultě informačních technologií Vysokého učení technického v Brně, dne 1. června 2005.

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Petra Motlíčka, Ph.D. a Ing. Lukáše Burgeta Ph.D.
Další informace mi poskytli ostatní členové skupiny zabývající se zpracováním signálů..
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Jaroslav Šebesta
Datum

Abstrakt

Ročníkový projekt je zaměřen na problematiku zpracování řečového signálu , jeho parametrizaci a následné použití těchto parametrů v úloze robustního rozpoznávání řeči pomocí neuronové sítě. Hlavním úkolem je použití nelineárního spektrálního vyhlazování, zjištění přínosu tohoto vyhlazení na robustnost rozpoznávače, porovnání této metody úpravy signálu s jinými používanými postupy a možnost použití toho algoritmu i v rozpoznávačích s větším počtem neuronů.

Klíčová slova

Řečový signál, parametrizace, HTK toolkit, zpracování signálu, cepstrum, HMM rozpoznávání, robustní rozpoznávač, MFCC, rozpoznávání řeči, nelineární spektrální vyhlazování, RASTA filtr

Obsah

Zadání

Téma práce: **Robustní parametrizace řeči ve frekvenční oblasti**

Vedoucí:

Motlíček Petr, Ing., PhD., UPGM FIT VUT

Přihlášen:

Šebesta Jaroslav

Zadání:

1. Seznamte se s klasickými přístupy v parametrizaci řeči ve spektrální oblasti.
2. Zaměřte se na nové algoritmy parametrizace pomocí nelineárního spektrálního vyhlazování.
3. Otestujte nově odvozené řečové parametry v úloze robustního rozpoznávání řeči.

Kategorie:

Signály a systémy

Implementační jazyk:

C, Matlab, bash

Operační systém:

Linux, Windows

Komerční software:

matlab

Literatura:

- B. Gold, N. Morgan. Speech and Audio Signal Processing. John Wiley & Sons, Inc. 1999.
- M. K. Carey. Robust Speech Recognition Using Non-Linear Spectral Smoothing, Eurospeech 2003, Geneve, CH, September 2003.

Komentář:

1. Proveďte první experimenty s klasickými řečovými parametry používanými v rozpoznávání řeči.
2. Použijte databázi TI-DIGITs, proveďte první experimenty s různými úrovněmi SNR.
3. Proveďte první experimenty s nelineárním vyhlazováním ve spektrální oblasti.

1 Úvod

Téma tohoto ročníkového projektu jsem si vybral z důvodu zájmu o zpracování signálů, zvláště pak řeči. Hlavní částí zpracování řeči je její parametrizace, jejíž výsledek může být následně použit k různým účelům ať už pro přenos signálu, nebo např. pro rozpoznávání mluveného textu. Právě rozpoznávání mluveného slova pro mě bylo neznámou a proto jsem se rozhodl více tuto oblast prozkoumat.

Mým úkolem bylo seznámit se s problematikou zpracování řečového signálu, zvláště pak zpracování tohoto signálu za účelem dalšího použití pro jeho rozpoznávání a použití nelineárního spektrálního vyhlazování pro zlepšení rozpoznávání.

Nejprve se zmíním o zpracování a parametrizaci řečového signálu obecně, potom podrobněji rozeberu použití nelineárního spektrálního vyhlazování. V dalších kapitolách pak použití této parametrizace v úloze robustního rozpoznávání řeči. V závěrečné kapitole zhodnotím práci a výsledky, kterých jsem dosáhl a také možnost dalšího vývoje tohoto projektu vzhledem k diplomové práci.

2 Zpracování řečového signálu

2.1 Obecná charakteristika

Digitální zpracování řečového signálu se stalo běžnou součástí dnešního života. Mobilní telefon již používá téměř každý, ale v čím dál větší míře se setkáváme také s automatickým rozpoznáváním řeči. Hlasové ovládání mobilního telefonu, počítače, či automatické rozpoznání čísla, znaku nebo celého slova automatem na druhém konci telefonního spojení je spojeno s touto oblastí zpracování zvukového signálu.

Nároky kladené na rozpoznání mluveného textu jsou hlavně na úspěšnost správného rozpoznání, popřípadě schopnost určit s jakou přesností byl výsledek vyhodnocen.

Pro komerční využití je tedy žádoucí aby při dostatečné době potřebné k rozpoznání byla zachována co největší přesnost výsledku(požadavky se samozřejmě mohou lišit).

Jedním možným způsobem, jak rozpoznat mluvený text, je mít rozsáhlou databázi na základě které porovnáme rozpoznávanou řeč se vzorky v této databázi. Řečový signál je však příliš závislý na konkrétním řečníkovy a ani stejný řečník neřekne jednu větu dvakrát stejně, proto provádíme parametrizaci.

2.2 Parametrizace

Úkolem parametrizace je vyjádřit řečový signál omezeným množstvím hodnot. Parametrizací úseku řečového signálu získáme daný počet hodnot. Pro více řečových úseků řadíme tyto hodnoty do matice. Pro rozpoznávače jsou vhodné parametry, které jsou:

- gaussovského rozložení
- dekorelované
- málo dimenzionální

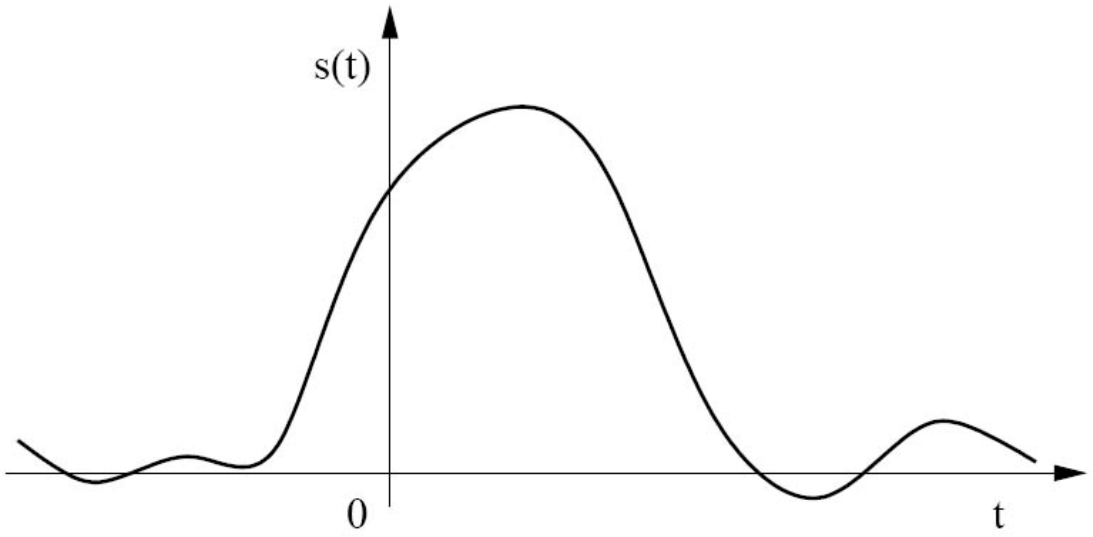
2.2.1 Robustní parametrizace

V praxi je důležité, aby rozpoznávač pracoval spolehlivě nejen se signály, které nejsou nijak rušené, ale naopak. Nejčastěji se setkáváme se signály, které jsou rušené nežádoucím hlukem nebo ztrátami dat při vedení signálu. Rozpoznávač tedy musí být robustní vůči těmto rušivým vlivům různé intenzity. Robustnost parametrizace tedy závisí na schopnosti odolat těmto poruchám.

2.2.2 Zpracování signálu

Původní signál

Za původní signál můžeme považovat samotnou řeč. Jedná se tedy o analogový signál. Není časově omezen.



Příklad části analogového signálu

Vzorkování signálu

Abychom mohli signál uložit a dále zpracovávat, je převeden do digitální podoby. Z analogového signálu získáme digitální signál jeho vzorkováním. Pro nevzorkování signálu vynásobíme původní analogový signál periodickým sledem Diracových impulsů. Po vynásobení dostaneme opet periodickou posloupnost Diracových impulsů, nyní již však s mocnostmi odpovídajícími původnímu signálu. Násobíme Diracovými impulsy s periodou T, tedy výsledný signál bude mít odpovídající hodnoty v bodech nT.

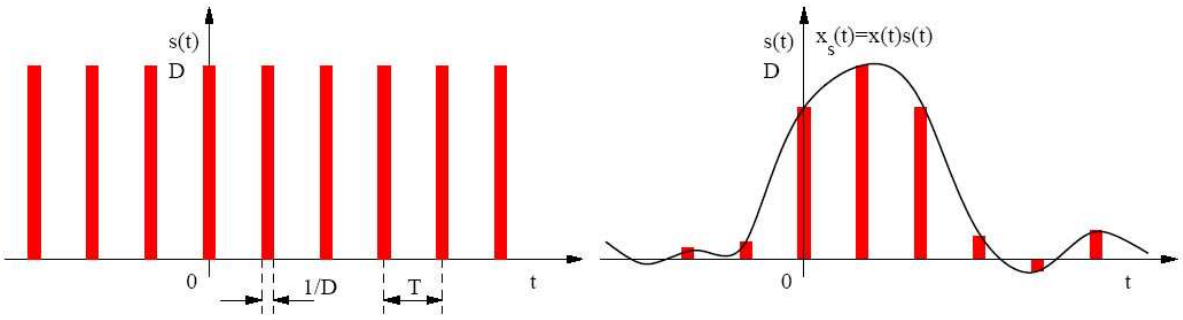
T je vzorkovací perioda

$F_s = \frac{1}{T}$ je vzorkovací frekvence

Vzorkovací frekvenci volíme větší než dvojnásobek maximální frekvence obsažené v původním spektru analogového signálu.

