en cs de sk es

TECH pedia

MULTIMODÁLNÍ ROZHRANÍ MMI (MULTIMODAL INTERFACES) PRO INTERAKTIVNÍ KOMUNIKACI A NAVIGACI

JURAJ KAČUR, GREGOR ROZINAJ, RENATA RYBÁROVÁ

Název díla: Multimodální rozhraní MMI (MultiModal Interfaces) pro interaktivní komunikaci a navigaci

Autor: Juraj Kačur, Gregor Rozinaj, Renata Rybárová

Přeložil: Ivan Pravda

Vydalo: České vysoké učení technické v Praze

Fakulta elektrotechnická

Kontaktní adresa: Technická 2, Praha 6

Tel.: +420 224352084

Tisk: (pouze elektronicky)

Počet stran: 41

Edice (vydání): 1. vydání, 2017

ISBN 978-80-01-06241-8

TechPedia
European Virtual Learning Platform for Electrical and Information Engineering

http://www.techpedia.eu

Erasmus+

Tento projekt byl realizován za finanční podpory Evropské unie.
Za obsah publikací odpovídá výlučně autor.
Publikace (sdělení) nereprezentují názory Evropské komise a Evropská komise neodpovídá za použití informací, jež jsou jejich obsahem.

Vysvětlivky

	Definice
	Zajímavost
	Poznámka
	Příklad
	Shrnutí
	Výhody
	Nevýhody

ANOTACE

Multimodální rozhraní MMI (*MultiModal Interface*) se skládá z různých částí a modulů, které slouží pro přirozenou a intuitivní komunikaci s vlastním systémem. Všechny dohromady představují funkcionalitu MMI. Do základního souboru modulů, které mohou být součástí multimodálního rozhraní, mohou patřit např. rozeznání řečníka, rozeznání řeči a hlasových povelů, rozeznání obličejů, navigace pomocí gest, navigace pohybem očí a syntéza řeči.

CÍLE

Hlavním cílem tohoto výukového kurzu je seznámit studenty se základy multimodálního rozhraní a multimodálními aplikacemi. Studentovi jsou představeny základní principy rozeznávání tváří, identifikace mluvčího, navigace pomocí hlasových povelů a gest. Konec modulu je věnován možnému vyžití popsaných modalit v multimodální aplikaci.

LITERATURA

[1] Probst, Michael; Schumann, Sebastian; Rozinaj, Gregor; Minarik, Ivan; Rybárová, Renata; Oravec, Miloš. EVALUATION: Final Multimodal Interface for User/Group-Aware Personalisation, Deliverable 5.5.1, available at http://www.hbb-next.eu/index.php/documents, Decmber 2013.

[2] Vančo, Marek; Minárik, Ivan; Rybárová, Renata. Evolution of static gesture recognition. In: Redžúr 2014 proceedings; 8th International Workshop on Multimedia and Signal Processing; 13 May 2014, Dubrovnik, Croatia. Bratislava: Nakladateľstvo STU, 2014, p. 41-44. ISBN 978-80-227-4162-0.

[3] Rautaray, Siddhart S., Agrawal, Anupam. Vision based hand gesture recognition for human computer interaction: a survey, Springer, 2012, pp 106-113. ISBN 978-3-642-27386-5

[4] Kristensson, Per Ola; Denby, L.C. Continuous recognition and visualization of pen strokes and touch-screen gestures. In Proceedings of the Eighth Eurographics Symposium on Sketch-Based Interfaces and Modeling, 2011, pp. 95-102.

[5] Gülbahar, Mark; Probst, Michael; Schumann, Sebastian; Rozinaj, Gregor; Minarik, Ivan; Rybárová, Renata; Oravec, Miloš. Design and Protocol: Intermediate Multimodal Interface and Context Aware Recommendation Engine, Deliverable 5.3.1, available at http://www.hbb-next.eu/index.php/documents, October 2012.

[6] T. Kinnunen, H. Li, An overview of text-independent speaker recognition: from features to supervectors, Speech communication, Vol. 52, pp. 12-40, Elsevier, 2010

[7] Benesty J., Sondhi M. M., Huang Y., Springer Handbook of Speech Processing, Springer-Verlag, ISBN 978-3-540-49125-5, 2008

[8] Uhlíř, J. - Sovka, P. - Pollák, P. - Hanžl, V. - Čmejla, .: Technologie hlasových komunikací. Nakladatelství ČVUT, 2007.

[9] Psutka, J. - Müller, L. - Matoušek, J. - Radová, V.: Mluvíme s počítačem česky. Academia 2006.

Obsah

1 MMI architektura

V současné době je nejrozšířenějším rozhraním sloužícím ke komunikaci člověka s počítačem klávesnice s počítačovou myší anebo dotykový displej. Tato zařízení představují adaptaci lidí úrovni počítačů ve spojení s přirozeným způsobem vzájemné komunikace a ovládání. V posledních letech se však začaly objevovat požadavky lidí na potřebu komunikovat s počítači rovnocenným způsobem jako mezi sebou samotnými, tj. řečí, mimikou a gesty. Takto totiž dokáží lidé přenést mnohem více informací než použitím periferních zařízení. Toto nás přivádí k pojmu multimodální rozhraní **MMI** (*MultiModal Interface*).