Fs > 2*fmax (Shanonův vzorkovací teorém)

Při nedodržení této podmínky dochází k tzv. aliasingu. To znamená, že spektrum tohoto signálu má jiný tvar než původní spektrum a signál již nemůžeme znovu rekonstruovat. Nevzorkovaný signál je posloupnost diskrétních hodnot. Musíme však zachovat informaci o použité vzorkovací frekvenci. Ve frekvenční oblasti se signál nevzorkováním periodizuje. Ve spektru tedy máme k dispozici N hodnot od 0 do Fs a horní polovina tohoto spektra je symetrická s dolní polovinou. Tudíž nás dále zajímá pouze polovina tohoto spektra.



Příklad vzorkovaného signálu

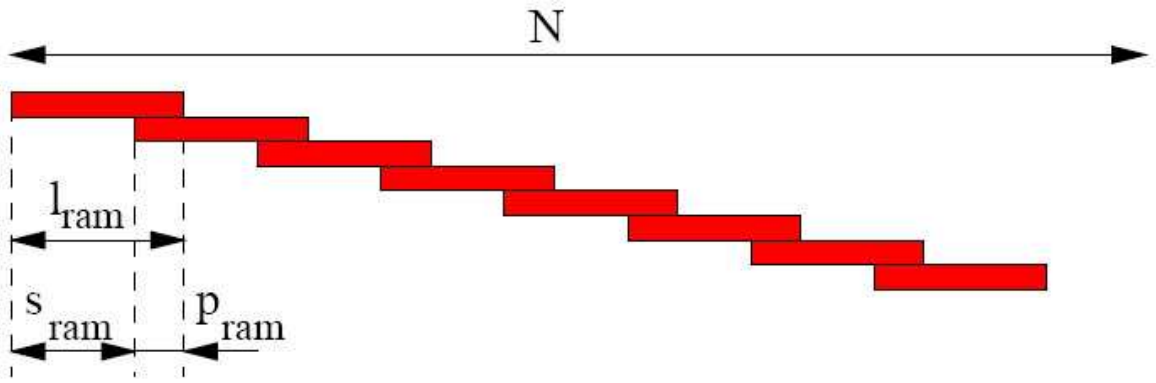
Rozdělení na rámce

Protože původní signál je pro zpracování příliš dlouhý a nestacionární, dělíme ho na rámce.

Délka rámce by měla být dostatečně malá, abychom mohli signál považovat za stacionární a zároveň dostatečně velká, aby mohli být požadované parametry odhadnuty s dostatečnou přesností. Setrvačnost hlasového ústrojí je přibližně 20 - 25ms což při vzorkovací frekvenci 8kHz odpovídá 160 - 200 vzorkům.

Důležité je také zvolení **překrytí** rámců. Pokud zvolíme malé překrytí rámců, hodnoty parametrů se mohou mezi jednotlivými rámci značně lišit. Při použití velkého překrytí rámců jsou si parametry více podobné, ovšem za cenu většího nároku na hardware. Je tedy opět nutné volit kompromis. V našem případě je překrytí 80 rámců za sekundu což odpovídá délce 10ms.

Pro jednotlivé segmenty řeči budeme chtít sčítat jejich spektra. Diskontinuity na krajích těchto segmentů vedou k zašumění spekter. Pro odstranění tohoto jevu sáhujeme segment hammingovým oknem.



Rozložení rámců v signálu

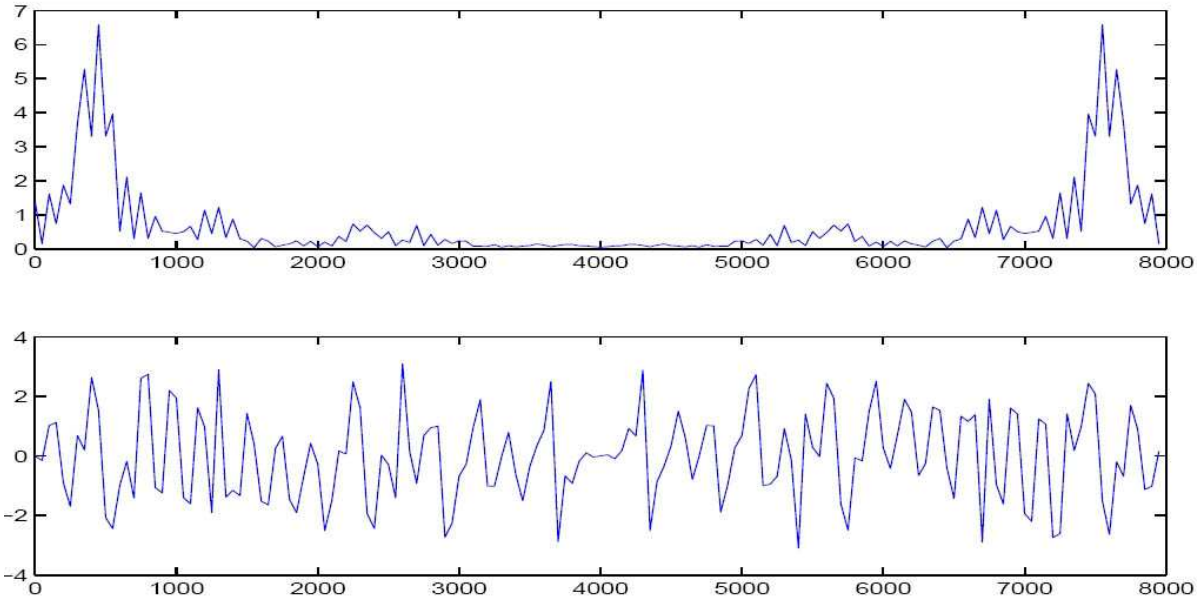
l_{ram} – délka rámce

s_{ram} – posun

p_{ram} – překrytí rámců

Spektrogram

Spektrum (spektrální hustota výkonu) odhadneme pro každý řečový rámec. Všechna spektra zobrazíme jako matici. Vodorovná osa odpovídá číslu rámce, na svislou osu vynášíme frekvence. Pro jednotlivé segmenty řeči budeme chtít sčítat jejich spektra. Diskontinuity na krajích těchto segmentů vedou k zašumění spekter. Pro odstranění tohoto jevu sáhujeme segment hammingovým oknem. Spektrogram je velice vhodným prístupem pro analýzu recového signálu. Je to popis, který nám dává velice dobré vizuální informaci o recovém signálu i v prípade, kdyz je signál prekryt znacným hlukem. Lze presne videt, v kterém okamziku a pri které frekvenci se objeví hluk v záznamu.



Příklad spektra jednoho rámce

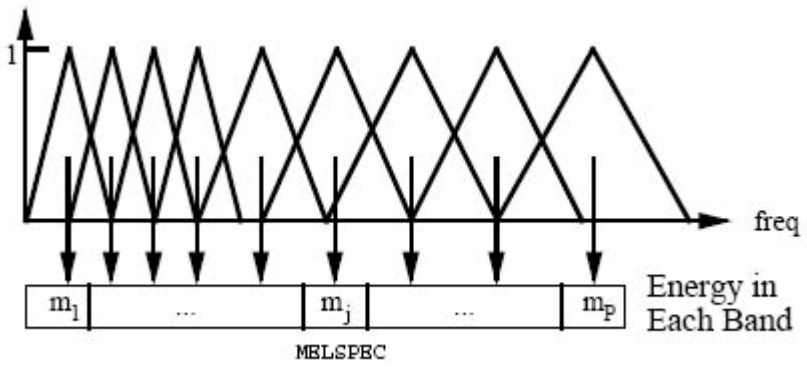
Cepstrum

Pro kódování je vhodné oddělit buzení a modifikaci. Buzení (základní tón) je velmi závislé na konkrétním řečníkovi(na pohlaví, věku, náladě,…). Pro rozpoznávání jej tedy nepotřebujeme vůbec.

MFCC – Mel-frequency cepstrum

Lidské ucho má na rozdíl od toho „strojového“ jinou rozlišovací schopnost na nízkých a vysokých frekvencích. Nízké frekvence rozlišujeme lépe než vysoké frekvence. Pro rozpoznávání řeči se snažíme cepstrum přiblížit lidskému slyšení. Na frekvenční osu nelineárně rozmístíme filtry a změříme energii na jejich výstupu. Prakticky provedeme výpočet MFCC tak, že provedeme Fourierovu transformaci, signál umocníme, vynásobíme trojúhelníkovým oknem a sečteme. Tento způsob je také použit v HTK toolkit použitého v tomto projektu. Tímto odstraníme jemnou strukturu spektra nesoucí informaci o nedůležitém základním tónu.

Výstup z banky filtrů je logaritmován, což odpovídá logaritmickému vnímání hlasitosti lidmi a vede k tomu, ze rozložení koeficientů se více blíží gaussovskému rozložení.´



Banka filtrů

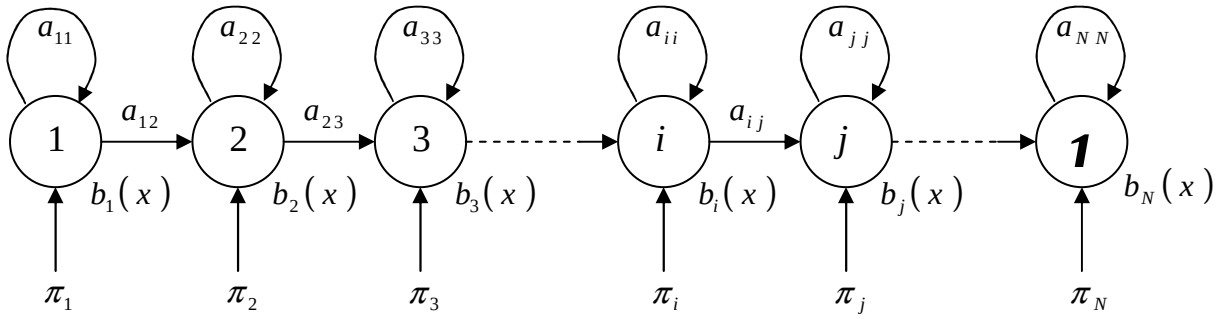
Dekorelace

Logaritmický výstup je promítnut do kosinových bází. Tím se tyto koeficienty dekorelují.