Multimodální rozhraní se skládá z různých částí a modulů, které slouží pro přirozenou a uživatelsky příjemnou formu komunikace, resp. interakce se systémem. Spolu tyto moduly představují funkce MMI. Součástí **MMI** mohou být tyto moduly:

- hlasová identifikace
- rozeznání řeči a hlasových povelů
- rozeznání více tváří
- rozeznání gest a navigace pomocí gest
- navigace očima
- syntéza řeči
- predikční modul (modul anebo systém pro filtrování informací, a který se snaží predikovat ohodnocení nebo preferenci uživatele, kterou by dal zvolené položce)

Všeobecně se architektura multimodálního rozhraní skládá z několika vrstev. Fyzická vrstva představuje hardwarová vstupní a výstupní zařízení umožňující přímou interakci s okolím. Multimodální údaje poskytované vstupními zařízeními (např. kamery, senzory, mikrofony, atd.) jsou zpracované paralelně každým modulem zvlášť (moduly jsou nad fyzickou vrstvou, tato vrstva se nazývá vrstvou modalit). MMI řadič (na vrstvě **MMI**) sbírá výstupní data ze všech modulů, vyhodnocuje je a následně spojuje do jednoho výstupního datového toku. Výsledný tok informací obsahuje informace o uživatelích a autorizaci o jimi požadovaných akcií. Tento tok se přenáší až do aplikační vrstvy.

2 Identifikace uživatele

V současné době se informační technologie dostávají stále víc a víc do popředí. Setkáváme se s nimi v našem každodenním životě od malých „chytrých“ telefonů až po inteligentní budovy.

Každé zařízení pracující s našimi osobními údaji by proto mělo být chráněno určitou formou a úrovní zabezpečení. Nejrozšířenějším typem ověření neboli autentifikace je dvojice uživatelského jména a hesla, avšak s rozvojem informačních technologií a stále rostoucí skupinou dostupných algoritmů pro zajištění ochrany a bezpečnosti, začínají používat informační systémy a aplikace autentifikaci na základě biometrických údajů. Například autentifikace na základě rozeznání obličeje je součástí systému Windows 10 pod názvem Windows Hello. Využití biometrických údajů přináší několik výhod. Biometrické údaje jsou jedinečné pro každou osobu, takže je postačující přítomnost daného člověka před senzorem či snímačem a není tak nutné si pamatovat několik různých hesel či PINů. Informační systémy dnes mohou pracovat s různými biometrickými principy identifikace osob, např. pomocí duhovky očí, otisků prstů, barvy hlasu, výrazu tváře, atd. V další části se všeobecně zaměříme na systémy pro rozeznávání obličejů a identifikaci mluvčího.

2.1 2D a 3D rozeznání obličeje

Rozeznání obličeje je podmnožinou velké oblasti výzkumu a technologií zaměřených na rozeznávání vzorů. Rozeznání lidského obličeje je jednou z nejdůležitějších metod autentifikace pomocí biometrických údajů užívanou v posledních několika desetiletích, a to především díky velkému potenciálu a uplatnění v široké škále aplikací a různých oblastech využití (bezpečnostní dohled, zabezpečení domácnosti, hraniční kontroly, atd.) Biometrické systémy pro identifikaci osob, které jsou vyvíjeny různými společnostmi, dosahují vysoké přesnosti rozeznání obličeje. Většina těchto aplikací musí splňovat [1]:

- dokáže rozeznat několik obličejů z jednoho video záběru anebo jednoho obrázku
- vysokou úspěšnost rozeznání
- nezávislost na světelných podmínkách
- stabilitu při změně výrazu tváře anebo pózy
- rozeznání v reálném čase, atd.

Hlavní výhodou v porovnání s ostatními přístupy je to, že rozeznání obličeje nevyžaduje explicitní souhlas uživatele, protože obrazy obličeje je možné získat kamerou i z větší vzdálenosti. Další výhodou je, že záznamová zařízení (např. kamery) jsou laciné a stávají se zcela běžnou a dostupnou komoditou.

Hlavní nevýhodou rozeznání obličeje je relativní jednoduchost, která však může potenciálně způsobit její nepřesnost.

V porovnání s identifikací mluvčího, rozeznání obličeje dosahuje mnohem lepších výsledků. Všeobecně existují tři hlavní přístupy založené na typu dat, které se používají v procesu rozeznávání. Patří sem metody založené na 2D obrazech, na údajích získaných z 3D tváří a techniky, které používají oba typy dat. Celý proces rozeznávání se skládá z 3 hlavních fází. První z nich je získání a předzpracování dat, druhou je registrace dat a třetí fází je vlastní rozeznání. Detailní popis 2D a 3D metod na rozeznání obličeje je možné najít v modulu *Identifikace uživatele*.

U moderních systémů je možno implementovat rozeznání více obličejů současně. V tomto případě jsou detekovány všechny tváře na snímcích nebo v obrazu. Polohy detekovaných tváří jsou zapamatovány pro daný obraz. Dále je obraz rozdělen na několik částí (vzorků) na základě pozice detekovaných tváří. Takto generované vzorky reprezentují všechny tváře v obrazu a umožní sledovat každý obličej samostatně. Rozdělení obrazu je hlavním krokem vedoucím k rozeznání více obličejů v systému nebo aplikaci.

2.2 Identifikace mluvčího

Stručně řečeno, cílem identifikace mluvčího je na základě hlasu určit jeho totožnost. Rozhodovací proces je realizován na základě množiny hovorových vzorků mluvčích, které jsou uloženy do databáze během fáze tzv. trénování. Systém je tedy schopen rozpoznat pouze ty uživatele, jejichž vzorky jsou uloženy v databázi, resp. rozeznání není provedeno, pokud je věrohodnost jeho výsledku příliš nízká.

Protože je počet uživatelů v databázi omezený, tato úloha se též označuje jako tzv. problém uzavřené skupiny. Identifikací na základě hlasu se věda věnuje už víc než 40 let a vývoj v této oblasti stále intenzivně pokračuje. Příchodem nových a dostupných technologií proniká identifikace mluvčího do velmi mnoha nových oblastí, ze kterých mezi nejvýznamnější patří:

- **Kriminalistika**
- **Automatická indexace hovorových a audio nahrávek**
- **Neinvazivní kontrola přístupu k citlivým datům, službám, atd.**
- **Herní průmysl**
- **Pomůcky pro postižené**

Identifikace mluvčího je poměrně složitou úlohou, a to z hlediska mnoha důvodů. Mezi ty nejvýznamnější patří:

- **Akustická variabilita mluvčího**
- **Změny hovorového projevu vlivem zdravotního stavu, nálady, psychického stavu, atd.**
- **Přítomnost šumu pozadí**
- **Vlivy zapříčiněné nedokonalostmi nahrávacích zařízení a okolním prostředím**
- **Zjednodušená reprodukce zvuku pomocí nahrávacích zařízení**

Na druhou stranu použití řeči jakožto biometrického signálu má následující výhody:

- **Není třeba používat heslo**
- **Jedná se o neinvazivní metodu**
- **Jednoduchá metoda získávání biometrického signálu**

Existuje velmi mnoho aplikací sloužících k identifikaci uživatele, které se mezi sebou liší komplexností, požitými metodami, v koncepci, ve způsobu činnosti,

požadavcích a omezeních, atd. Proto se zažily jisté kategorie, a mezi ty nejvýznamnější patří:

- Textově nezávislé systémy – tyto systémy pracují nezávisle na obsahu hovoru, tj. nevyžadují předepsaná slova nebo fráze, vyžadují pouze dostatečnou délku hovoru. Úspěšnost těchto systémů se v závislosti na počtu uživatelů a typu prostředí pohybuje v rozmezí od 80 až 95%. Tyto systémy jsou nejběžnější, a to i přesto, že vykazují nižší úspěšnost identifikace mluvčího.

- Textově závislé systémy – požadují pro svou správnou činnost využití předepsaných frází nebo hesel v hovoru. Tyto systémy dále umožňují využití i koartikulačních efektů, ke kterým dochází při výslovnosti sousedních hlásek a případně je možné využít i neznámého hesla. Tyto systémy proto dosahují vyšší úspěšnosti v rozmezí od 95 do 99%. Ve speciálních případech tyto systémy mohou měnit očekávaný text pomocí různých doprovodných otázek.

- Systémy pracující s akustickými příznaky – využívají extrahované akustické informace hovoru, jako jsou různá modifikovaná krátkodobá spektra hovoru vystihující formantové frekvence hlasového traktu. Tyto příznaky v sobě odrážejí fyzické vlastnosti hlasových orgánů, jako jsou jejich velikost, tuhost, tvar atd. Jednoduše se získávají a jsou základem skoro všech systémů.

- Systémy pracující s prozodickými informacemi (rytmickými vlastnostmi jazyka) – extrahují a dále zpracovávají specifické informace jako je dynamika řeči, rychlost promluvy, modulace hlasu, pauzy, přízvuk, atd. Určují se zpravidla z delších intervalů hovoru, ale jsou obtížněji detekovatelné a kvantifikovatelné.

- Systémy pracující s příznaky vyšší úrovně – používají informace jako je dialekt, jazyk, slovná zásoba, atd., ze kterých lze určit původ a sociálně-vzdělanostní postavení jedince. Jsou však málo diskriminující.

- Systémy pracující v reálném čase – poskytují skoro okamžitou informaci o mluvčím se zpožděním pouze několika sekund.

- Systémy nepracující v reálném čase – poskytují informaci až po vyhodnocení celého hovoru a proto mohou používat i složitější algoritmy, delší časové intervaly a dosahovat tím lepších a přesnějších výsledků.

Tato základní dělení, koncepty a tvrzení v sobě zahrnují pouze výčet metod spolu se zjednodušeným přehledem jejich vlastností a omezení. Proto je kvůli podrobnějšímu výkladu dané problematiky možné využít přehledové články jako např. [6].

2.3 Víceúrovňová identifikace

K některým aplikacím (např. *AppStore*) může přistupovat více uživatelů z jednoho nebo více typů zařízení. Pro tento účel je zavedená tzv. víceúrovňová identifikace.

Víceúrovňová identifikace se zabývá různými požadavky na přístupová práva nebo kontrolu prostřednictvím kombinace několika biometrických způsobů identifikace uživatele v kombinaci s uživatelským jménem, heslem nebo PINem. Ze skupiny možných biometrických způsobů identifikace uživatele slouží k jeho identifikace zejména hlas a obličej. V závislosti na výsledku identifikačního procesu může být přístup k systému nebo aplikaci přímo umožněn, nebo může být požadováno další ověření pomocí přihlašovacích údajů – tj. uživatelským jménem, heslem nebo PINem. V budoucnosti může být identifikace v systému rozšířena použitím rozeznání duhovky, a nahradí tak použití PINu nebo hesla. Celkově se tím zvýší bezpečnost celého systému.

3 Komunikace pomocí hlasových povelů

Komunikace pomocí hlasových povelů spadá do rozsáhlé vědecké oblasti označované termínem *automatické rozeznávání řeči* **ASR** (*Automatic Speech Recognition*).

Tento termín označuje automatický (strojový) přepis mluveného slova. Vstupem je digitální hovorový signál, přitom na výstupu je textový řetězec označující rozpoznaná slova daného jazyka, která se nacházejí v automatem používaném slovníku. Slovník celého jazyka je však obrovský, proto se obyčejně používají jen jeho omezené verze určené pro konkrétní oblast nasazení. Výstupní text vyhovuje předepsané gramatice nebo běžně používaným slovním spojením existujícím v zvoleném jazyce.

ASR se odlišuje od úlohy porozumění obsahu hovoru, které pracuje na vyšší úrovni. Oblast, která se tímto problémem zaobírá, se nazývá umělá inteligence. Jejím vstupem je naopak text po zpracování systémem **ASR**.