3 Rozpoznávání řeči

3.1 HMM – Hidden Markov Model

Rozpoznávač řeči použitý v tomto projektu je založen na HMM neuronové síti. Skryté Markovovy modely (HMM - Hidden Markov Model) odpovídají statistickým modelům s konečným počtem stavů. Tyto modely mohou být užitečné pro statistický popis po částech stacionárních signálů jako je řeč. V teorii HMM je po částech stacionární signál modelován pomocí řetězce skládajícího se z N stavů, přičemž každý z nich má jiný soubor statistických charakteristik. N stavový Markovův model je uveden na obr.



N-stavový skrytý Markovův model je definován následujícím souborem parametrů $\lambda = \{\pi_i, a_{ij}, b_i(x), i, j = 1, ..., N\}$.

Význam jednotlivých parametrů:
π_i - počáteční pravděpodobnost stavu
a_{ij} - pravděpodobnost přechodu mezi stavy *i* a *j*
$b_i(x)$ - hustota pravděpodobnosti pro pozorovaný vektor *x* v daném stavu
N - počet stavů
i, j - označení daných konkrétních stavů

Hlavními parametry, které charakterizují *HMM* model jsou pravděpodobnosti přechodů mezi stavy a_{ij} a hustoty pravděpodobnosti pro pozorované vektory *x* v daných stavech $b_i(x)$. Pravděpodobnosti přechodů mezi jednotlivými stavy vyjadřují různou délku trvání řečových segmentů, jejichž statistické parametry lze přiřadit jednotlivým stavům. Hustoty pravděpodobnosti pro pozorované vektory v daných stavech oproti tomu vyjadřují změny ve spektrálním obsahu řečových segmentů přiřazených k danému stavu. Vzhledem k tomu, že přechod mezi jednotlivými stavy se může konat pouze z levé části modelu do pravé, se pro zpracování řečových signálů používá tzv. levopravý *HMM*.

Pravděpodobnost, že sekvence pozorovaných vektorů $x=[x_1, x_2, \mathrm{K}, x_T]$ odpovídá akustické realizaci slova popsaného modelem λ, je obecně dána sečtením pravděpodobností sekvence pozorovaných vektorů přes všechny možné sekvence stavů q. Toto lze vyjádřit následujícím vzorcem:

$$p_\lambda(x)=\sum_{\forall q} p_\lambda(x|q)\,p_\lambda(q)=\sum_{\forall q_1,\mathrm{K},q_T} \pi_{q1}\,b_{q1}(x_1)\,a_{q1q2}\,b_{q2}(x_2)\Lambda\; a_{qT-1qT}\,b_{qT}(x_T) \qquad (1)$$

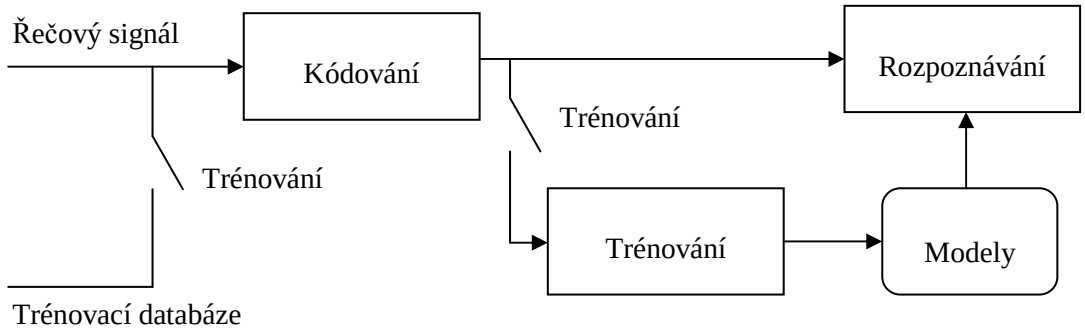
Pro vyčíslení předchozí pravděpodobnosti je nutné umět vypočítat hustotu pravděpodobnosti pro pozorovaný vektor *x* v daném stavu $b_i(x)$. Tato hustota je zpravidla modelována pomocí *M*-složkové směsi hustot pravděpodobnosti podle Gaussova rozložení. To vyjadřuje vzorec 2.

$$b_i(x_k)=\sum_{m=1}^{M} P_{im}\,\mathrm{N}(x_k,\mu_{im},\Sigma_{im}) \qquad (2)$$

kde P_{im} jsou váhové koeficienty stavu i a $\mathrm{N}(x_k,\mu_{im},\Sigma_{im})$ jsou jednotlivé hustoty pravděpodobnosti pro daný pozorovaný vektor x_k, μ_{im} je vektor středních hodnot a Σ_{im} označuje kovarianční matici.

3.1.1 Jednotlivé části HMM rozpoznávače řeči

Zjednodušené schéma rozpoznávače je uvedeno na obr. První operací, kterou řečový signál prochází je kódování. Při něm se zpravidla signál segmentuje na stejně dlouhé segmenty z kterých se vypočítá spektrum a to se dále zpracovává. Použité kódování bylo popsáno výše.



Blokové schéma rozpoznávače

Před tím, než je rozpoznávač schopen rozpoznávat jednotlivé promluvy (např. slova) je třeba nejdříve natrénovat jeho modely jednotlivých promluv. K tomu slouží tzv. trénovací část databáze, kde se například vyskytuje mnoho různých realizací od každého slova, které chceme mít ve slovníku rozpoznávače. Tato etapa přípravy rozpoznávače se nazývá trénování.

Vlastní rozpoznávání se děje vyhodnocováním pravděpodobnosti, že daná promluva (sekvence pozorovaných vektorů *x*) byla generována daným modelem. Model, který generuje danou promluvu s nejvyšší pravděpodobností reprezentuje danou promluvu.

Šum ovlivňuje proces kódování, trénování modelů i vlastní rozpoznávání. Rozpoznávače řeči dosahují nejlepších výsledků pokud jsou trénovány a testovány na promluvách s přítomností stejného druhu šumu (co do množství i spektra).

Šum ovlivňuje jak pravděpodobnost přechodů mezi jednotlivými stavy, tak i hustoty pravděpodobnosti pozorovaného vektoru v daných stavech. Vliv šumu na hustoty pravděpodobnosti pozorovaného vektoru v daných stavech se zpravidla považuje za více významný než jeho vliv na pravděpodobnosti přechodů.

K definici, trénování a k rozpoznávání pomocí HMM (hidden Markov model) slouží HTK. HTK(Hidden Markov model toolkit) slouží také pro vyhodnocování výsledků rozpoznávání. HTK je napsáno v jazyce C a pro nekomerční použití je volně stažitelný.

Použité funkce HTK toolkit **:**

HList – zobrazuje soubor s parametry

HCopy – kopíruje, přičemž dokáže provést i konverzi dat. Jelikož konverzi dat jsme upravovali, není v projektu použit

HCompV – inicializuje parametry vysílacích hustot rozdělení pravděpodobností ve stavech modelu na globální hodnoty pro dané slovo

HRest – přetrénování modelu. Provádí výpočet funkce přiřazení jednotlivých vektorů ke stavům s následným přetrénováním modelů

HVite – Viterbiho dekodér. Rozpoznávač – pro natrénované modely a neznámá slova provede výpočet Viterbiho pravděpodobností a výběr maxima. Model, který vyslal slovo s největší pravděpodobností je rozpoznán

HResults – na základě správných předpisů rozpoznaných slov vyhodnocuje chybovost

3.1.2 Testovací databáze

Testování probíhalo na databázi TIDIGITS. Databáze obsahuje různé posloupnosti číslic s různými typy a intenzitami rušení.

Typy rušení – různé hluky na pozadí řeči. Obsahuje hluky odpovídající hluku na ulici, v metru, v autě, v restauraci, atd…

Intenzity rušení – od čistého řečového signálu až po signál, kde intenzita rušení dosahovala hlasitosti řeči. Databáze TIDIGITS obsahuje 6 stupňů intenzity rušení :

- **clean** – čistý signál bez rušení

- **NSR20** – odstup řečového signálu od šumu 20dB – málo zašuměný signál
- **NSR15** - odstup řečového signálu od šumu 15dB
- **NSR10** - odstup řečového signálu od šumu 10dB
- **NSR5** - odstup řečového signálu od šumu 5dB
- **NSR0** - odstup řečového signálu od šumu 0dB – velmi zašuměný signál

3.1.3 Možné způsoby zlepšení rozpoznávacích schopností

- Adaptace (Přizpůsobení) modelů jednotlivých promluv tak, aby zahrnuly vliv šumu

  Jak již bylo řečeno, rozpoznávače řeči dosahují nejlepších výsledků pokud jsou trénovány a testovány na promluvách s přítomností stejného druhu šumu.

- předzpracování řečového signálu před vlastním rozpoznáváním.

  Použití různých algoritmů pro úpravu řečového signálu. Toto se může dít před vlastním kódováním, případně může být daný algoritmus zařazen do části kódování. Použité metody jsou popsány v následující kapitole.

- nastavení neuronové sítě umožňující lepší (přesnější) natrénování této sítě.

3.1.4 Algoritmy pro úpravu signálu

Všechny zmíněné úpravy jsou součástí kódovacího procesu

3.1.4.1 Nelineární vyhlazování signálu ve frekvenční oblasti

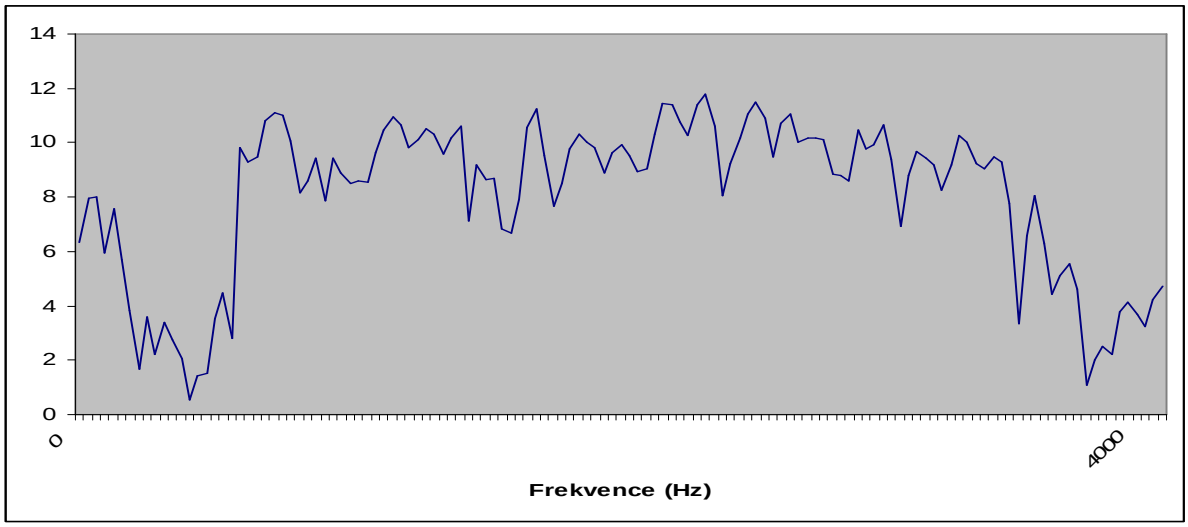
Jedná se o jednoduchý, podle autorů však účinný, algoritmus pro zvýšení přesnosti rozpoznávače. Spektrum každého rámce je upraveno následujícím způsobem

F(i) = max (f(i), f(i+1)*c1, f(i-1)*c2)

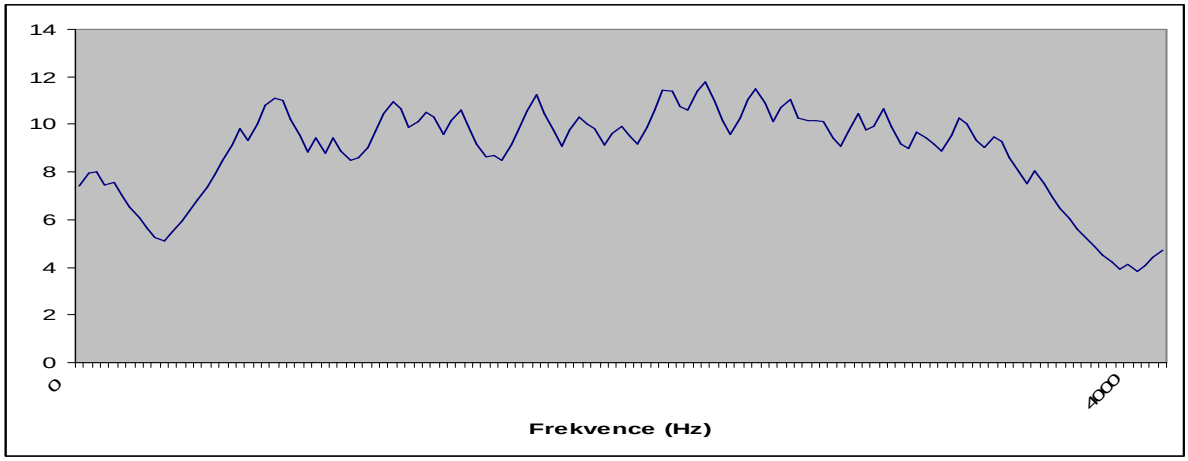
Je tedy zajištěno, že dvě sousední hodnoty spektra daného rámce se liší maximálně o (1-c1) resp. (1-c2) násobek jedné z těchto hodnot. Tento algoritmus tedy částečně vyhladí spektrum každého řečového rámce. Stupeň tohoto vyhlazení lze korigovat vhodně zvolenými koeficienty c1 a c2.

Příklad.

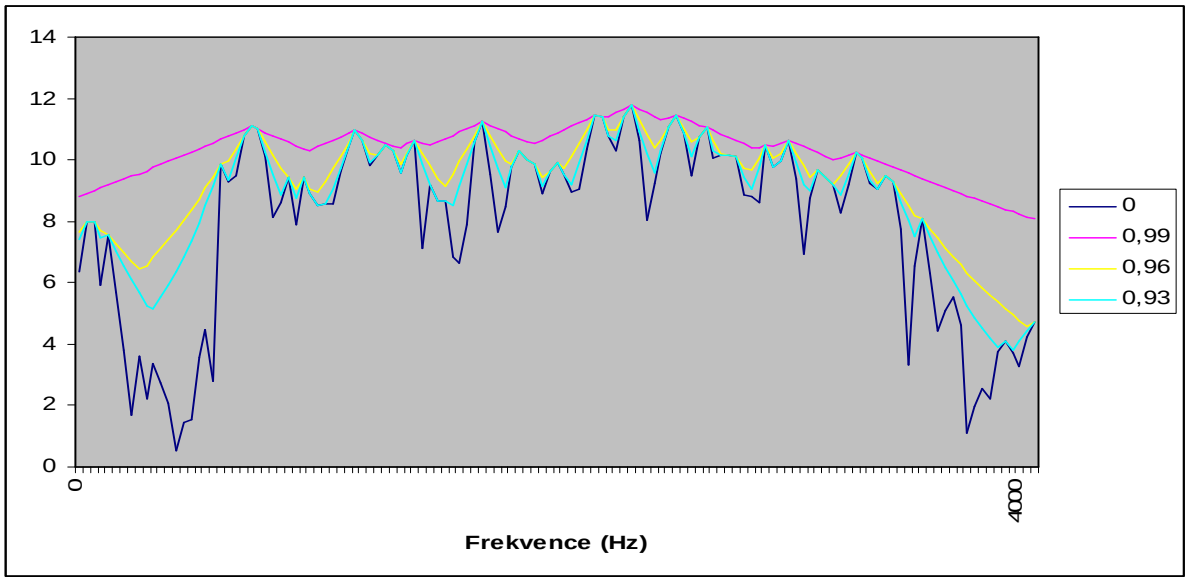
Původní spektrum jednoho rámce řečového signálu



Upravené (vyhlazené) spektrum jednoho rámce řečového signálu



Ukázka vlivu zvolení koeficientů na stupeň „vyhlazenosti“ signálu



Jak je vidět z průběhu signálu, je výsledek závislý na výrazných frekvencích spektra a méně výrazné frekvence jsou překryty. Jak jsem se již zmínil dříve, rozpoznávač dosahuje lepší efektivity, pokud trénování i testování probíhá s daty zašuměnými stejným šumem. Ze stejného důvodu provádíme nelineární vyhlazování na trénovacích, tak i testovacích datech.

Dosažené výsledky:*

SNR – signal to noise ratio – odstup řečového signálu od šumu (0,5,10,15,20 dB).

Clean – čistý signál bez přidání aditivního šumu.

Efektivita – udává kolik procent rozpoznaných slov bylo rozpoznáno správně.

Referenční hodnota - efektivita samotného rozpoznávače, bez použití algoritmu pro vyhlazování.

SNR (dB)	Clean	20	15	10	5	0
Efektivita (%)	99,025	98,265	97,595	95,655	88,905	61,035

Hodnoty dosažené použitím nelineárního vyhlazování

Koef	Clean	20	15	10	5	0
0,85	98,89	98,37	97,70	95,93	88,98	62,11
0,90	98,95	98,48	97,60	96,06	89,45	64,14
0,91	99,03	98,51	97,71	96,03	89,67	64,85
0,92	99,04	98,53	97,80	96,31	89,97	65,33
0,93	99,06	98,51	97,81	96,13	90,05	65,50
0,94	99,01	98,32	97,74	96,11	90,37	66,08
0,95	98,86	98,30	97,66	95,92	90,28	66,00
0,96	99,03	98,42	97,70	95,95	89,87	66,34
0,97	98,95	98,18	97,38	95,71	89,21	65,35
0,98	98,63	97,77	96,70	94,75	87,56	64,40
0,99	97,22	95,86	94,49	91,60	82,74	58,02

Relativní zlepšení odpovídající naměřeným hodnotám**

Koeficient	SNR Clean	20	15	10	5	0	Průměrná hodnota
0,85	-14	6	4	6	1	3	1,1
0,90	-8	12	0	9	5	8	4,5
0,91	0	14	5	9	7	10	7,3
0,92	2	15	8	15	10	11	10,1
0,93	3	14	9	11	10	11	9,8
0,94	-2	3	6	10	13	13	7,3
0,95	-17	2	2	6	12	13	3,1
0,96	0	9	4	7	9	14	7,1
0,97	-8	-5	-9	1	3	11	-1,2
0,98	-41	-29	-37	-21	-12	9	-21,8
0,99	-185	-139	-129	-93	-56	-8	-101,6