V průběhu několika posledních desetiletí probíhá intenzivní výzkum v oblasti konstrukce **ASR** systémů, které se využívají v následujících oblastech: systémy poskytování informací, dialogové systémy, pomůcky pro hendikepované, přepisy mluveného slova, pro účely archivace atd. I přesto je to teprve nedávno, kdy byly předvedeny první použitelné systémy, které se dají využít i mimo kontrolované prostředí laboratoří. Zvětšuje se jejich komplexnost a aplikace přepisu diktovaného slova nebo dokonce i plynulé řeči či běžné konverzace už nejsou úplnou raritou. Tato problematika je velmi složitá a musí řešit problémy zasahující do většího počtu vědních oborů a disciplín. Praktické systémy musí být schopny pracovat i v nepříznivých podmínkách (změna pozadí a prostředí, změna nahrávacích zařízení, atd.), musí podporovat rozsáhlou gramatickou pestrost jazyka, variabilitu mluvčích (akustické podoby), rozsáhlé slovníky, atd. Navíc běžný uživatel vyžaduje okamžitou reakci systému, není ochoten měnit svůj způsob přirozeného projevu, omezit se na danou slovní zásobu a také rychle ztrácí trpělivost v případě možných chyb **ASR** systému.

Tak jako existuje rozličné množství **ASR** aplikací s různým stupněm složitosti stanovených požadavky a omezeními, tak byly vytvořeny klasifikace těchto systémů dle specifických kritérií. Tím nejdůležitějším z nich je velikost podporovaného slovníku. Rozlišují se následující dělení:

- **Malý slovník –** obsahuje několik desítek až stovek slov
- **Střední slovník –** obsahuje několik stovek až tisíců slov
- **Velký slovník –** několik desítek až stovek tisíc slov

Toto dělení se přirozeně s růstem technologie mění.

Systémy mohou být dále závislé nebo nezávislé na mluvčím. To znamená, že systém nezávislý na mluvčím musí pracovat nezávisle na tom, který mluvčí právě hovoří. To si žádá existenci vzniku všeobecných modelů získaných z velkého množství mluvčích. Systémy závislé na mluvčím jsou přizpůsobeny pro konkrétního uživatele, a proto obvykle dosahují lepších výsledků, ale jen v kontextu pro daného mluvčího. Systémy, které poskytují okamžitou odpověď, se nazývají systémy pracující v reálném čase. Významnou roli zde hraje i to, v jakém tvaru je řeč očekávaná na vstupu. Podle toho je možné odlišit následující scénáře:

- Rozeznání izolovaných slov – na vstupu je očekáváno jedno slovo ze slovníku – jedná se především o povely
- Přepisové systémy – předpokládají, že je za sebou v promluvě vetší množství slov, ale je mezi nimi dostatečná, dobře detekovatelná mezera.
- Plynulá řeč – kde jdou slova plynule za sebou, ale jsou zde víceméně uplatňována omezující pravidla na použitou gramatiku.
- Přirozená řeč – jedná se o nejsložitější úlohu, která obsahuje plynulou řeč bez omezující gramatiky a uskutečněnou v reálném prostředí.

Nakonec je možné systémy dělit podle toho, jakou část řeči modelují (fonémy, slabiky, slova, fráze, atd.) a jakým způsobem ji modelují, např. statistický přístup pomocí skrytých Markovovských modelů **HMM** (*Hidden Markov Models*).

Hovorový signál je produkován hlasovými orgány člověka a je pozorovatelný jako vibrování zvuku. Mimo jiné nese lexikální informaci (tj. význam toho, co bylo řečeno). Ta je reprezentovaná sekvencí vhodných zvuků, které mají specifické akustické formy. Ty akustické jednotky, které stačí na vytvoření daného jazyka, se nazývají fonémy. Jejich počet se může měnit a pohybuje se od 40 do 60 v závislosti na konkrétním jazyce. Samozřejmě jejich aktuální akustická podoba se mění od mluvčího k mluvčímu a je významně ovlivněna okolními hláskami, které vytváří tzv. koartikulační efekt.

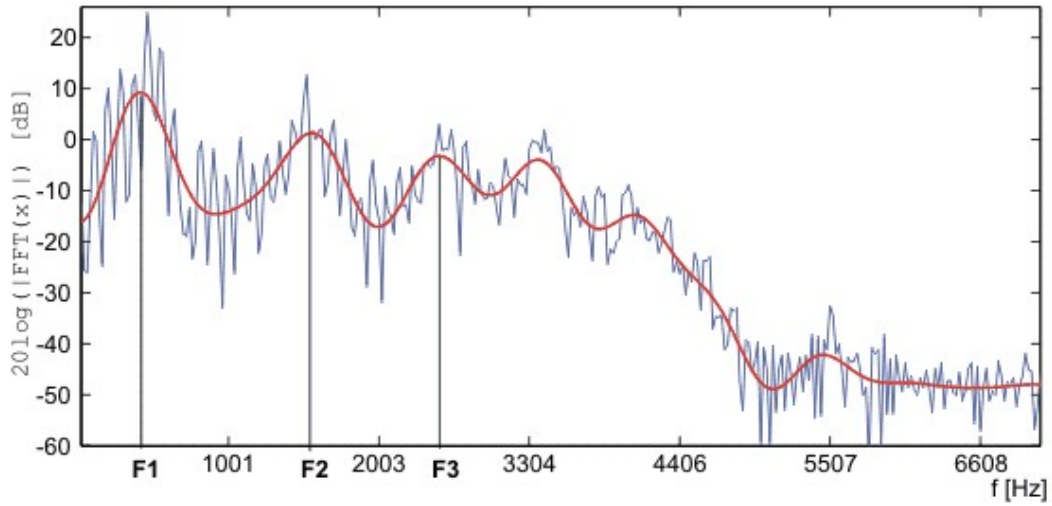
Na potlačení té části informace obsažené v promluvě, která neobsahuje lexikální informaci, je třeba použít metodu extrakce vhodných příznaků. Nadbytečná informace může totiž značně zhoršit samotný způsob rozeznání jednotlivých slov. Informace o mluvčím (jeho náladě, zdravotním a aktuálním fyzickém a emočním stavu) je tedy na překážku. Primární úlohou systémů extrakce vhodných příznaků je tedy výběr jen těch nejlepších lexikálních informací v promluvě, což ve skutečnosti není úplně jednoduché. Hovorový signál obsahuje cca 100 kbit/s informací, pro porovnání lexikální obsah obsahuje jen cca 10 bit/s informace. Proto algoritmy extrakce vhodných příznaků pracují jako účinné kompresory informací.