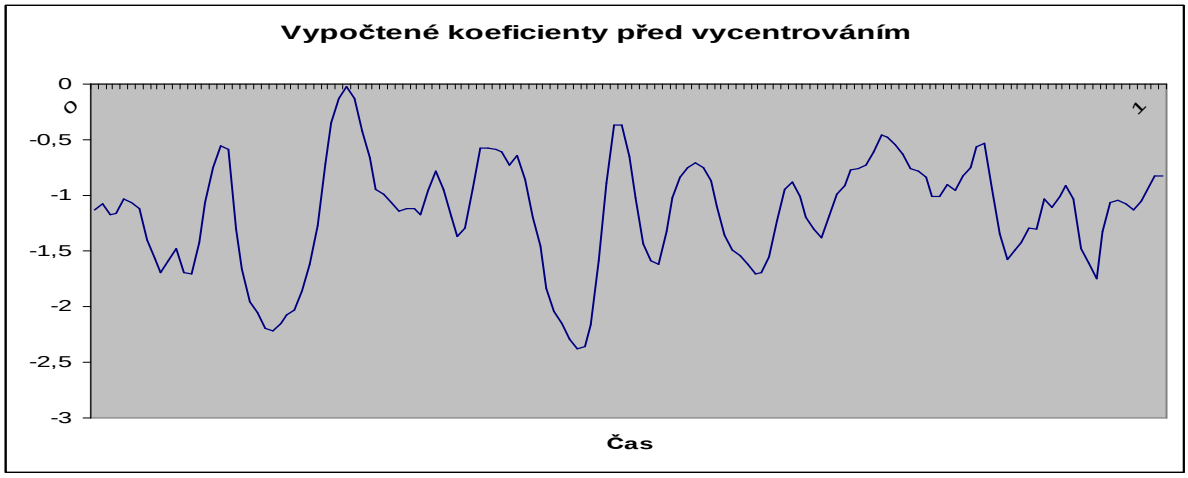
Autoři článku (Non-linear spectral smoothing) doporučují nastavení koeficientů v rozmezí hodnot 0,95 a 0,99. Ze získaných dat vyplývá, že můžeme doporučit hodnoty v rozmezí 0,91 až 0,93, pokud budeme uvažovat o zlepšení pro všechny druhy šumu. Za předpokladu, ze rozpoznávám bude jen zašuměný signál popřípadě při znalosti míry zašumění můžeme volit hodnoty 0,90 – 0,96. Zprůměrováním relativních zlepšení se jeví jako nejvýhodnější hodnota nastavení koeficientu na hodnotu 0,92. Časté zhoršení výsledků pro čistý signál je pochopitelné, neboť do něj vkládáme i frekvence, které v signálu původně vůbec nebyly, nebo byly zastoupeny jen velmi nevýrazně.

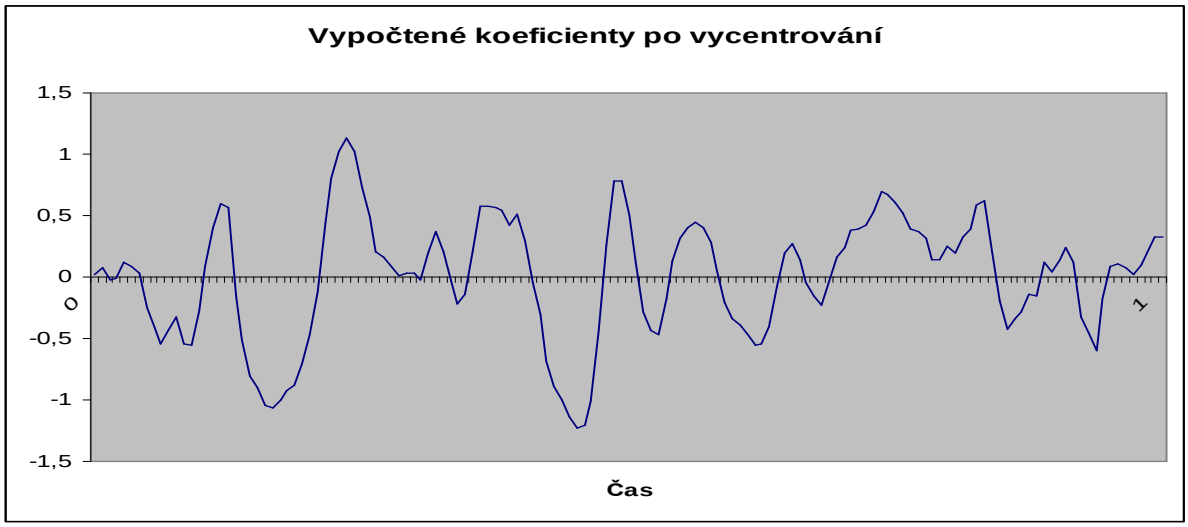
**Pro přehlednost získaných výsledků jsou výsledné hodnoty průměrovány přes všechny typy použitých šumů.*

***Udává, o kolik procent se zmenšila/zvětšila část špatně rozpoznaných slov.*

3.1.4.2 Centrování získaných parametrů v časové ose

Jak jsem se zmínil v kap.Zpracování signálu, Každý řečový signál je nejprve digitalizován. Výsledná podoba digitálního signálu je vždy závislá na použité aparatuře. Dvě různé nahrávací aparatury mohou být také různě citlivé. I při zachování charakteru signálu (stoupání, klesání, maxima či minima, …) mohou mít signály různou intenzitu danou právě citlivostí našeho systému. Pokud přejdeme od lineárních hodnot k logaritmickým, odpovídá tato různá intenzita (například signál jehož hodnoty jsou 2* větší) pouhému posunutí hodnot o předem neznámou konstantu. Opět se zde snažíme o to, aby si testovací a trénovaní signály byly co nejvíce podobné. Všechny tyto signály tedy v logaritmickém měřítku posuneme tak, aby měly stejnou střední hodnotu. Pracujeme s již známými spektry signálu. Centrování tedy provádíme pro každou frekvenci přes celý signál. Jelikož při kódování provádíme pouze lineární operace, můžeme centrování provádět na závěr celého kódování, čímž tento proces urychlíme, neboť nebudeme centrovat signál přes všechny frekvence, ale jen přes získané koeficienty.





Dosažené výsledky:

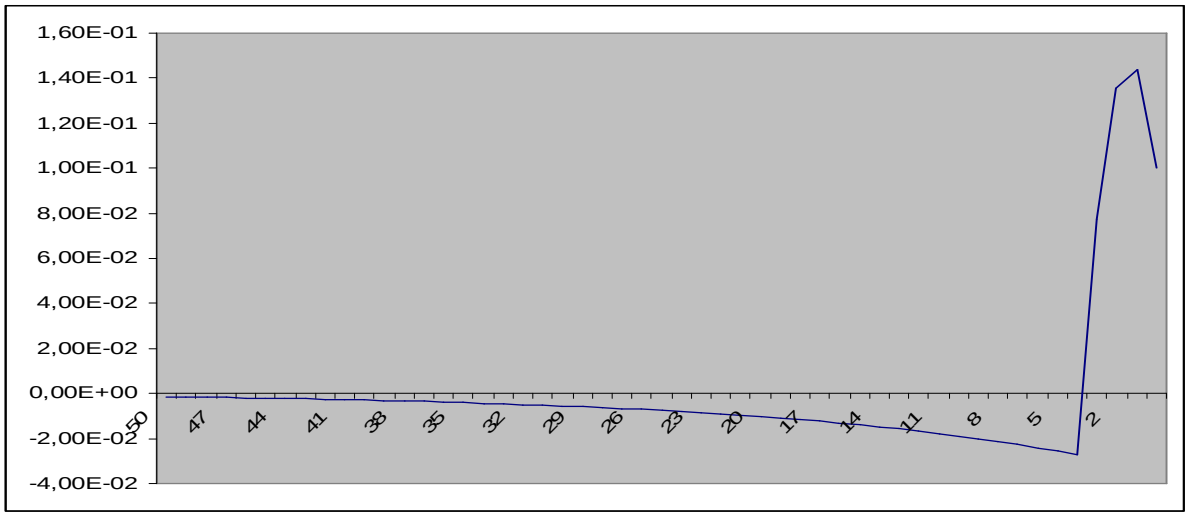
	SNR (dB)					
Efektivita (%)	Clean	20	15	10	5	0
Bez použití centrování	99,02	98,26	97,59	95,65	88,90	61,03
S použitím centrování	99,03	98,32	97,68	95,86	89,18	61,78
Relativní zlepšení	1,0204	3,4483	3,7344	4,8276	2,5225	1,9246

Centrováním signálu došlo k částečnému zlepšení o cca 3%

3.1.4.3 Použití RASTA filtru

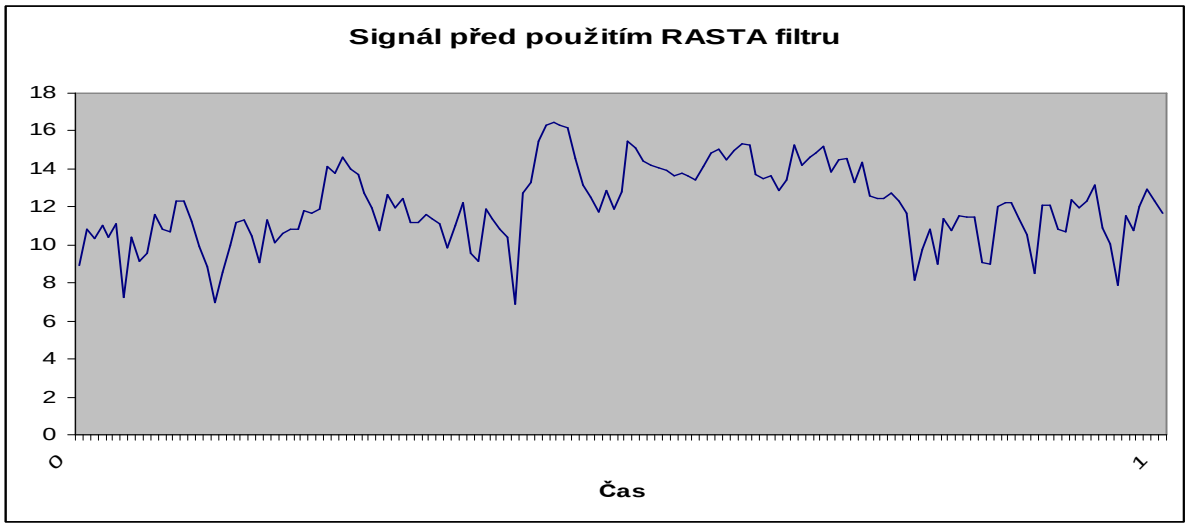
Během provádění záznamu řeči může být signál rušen vnějším šumem odpovídajícím rychlým změnám spektra signálu. I přes jistou charakteristiku šumového signálu se může na pozadí vyskytnout nežádoucí ruch, který svojí intenzitou tento šum převyšuje. Pokud je však takovýto ruch dostatečně krátký, můžeme na základě znalosti lidské řeči (lidská řeč má jistou setrvačnost, není možné měnit charakteristiky řeči okamžitě) označit tuto část signálu za nežádoucí. Dalším rušivým jevem mohou být dlouhodobé změny signálu. Tuto změnu může způsobit například změna polohy řečníka vzhledem k nahrávacími zařízení.