Dobrou ideou je tedy modelování sluchového aparátu člověka, matematicky ho popsat, zjednodušit pro praktické použití a adaptovat jej pro snadnou spolupráci s nadřízenými systémy rozeznání řeči.

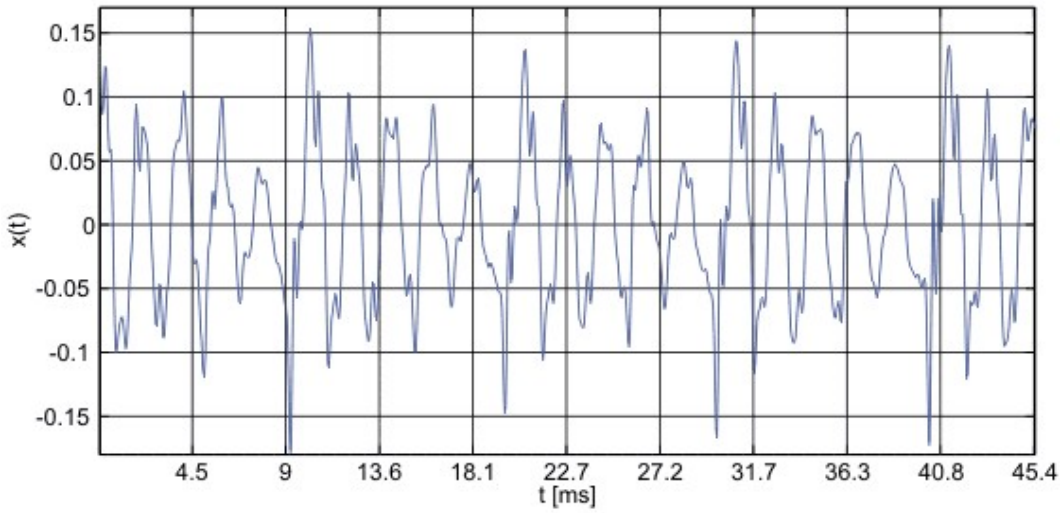
Existuje mnoho úspěšných metod modelujících buď sluchový aparát člověka (kritická pásma, křivka hlasitosti, atd.) nebo proces predikce řeči (buzení, hlasový trakt,…). Výhodou metod modelujících sluchový aparát člověka je však v tom, že mají v sobě průběhem jejich evoluce zakomponovány vhodné mechanizmy pro

zvýraznění řeči v porovnání s ostatními běžnými nepříznivými vlivy okolního prostředí. Potlačují tedy běžné šumy pozadí a zkreslení prostředí.

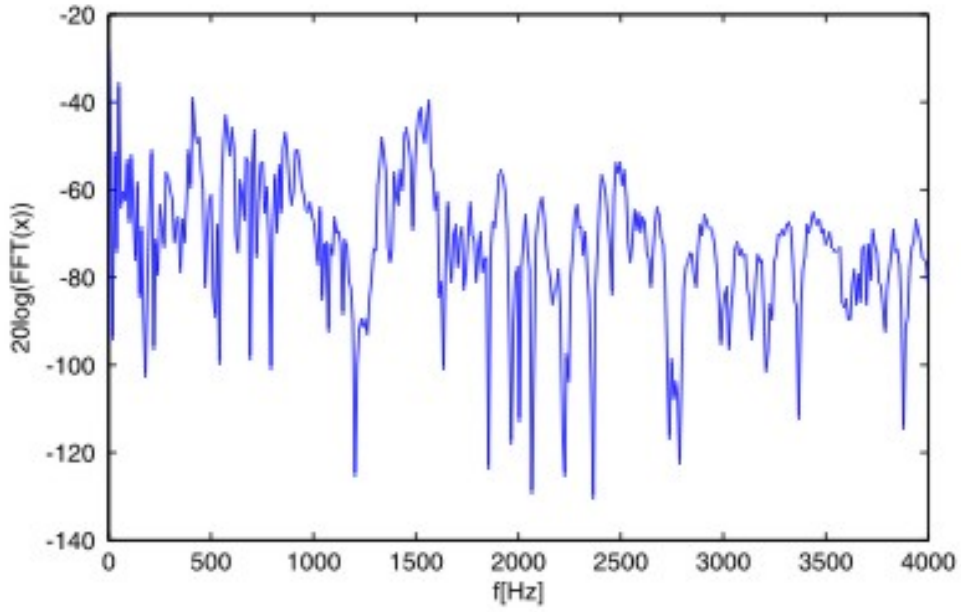
V průběhu výzkumu se zjistilo, že podstatná část lexikální informace je ukryta v obálce frekvenčního spektra signálu.Ta např. slouží ke klasifikaci jednotlivých foném mezi sebou. Přesněji je tato informace ukrytá v poloze, počtu a tvaru dominantních frekvenčních složek. Pro lepší demonstraci tohoto faktu je na obr. 3.1 zobrazeno spektrum samohlásky "*E*", obálka spektra a polohy tzv. formantových frekvencí (dominantních frekvencí, které souvisí s rezonancemi v hlasovém traktu). Na obr. 3.2 je zobrazen časový průběh stejné hlásky.