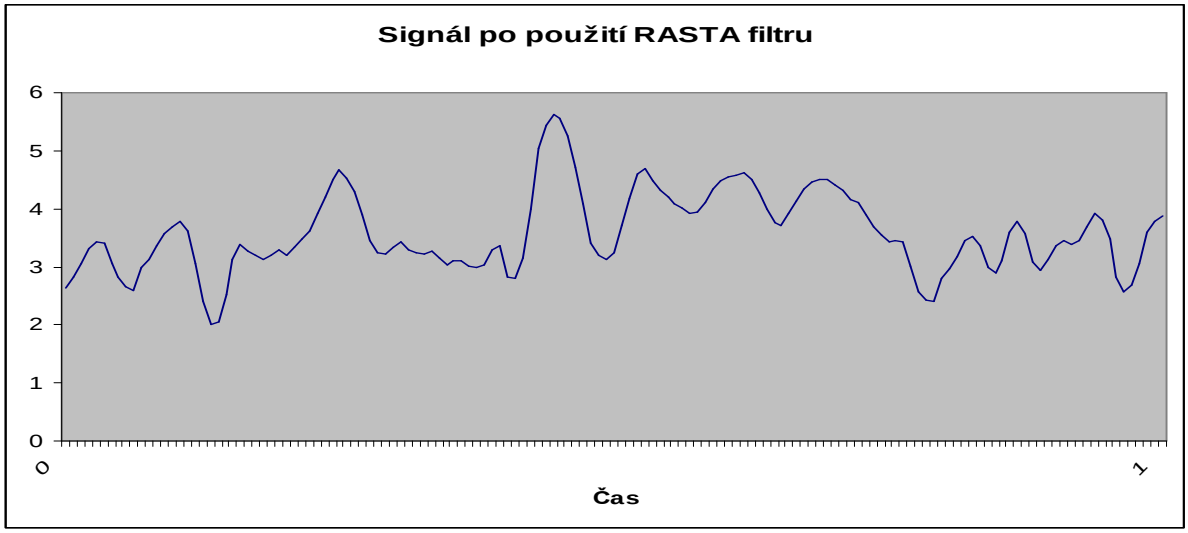
Oba tyto jevy, krátkodobé i dlouhodobé, můžeme částečně eliminovat použitím filtru RASTA. Jedná se vlastně o jakousi pásmovou propust. RASTA filtr má následující impulsní odezvu



Nová hodnota po projití signálu tímto filtrem tedy není dána pouze okamžitou hodnotou signálu, ale je závislá také na předchozím

průběhu signálu. Na následujících obrázcích vidíme jaká vliv má tento filtr na řečový signál.





Použitím filtru tedy dojde k částečnému rozmazání a narovnání signálu v časové ose.

Výsledky dosažené použitím RASTA filtrování:

	SNR (dB)					
Efektivita (%)	Clean	20	15	10	5	0
Bez použití RASTA filtru	99,03	98,27	97,60	95,66	88,91	61,04
S použitím filtru	98,68	98,25	97,69	96,12	90,33	68,23
Relativní zlepšení	-35,38	-0,86	3,95	10,70	12,84	18,47

Použití RASTA filtru tedy vedlo ke zlepšení výsledků u zašuměných signálů.

3.1.4.4 Kombinace předešlých algoritmů

Algoritmy pro úpravu signálu, respektive parametrů používaných pro rozpoznávání můžeme také s úspěchem kombinovat. Kombinace těchto algoritmů vede k dalšímu zlepšení efektivnosti rozpoznávače.

Výsledky dosažené kombinací všech předešlých postupů

	SNR (dB)					
	Clean	20	15	10	5	0
Relativní zlepšení (%)	-3	4	-2	11	20	29
Efektivita (%)	99,00	98,33	97,54	96,15	91,10	72,25

Pro porovnání výsledků bylo použito RASTA filtrování, nelineární vyhlazení s koeficientem vyhlazeni 0,93 a následně centrování hodnot vypočtených parametrů.

Jak jsme již řekli, použití RASTA filtrování je vhodné zejména pro více zašuměné signály. Z tohoto důvodu je hlavní oblast zlepšení rozpoznávání opět mezi silně zašuměnými signály.

3.1.5 Nastavené neuronové sítě

Neuronová síť používaná pro rozpoznávání řeči má jistý počet skrytých (vnitřních) stavů – v našem případě 3. Jak jsem se zmínil výše, každý z těchto stavů je charakterizován pomocí pravděpodobnosti setrvání v tomto stavu. Každý tento stav je tedy charakterizován gaussovým rozložením pravděpodobnosti. Jelikož používáme 13 parametrů, které jsme získali při kódování, je toto rozložení pravděpodobnosti dáno 13-ti rozměrnou gaussovou plochou. Pro lepší pokrytí trénovaní množiny můžeme použít více těchto gaussových ploch. Tím tedy zvětšíme počet vnitřních stavů naší neuronové sítě. Nejedná se zde tedy o změnu ve vstupním signálu, ale přímo o změnu neuronové sítě.

Výsledky dosažené zvýšením počtu neuronů:

	SNR (dB)					
Efektivita (%)	Clean	20	15	10	5	0
3 gaussovy plochy	99,03	98,42	97,70	95,95	89,87	66,34
4 gaussovy plochy	99,15	99,03	98,62	97,06	92,15	68,71
5 gaussových ploch	99,36	99,15	98,89	97,54	92,7	70,86

Což odpovídá relativnímu zlepšení:

	SNR (dB)					
Relativní zlepšení (%)	Clean	20	15	10	5	0
4 gaussovy plochy	12,821	38,608	40	27,407	22,546	7,0548
5 gaussových ploch	34,359	46,203	51,739	39,259	27,972	13,441

Zvětšení počtu stavů neuronové sítě tedy vede ke zlepšení výsledků, zároveň však stoupá časová náročnost trénovacího procesu. Jelikož se nejedná o algoritmus pro úpravu signálu, výsledky dosažené pomocí zvětšení počtu neuronů jsou uvedeny pouze pro porovnání.

Po použití nelineárního vyhlazování, RASTA filtru a centrování signálu na signál, který byl vstupem takto upravené neuronové sítě jsme dosáhli následujících výsledků:

Pro 4 vnitřní stavy neuronové sítě

	SNR (dB)					
Efektivita rozponávače (%)	Clean	20	15	10	5	0
Původní hodnoty	99,15	99,03	98,62	97,06	92,15	68,71
Po použití zkoumaných algoritmů	99,31	99,10	98,80	97,53	93,26	73,50
Relativní zlepšení (%)	19	7	13	16	14	15

Pro 5 vnitřních stavů neuronové sítě

	SNR (dB)					
Efektivita rozponávače (%)	Clean	20	15	10	5	0
Původní hodnoty	99,36	99,15	98,89	97,54	92,70	70,86
Po použití zkoumaných algoritmů	99,35	99,30	98,92	97,60	93,23	74,68
Relativní zlepšení (%)	-2	18	3	2	7	13

Jak je tedy z těchto hodnot patrné, nelineární vyhlazování spektra signálu (v kombinaci s filtrováním a centrováním koeficientů) je vhodné i při použití neuronové sítě s větším počtem vnitřních stavů. I při zvýšení počtu stavů sítě dosahujeme tímto algoritmem zlepšení o cca 10%.

4 Závěr

Projekt byl zaměřen zejména na seznámení s parametrizací řečového signálu a na použití nelineárního spektrálního vyhlazování pro zlepšení výsledků rozpoznávání a porovnání tohoto algoritmu s ostatnímu metodami používanými pro zlepšení výsledků rozpoznávání řeči. Jelikož procentuální úspěšnost rozpoznávání daných slov z databáze TIDIGITS se použitím nelineárního vyhlazování zlepšila o několik procent, ukázal se tento algoritmus, vzhledem k jeho nepříliš složité implementaci, jako účinná pomoc při parametrizaci řeči. Kombinací s ostatními způsoby úpravy signálu před samotným rozpoznáváním jsme dosáhli zajímavého zlepšení rozpoznávací schopnosti zejména pro značně zašuměné signály. Zejména pro SNR 0 a 5 dB jsme dosáhli téměř stejného zefektivnění rozpoznávání jako při použití složitější neuronové sítě.

Jako další zlepšení rozpoznávací schopnosti se jeví kombinace mezi rozšířením neuronové sítě a použitím výše uvedených algoritmů pro úpravu signálu. Tuto kombinaci je však třeba volit uvážlivě v závislosti na požadované časové náročnosti trénování systému.

Zpracování řeči je velmi zajímavou oblastí zpracování řečového signálu. Díky HMM je úzce spojena s neuronovými sítěmi. V diplomovém projektu bych se proto rád dále zabýval touto oblastí. Zajímavá se mi jeví také implementace hlasového ovládání do nových, nebo i stávajících programů, či automatický přepis mluveného slova do textové podoby spolu s identifikací řečníka podle hlasu.

Literatura

[1] Přednášky kurzu Číslicové zpracování řeči

[2] HTK book

[3] Brian C. J. Moore: An introduction to the psychology of hearing

[4] Přednášky kurzu Neuronové sítě

[5] B.P. Milner, S.V. Vaseghi: Comparison of some noise-compensation methods for speech recognition in adverse environments

[6] Young, S. J., HTK: ‘Reference and User Manual’

Příloha A

Parametrizace zvukového souboru

Pro parametrizaci zvukového souboru je použit program fbout.
Tento prgram provede parametrizaci souboru s využitím centrování parametrů a RASTA filtru. Pokud zadáme parametry pro nelineární vyhlazeni signálu, bude použito i toto vyhlazování.

Použití
fbout input_file output_file [constA constB]

Pro úpravu většího počtu vstupních souborů je použit script param.sh, který spustí fbout na všechny soubory ve vstupní složce a uloží výstup do výstupní složky.