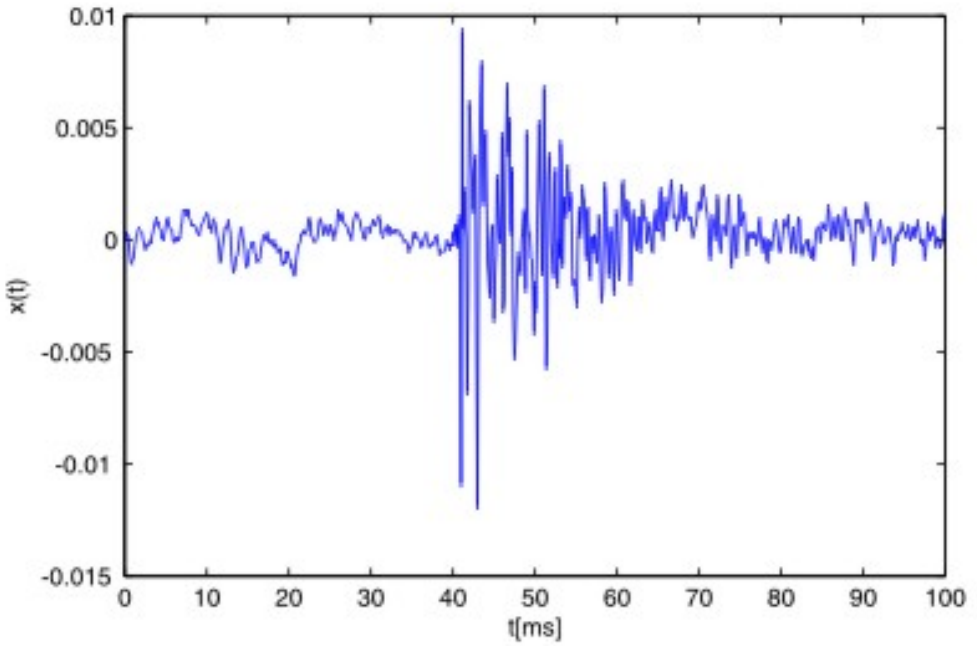
Obrázek 3.1. Spektrum formantové frekvence a obálka spektra pro hlásku "E"



Obrázek 3.2. Signál hlásky "E"



Obrázek 3.3. Spektrum hlásky "T"



Obrázek 3.4. Signál hlásky "T"

Aby byl rozdíl v hláskách patrný, je na obr. 3.3 zobrazeno spektrum hlásky "*T*" a na obr. 3.4 její časový průběh.

Kvůli sumarizaci je v následující tabulce zobrazeno rozložení prvních 2 formantových frekvencí pro slovenské samohlásky, a to odděleně pro muže a ženy, a stanovené průměrem z populace. Tato tabulka poskytuje velmi hrubý a jednoduchý způsob jak klasifikovat samohlásky na základě jejich frekvenční podoby.

Tabulka 3.1. První dvě formantové frekvence pro samohlásky, odděleně pro muže a ženy

samohláska	Muži		Ženy	
	F_1 [Hz]	F_2 [Hz]	F_1 [Hz]	F_2 [Hz]
A	730	1100	850	1200
E	530	1850	600	2350
I	400	2000	430	2500
O	570	850	590	900
U	440	1000	470	1150

Vybrané příznaky by tedy měly být schopny odhadnout rozdíly v polohách formantových frekvencí. Na druhé straně musí být „hluché“ vůči změnám, které jsou přirozené a neslyšitelné. V následující tabulce jsou uvedeny nejslyšitelnější a nejméně slyšitelné změny spektra odlišitelné sluchem.

Tabulka 3.2. Slyšitelné a neslyšitelné změny spektra

Typy změny	
Slyšitelné	**Neslyšitelné**
Počet formantových frekvencí	Celkový sklon spektra
Poloha formantových frekvencí	Frekvence pod první formantovou frekvencí
Šířka formantových frekvencí	Frekvence nad třetí formantovou frekvencí
-	Úzko-pásmové filtrování

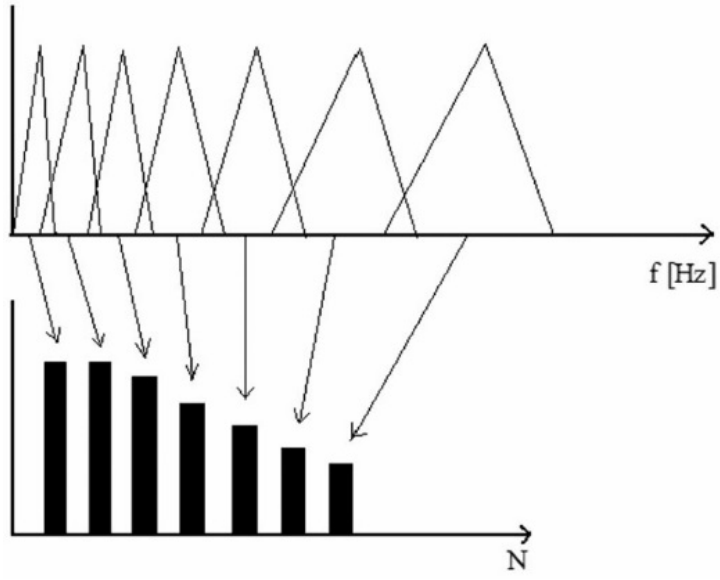
Navíc intenzita zvuku je vnímána nelineárně, což lze aproximovat pomocí logaritmické funkce.

Z pohledu rušení by měli být příznaky odolné vůči aditivnímu stejně jako konvolučnímu šumu. V neposlední řadě se musí dát příznaky jednoduše implementovat, matematicky popsat, mít kompaktní vyjádření atd. Obyčejně je dobré, pokud jsou příznaky mezi sebou lineárně nezávislé, a to kvůli redukci informačního obsahu a následnému snadnějšímu způsobu modelování.

3.1 Parametry pro rozeznání řeči

Na úvod je třeba poznamenat, že zatím neexistuje parametrizace řeči, která by splňovala všechny vlastnosti potřebné pro optimální parametrizaci uvedenou v předešlé části. Právě proto je výzkum v této oblasti stále aktivní a bylo nalezeno velmi mnoho různých způsobů a metod parametrizace. Mezi ty nejznámější a nejlepší patří Melovy frekvenční kepstrální koeficienty **MFCC** (*Mel-Frequency Cepstral Coefficients*) a perceptuální lineární predikce **PLP** (*Perceptual Linear Prediction*). **PLP** a **MFCC** se snaží modelovat sluchový aparát člověka, což vede k dobrým výsledkům v oblasti rozeznávání řeči. Obě dvě metody jsou schopny postihnout polohu, počet a šířku formantových frekvencí, které jsou vnímatelné (slyšitelné). I přes jistou podobnost se však liší v psychoakustických principech, které v sobě integrují.

Metoda **MFCC** filtruje signál filtrem typu horní propust (potlačení vlivu radiace (zakřivení) zvuku z úst), segmentuje řeč tzv. Hammingovým oknem, a následně počítá spektrum pomocí rychlé Fourierovy transformace **FFT** (*Fast Fourier Transform*). Poté se spektrum nelineárně upravuje, a to do Melovy stupnice (psychoakustická stupnice odpovídající vnímání jednotlivých frekvencí člověkem), která se rovnoměrně rozdělí na překrývající se pásma s 50% překryvem, ze kterých se následně vypočítá výkon. Tím se modeluje psychoakustická banka filtrů, viz obr. 3.5. V poslední fázi se aplikuje logaritmická aproximace a diskrétní kosinová transformace *DCT* (*Discrete Cosine Transform*). Logaritmická aproximace upravuje intenzitu zvuku tak, jak ji přibližně vnímá člověk a **DCT** snižuje závislost koeficientů mezi sebou – komprese.



Obrázek 3.5. Rozdělení spektra do bank filtrů v Melové stupnici s použitím trojúhelníkových oken

PLP příznaky se liší v několika aspektech: používají Barkovu frekvenční míru, vyhlazují spektrum a vzorkují ho v intervalech 1 Bark, čímž modelují kritická pásma pro lidské vnímání (rozlišení), váhování křivkou rovnoměrné hlasitosti,

transformace energie do hlasitosti, výpočet koeficientů modelu řeči a jejich transformace na kepstrum.

Jak je vidět, metoda **PLP** je složitější, a obsahuje víc psychoakustických principů než metoda **MFCC**, ale v příznivých laboratorních podmínkách dosahují obě metody srovnatelných výsledků.

Dynamické příznaky a energie

Řeč je posloupností různých zvuků, a proto je vhodné vyhodnocovat jejich vývoj v čase a sledovat změny, které vznikají. Nejběžnější metodou jak toto zajistit, je sledování a vyhodnocování rozdílových a akceleračních koeficientů, které se počítají v čase ze základních příznaků řeči. V principu lze toto realizovat jako jednoduchý rozdíl sousedních bloků, ale častěji je to uskutečněno jako lineární kombinace rozdílů pokrývajících širší časový úsek. Dále se prokázalo, že jednoduchá energie může dobře lokalizovat polohu samohlásek díky jejich velké energii, což zvyšuje celkovou diskriminační schopnost základních příznaků řeči. Proto se (normalizovaná) energie často přidává do výsledného vektoru základních příznaků řeči.

3.2 Metody rozeznávání řeči

Po parametrizaci je řeč daná posloupností vektorů základních příznaků řeči, např. **MFCC**. Zjednodušeně řečeno proces rozeznávání poté vezme vzorky nebo modely známé řeči (její části: fonémy, slabiky atd.) a porovná je s neznámou promluvou. Následně jsou vzorky, nebo modely s největší shodou prohlášeny za rozpoznaná slova. Díky specifickým vlastnostem řeči, jako jsou např.: různé promluvy stejného slova, se mohou tyto promluvy lišit délkou a frekvencí, atd. Navíc tato délková variabilita se nerovnoměrně mění i v rámci slova, proto mohou být některé části kratší nebo delší. Jednoduché lineární natáhnutí (lineární interpolace), resp. zkrácení pak nepřináší požadovaný efekt. Dále, výsledná promluva je daná jako řetězec jednotlivých částí (modulů) řeči, které musí na sebe navazovat. Tyto specifické vlastnosti dali vzniknout metodám rozeznávání (porovnávání vzorů) vhodných právě pro řeč. Nejznámějšími z nich jsou metody dynamické ohýbání času **DTW** (*Dynamic Time Warping*) a skryté Markovovské modely **HMM**. Samozřejmě existuje spousta modifikací či dokonce kombinací vhodných pro konkrétní aplikaci. V další části proto budou stručně popsány právě tyto dvě základní metody, tj. **DTW** a **HMM**.