Použití
Param.sh coefA coefB
Pokud chceme použít tento script bez využití nelineárního spektrálního vyhlazení, použijeme volání param.sh 0.0 0.0

Trénování a testování systému

Pro trénování a testování rozpoznávače je použit script train_recog_mutli, který provede trénování a následné testování rozpoznávače včetně statistického zpracování výsledků.

Všechny zdrojové texty potřebné pro parametrizaci signálů, spuštění a testování rozpoznávače jsou přiloženy na CD.

Příloha B

Tato příloha obsahuje článek „Robust Speech Recognition Using Non-Linear Spectral Smoothing“ na jehož základě byl zpracován a zkoumán algoritmus nelineárního spektrálního vyhlazování a jeho použití pro úlohu robustního rozpoznávání řeči.

Robust Speech Recognition Using Non-Linear Spectral Smoothing

Michael J. Carey

Department Electrical and Electronic Engineering
University of Bristol U.K.
michael.carey@iee.org

Abstract

A new simple but robust method of front-end analysis, non-linear spectral smoothing (NLSS), is proposed. NLSS uses rank-order filtering to replace noisy low-level speech spectrum coefficients with values computed from adjacent spectral peaks. The resulting transformation bears significant similarities with masking in the auditory system. It can be used as an intermediate processing stage between the FFT and the filter-bank analyzer. It also produces features which can be cosine transformed and used by a pattern matcher. NLSS gives significant improvements in the performance of speech recognition systems in the presence of stationary noise, a reduction in error rate of typically 50% or an increased tolerance to noise of 3dB for the same error rate in an isolated digit test on the Noisex database. Results on female speech were superior to those on male speech: female speech gave a recognition error rate of 1.1% at a 0dB signal to noise ratio.

1. Introduction

The filter-bank analyzer (FBA) has become the preferred method of feature extraction in speech recognition systems. It is often implemented as two stages of processing with a bank of frequency domain filters following the computation of the power spectrum using the fast Fourier transform (FFT) [1]. The reason usually given for the use of the FFT is that it is computationally more efficient than a direct realization of the filter-bank using a set of digital filters. Given the required spectral information is contained in the log-power FFT it is profitable to consider the reasons why it is necessary to include the frequency domain filter-bank processing stage. The logpower spectrum produced by the FFT, shown for example in Figure 1, is deficient as a feature set in the following respects,

- It is not pitch synchronous and contains the pitch structure of the speech,
- The logarithmic function gives equal weight to differences in the high and low power parts of the spectrum,
- The Fourier coefficients are linearly spaced in the frequency domain giving undue emphasis to the higher spectral features of the speech signal.

The human auditory [2] system does not share these deficiencies. In the auditory system the firings of the auditing nerves are pitch synchronous. Auditory masking suppresses the contributions of the low power parts of the spectrum and the frequency response of the basilar membrane is logarithmically spaced. The filter-bank analysis stage of the processing is used to try to correct the FFT deficiencies by bringing the feature extraction process into line with that of the auditory system.

The frequency domain filters used in the FFT filter-bank make a weighted summation of adjacent spectral coefficients. In so doing the pitch variation in the FFT coefficients is attenuated. By adding the coefficients in the linear power domain the effect of changes in the low-power coefficients on the magnitude of the features is reduced. The centre frequencies and bandwidths of the filters follow the Mel-scale which is approximately logarithmic giving more weight to spectral features below 1 kHz which contain important formant information. The filter bank is not totally effective because the critical-band filter spacing and bandwidth below 1 kHz is usually 100 Hz and the pitch frequency of female speech can exceed this. To remove pitch artefacts entirely further processing, such as discarding the higher cepstral components following a cosine transform, is required. A more serious concern is that the pitch peaks in the spectrum forming the spectral envelope which contains the speech information can be attenuated if they do not align with the centre frequencies of the filters. The intervening low-power parts of the spectrum are not attenuated. While both the filter-bank and the auditory system use critical-band filters the centre frequencies of these filters is fixed in the filter-bank. The auditory system has several thousand filters allowing the critical-band response to be centred on the pitch peaks in the spectrum.

An algorithm which estimates the spectral envelope of the speech signal using the peaks of the spectrum would more closely model the auditory system and should, by rejecting the parts of the spectrum where the noise dominates, be more robust to the effects of additive noise on the speech signal. When the energy of the speech sound is low this signal is lost entirely; when the energy is higher, on voiced sections of speech, the noise distorts low-level parts of the spectrum leaving the high energy parts of the spectrum unchanged. This is illustrated in Figure 1 which shows two different sets of noise samples added to the same frame of female voiced speech. While the low level parts of the spectrum are markedly different in the two traces the pitch peaks in the spectrum remain unaffected and the difference between them remains small.

In this paper we present in Section 2 a new algorithm for feature extraction which can be used independently or in combination with a filter-bank. This computationally simple algorithm is a superior model of the human auditory system. The experiments described in Section 3 show that the algorithm significantly improves the performance of an isolated digit recognizer in additive noise particularly for female speech. Section 4 contains a discussion of the results.

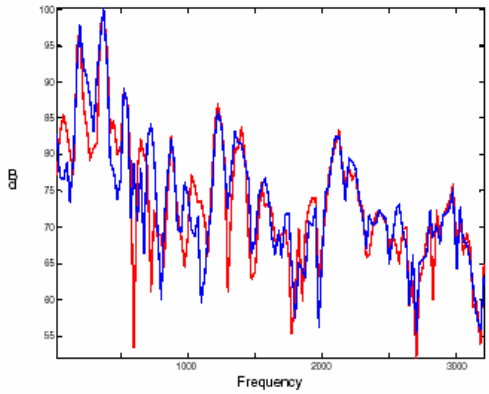


Figure 1: Spectra of a Frame of Voiced Speech at a 3dB Signal to Noise Ratio

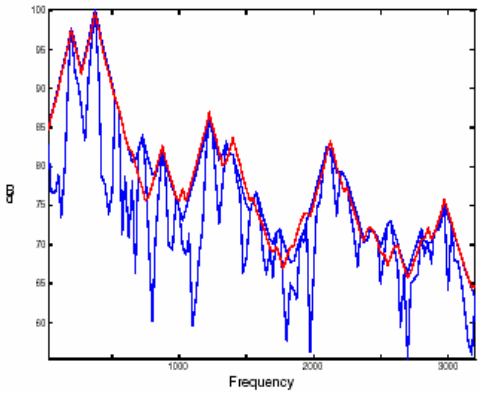


*Figure 2:*Original and Smoothed Spectra of a Frame of Voiced Speech at a 3dB Signal to Noise Ratio.

2. Non-Linear Spectral Smoothing

2.1. The Smoothing Algorithm

Figure 1 shows the logpower spectrum of a frame of voice speech, to which two different sets of noise samples at a level 3dB below that of the speech utterance were added. While there are significant differences between the low-level parts of the spectra the spectral peaks which contain the speech envelope information are little affected by the noise. A technique is required which transforms the spectrum so that the spectral peaks remain unchanged but the lower level parts of the spectrum are ignored. This can be achieved by nonlinear spectral smoothing (NLSS) defined as follows:

$$f'(i) = Max(f(i), s_l f(i-1), s_u f(i+1)) \qquad (1)$$

Where $f(i)$ is the original spectrum, $f'(i)$ is the smoothed spectrum and s_l and s_u are smoothing constants. These constants lie in the range 0.95 to 0.99. Each coefficient in the original spectrum either remains unchanged or is replaced by a neighbouring exponentially decayed coefficient if this is larger. The range in which this replacement occurs is determined by the values of the upper and lower smoothing constants. Smaller constants restrict the range of replacement to coefficients close to the spectral peaks while larger values extend their effect more widely across the spectrum.

Figure 2 shows one of the spectra of the speech frame of Figure 1 and both spectra after NLSS. The peaks of the spectrum have been retained but the low level parts of the spectrum are ignored. As can be seen there is now a reduction in difference between the two spectra. When used in a recogniser this should result in better matches between noise contaminated spectra. Readers familiar with analogue communication systems will recognize that the algorithm which is a form of rank-order filter [3], is analogous to an amplitude demodulator used to extract the information bearing envelope from the rectified carrier in an amplitude modulated system.

2.2. Auditory Masking

NLSS Has close parallels with auditory masking. This can be demonstrated with reference to Figure 3. This shows a spectrum comprising two tones and gaussian white noise. The smoothed spectrum accurately represents the spectral peak caused by the louder of the tones. Away from the spectral peak the smoothed spectrum tracks the peaks of the noise. The smooth spectrum however does not represent at all the secondary peak to the left of the main peak caused by the second tone or the noise peaks to the right of the main peak. The presence of these artefacts within the spectrum is masked by that algorithm's response to the dominant tone. This bears close comparison to the masking [4] which occurs in the human auditory system.

2.3. Mel Coefficients

NLSS was originally proposed as an intermediate processing stage lying between the FFT and the FBA. Figure 4 shows spectrograms of the utterance *'two'* spoken by a woman before and after NLSS. In Figure 4b the pitch information which dominates Figure 4a has been removed, the important

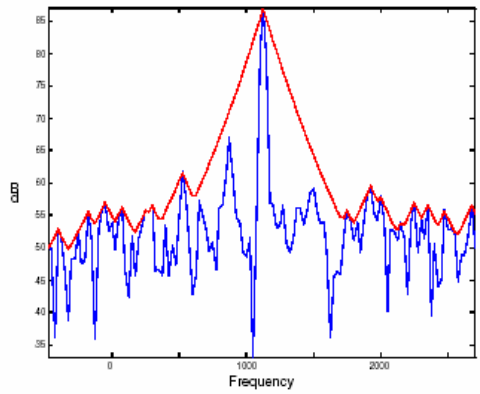


Figure 3: Original and Smoothed Spectra of Two Tones and Added Noise

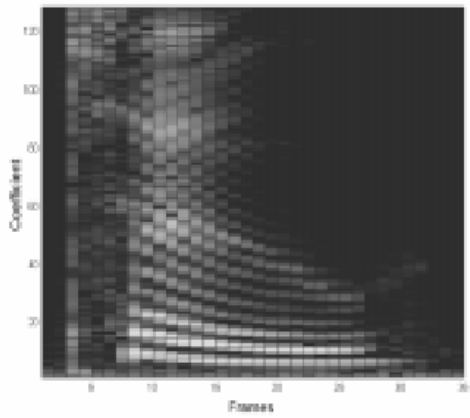


Figure 4a Original Spectra of a Frame of Voiced Speech at a 3dB Signal to Noise Ratio.

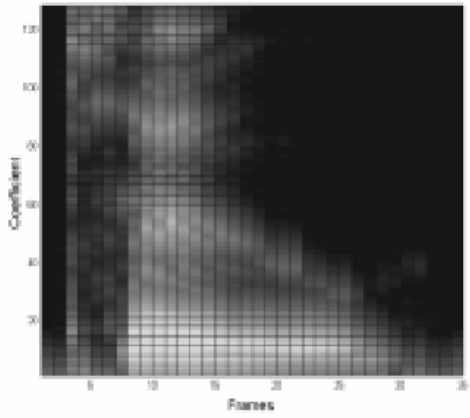


Figure 4b Smoothed Spectra of a Frame of Voiced Speech at a 3dB Signal to Noise Ratio.

spectral elements including the formant structure have been preserved.

This leads to the conjecture that it may be possible to use the smoothed spectrum directly dispensing with the filter bank entirely. As mentioned in Section 1 however the generation of features lying on a logarithmic frequency scale is an important attribute of the auditory system and the FBA. The coefficients produced by NLSS lie on a linear frequency scale. An approximation to logarithmically spaced coefficients can be made by dividing the frequency range into three bands, 0-1 kHz, 1-2 kHz and 2-4 kHz. All the coefficients lying in the first band are retained; the coefficients lying in the second band are decimated by two while those in the third band are decimated by four. This gives an equal weighting to frequencies in the first band and to those in the second and third bands combined, similar to a Mel scale filter bank.

3. Experiments

3.1. Experimental Set Up and Database

Experiments conducted during this investigation used the Noisex database [5]. This comprises files of six different background noises including car, tank, aircraft, helicopter, and operation room noise. In addition there is random noise with a spectral characteristic which is typical of a speech signal. Experiments were conducted using this speech spectrum noise (Noise06). The database also contains recordings of two speakers, a man and a woman, saying isolated and connected digits. The isolated digits were used for these experiments. Each speaker utters the ten digits twenty times. These were end-point analyzed and divided into ten repetitions of the digits for training and ten for testing. Noisy test and training files were generated by adding the noise signals in the noise files to the speech signals at different levels to give signal to noise ratios of +24dB to -6 dB. Noise was added to the training material from the second half of the noise files and to the test material from the first half of the noise files. Noise statistics for Parallel Model Combination (PMC) were also estimated from the second half of the noise files. The data as supplied is sampled at 16 kHz however for the purpose of these experiments the sampling rate was reduced to 8 kHz.

The reference system comprised a front end consisting of a 256 point FFT preceded by a pre-emphasis filter and followed by a nineteen-filter Mel-scale filter bank as described in reference [5]. In order to maintain the simplicity of the system and focus differences in performance on the front-end processing a simple dynamic programming algorithm was used for pattern matching. A discrete cosine transform was used to produce twelve Mel frequency cepstral coefficients, MFCCs. These were liftered prior to the distance calculation and Viterbi alignment. The training data was divided into ten sets of digit utterances to which each of the 100 tests digits were matched. This resulted in 1,000 trials for each of the conditions under investigation.

Experiments were carried out using the smoothed FFT output with 128 lineary spaced coefficients and with Mel spaced coefficients using the approximation described in section two. This gave 64 coefficients in total, thirty-two lying between 0 and 1 kHz, sixteen between 1 and 2 kHz and sixteen lying between 2 and 4 kHz. The upper and lower smoothing factors were set to be the same and in the range 0.95 and 0.99. The filter-bank analyzer was used with and without NLSS.

3.2. Results with Matched Noise

In the first set of experiments the training and testing noise conditions were matched. The results of these experiments are shown in Table 1. The first point of note is that the performance on female speech was superior to that on male speech which is at odds with most other experimental findings. This is believed to be an artefact of the Noisex database. Because women's voices are on average less loud than men's voices a poorer signal to noise ratio results when the background noise level is held constant. However in the Noisex database it is the signal to noise ratio which is normalized for both speakers giving the observed improvement in performance on the female voice.

The first two results show the reference FBA system with and without NLSS. There is a significant reduction in error rate when NLSS is used. The next result showed the effect of using the FFT coefficients directly; the poor performance

Test	Male	Female	Ave.
FBA	12.8	4.7	8.8
FBA NLSS	7.4	2.5	5.0
Lin FFT	18.5	13.0	15.8
Lin FFT NLSS	8.7	3.1	5.9
Mel FFT NLSS	7.0	2.0	4.5

Table 1. Error Rate with Matched Conditions and Speech Spectrum Noise at a 3dB Signal / Noise Ratio

achieved vindicates the use of the FBA in conventional recognition systems. However when NLSS is applied to the FFT coefficients there is a considerable improvement in performance but the linear spaced FFT coefficients do not out perform the FBA with NLSS. The last line of Table 1 shows that transforming the FFT coefficients on to a Mel scale produces results which are better than the other feature sets including the FBA with NLSS. This indicates that when NLSS is included in the system the FBA's only benefit is to effect a transformation on to a Mel scale. This raises the question as to whether the FBA is necessary when NLSS is included in the system.

3.3. Results with Parallel Model Combination

Table 2 shows the results achieved using Parallel Model Compensation [6] as a noise compensation algorithm. For each set of features a set of corresponding noise means were computed from a half second section of the noise file. This was scaled to match the level of noise added to the test utterances and combined with the templates by adding the FBA coefficients or the NLSS coefficients as appropriate in the linear power domain. Since the pattern matching stage assumes unit variances the variances of the new models were left unchanged. This is equivalent to the Weiner filtering technique described in Reference [7].

The reference FBA system was tested with and without

Test	Male	Female	Ave.
FBA –Comp	1.2	2.3	1.8
FBA – NLSS - Comp	1.2	0.7	1.0
Mel FFT NLSS Comp	0.9	0.3	0.6

Table 2. Error Rate with Parallel Model Combination Speech Spectrum Noise at a 3dB Signal to Noise Ratio

NLSS together with the Mel FFT system with NLSS. The performance of all three of these systems was significantly superior to the match noise results presented above. It is usually found that noise compensation using PMC gives results a little worse than those achieved under matched conditions. Again we find that the system with Mel FFT coefficients and NLSS out performed the reference FBA system even when the FBA included NLSS.

4. Discussion

Figure. 5 graphs the error rate performance of the different feature sets as the signal to noise ratio is varied between +18 and -6dB. It shows that the systems with NLSS achieve the same error rate when recognizing speech with signal to noise ratios 2-3dB worse than the reference FBA system. The

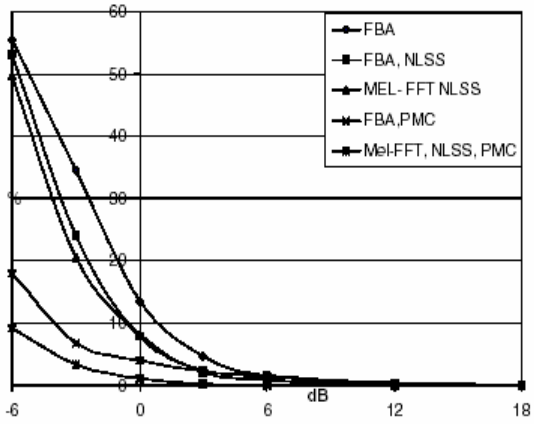


Figure 5: Error Performance on Female Speech of the Feature Sets as a Function of Signal to Noise Ratio

recognition results demonstrate that the technique gives a maximum benefit at signal to noise ratios lying between +3 and -6dBs. Figure 1 illustrates that at a noise level of 3dB the spectral maxima caused by the pitch pulses are unaffected by the noise which corrupts the spectrum has levels lying between 6 and 10dB lower. One surmises that at higher signal to noise ratios the noise is insufficient to affect adversely the estimation of the spectral envelope. Once the signal to noise ratio drops below 0dB weaker parts of the speech signal start to be completely masked by noise making it difficult to discriminate between words such as '*five*' and '*nine*' which only differ in low-level parts of the spectrum accounting for the observed increase in the error rate. Between these levels the noise-masking properties of NLSS give the performance gain observed above.

5. References

[1] Davis, S. and Mermelstein, P. ' Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Trans. Acoustics, Speech and Signal Processing, Vol.28*, pp357-366

[2] Kandel E.R., Shwartz, J.H., and Jessell T.M., ' *Principles of Neural Science*', 4th Edition, McGraw Hill, New York,2001 pp 591-602

[3] Yin L, Yang R, Gabbouj M, Neuvo Y., 'Weighted Median Filters: A Tutorial'. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal* Processing, 1996; Vol43(3):157-192

[4] Holmes, J. and W., *Speech Synthesis and Recognition*, 2nd Edition, Taylor and Francis, London, 2001, p 39

[5] Varga, A.P., Steeneken, H.J.M., Tomlinson, M., Jones, D., 'The Noisex-92 study on the effect of additive noise on automatic speech recognition', *Technical Report* Speech Research Unit, Defence Research Agency, Malvern, U.K.

[6] Gales, M.J.F., and Young, S. J. 'Robust speech recognition in additive and convolutional noise using parallel model combination', *Computer, Speech and Language*, *Vol.9*, pp 289-307

[7] Vaseghi, S. V. '*Advanced Signal Processing and Digital Noise Reduction*, J. Wiley, 1996, pp 135-6