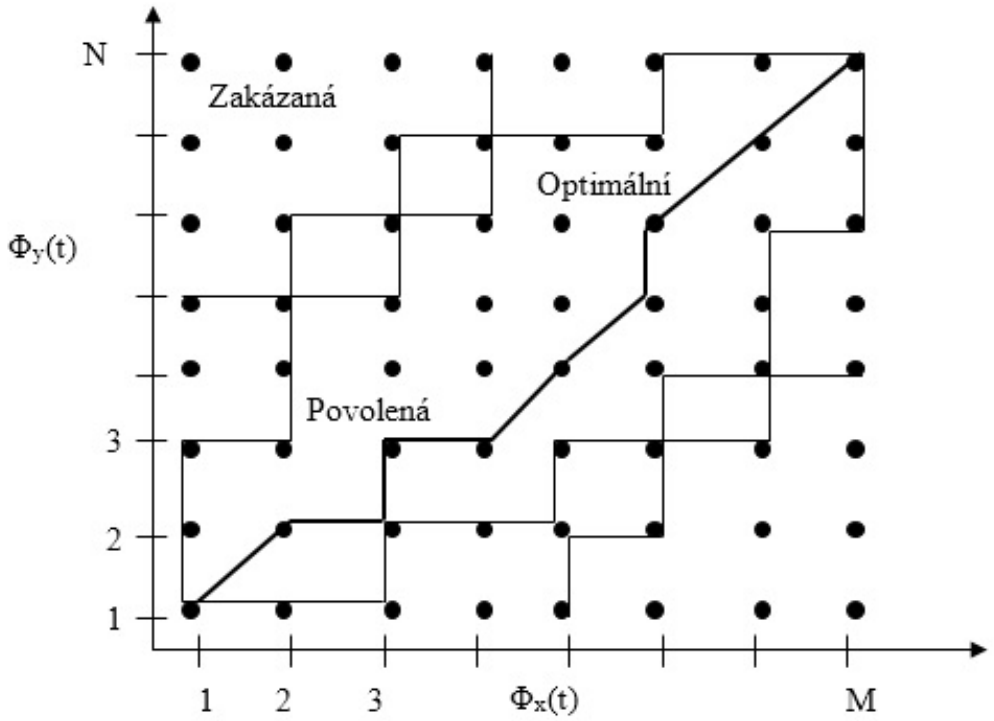
Dynamické ohýbání času (DTW)

DTW je zkratkou pro dynamické ohýbání času, které slouží pro akustické porovnání dvou všeobecně různě dlouhých sekvencí příznaků. Je založeno na nelineárním ohýbání časových indexů referenční a neznámé posloupnosti tak, aby vyhovovaly struktuře řeči, a dosahovaly co nejmenší vzdálenosti mezi sebou. Tím dochází k nelineární kompenzaci rozdílů mezi délkami obou posloupností.

Logickou podmínkou je to, aby počáteční a koncové vektory obou posloupností byly vzájemně na sebe zarovnány. To si vyžaduje využití algoritmu detekce začátku a konce promluvy, což může být samo o sobě poměrně složitou úlohou. Samozřejmě existují modifikace, které tuto striktní podmínku přesně nevyžadují.

Ve stručnosti se tato metoda snaží najít mapování mezi oběma sekvencemi tak, aby každý vektor z jedné posloupnosti měl přiřazen jeden vektor z druhé posloupnosti. To vede k závěru, že některé vektory jedné posloupnosti mohou být namapovány vícekrát na různé vektory druhé posloupnosti a naopak. Samozřejmě se tento proces musí řídit jistými základními logickými omezeními, zejména počáteční a koncové vektory jedné postupnosti musí být namapovány na svoje partnerské vektory druhé posloupnosti, ohýbací funkce nesmí být klesající (tj. nedojde ke změně pořadí vektorů) a je stanoven maximální rozdíl v jejich poloze, který ještě může být překlenut. Při výpočtu **DTW** se využívají dvě matice (matice lokálních vzdáleností a matice globálních vzdáleností). V matici lokálních vzdáleností se nacházejí akustické vzdálenosti mezi jednotlivými vektory příznaků referenční a neznámé promluvy. Globální matice vzdáleností slouží pro výpočet akumulované minimální vzdálenosti a nalezení cesty mezi oběma posloupnostmi. Tedy pro každý element globální matice vzdáleností existuje minimální vzdálenost a cesta spojující tento bod se začátkem matice (počáteční vektory jedné a druhé posloupnosti), který se nachází vlevo dole. Tato situace je graficky znázorněna na obr. 3.6. Samozřejmě

jsou stanovena přirozená pravidla, jakým způsobem se lze dostat z jednoho bodu matice globálních vzdáleností do druhého, např. cesta nesmí klesat. Pokud se proces dostane do konce (pravý horní roh), porovnání skončilo a nejmenší vzdálenost byla nalezena. Zpětným sledováním lze zjistit, jak vypadá minimální cesta. Tento proces je prováděn pro každé slovo ze slovníku a to, které má nejmenší vzdálenost s neznámým je prohlášeno za rozeznanou promluvu. Jak je z výkladu zřejmé, tato metoda je přímo předurčena pro úlohy rozeznávání izolovaných slov nejlépe závislých na mluvčím.



Obrázek 3.6. Matice globálních vzdáleností, optimální cesta a globální omezení na cestu

Metoda **DTW** měla významné postavení v oblasti rozeznávání řeči zejména v problematice izolovaných slov závislých na mluvčím. Jak se však požadavky zvyšovaly (nezávislost na mluvčím, plynulá řeč) ztrácela postupně svoje významné postavení a byla nahrazena metodou **HMM**.

Skryté Markovovské modely (HMM)

Skryté Markovovské modely patří do statistických modelovacích technik většinou používaných v oblasti rozeznávání řeči. Tyto metody z principu své konstrukce zahrnují jednoduše požadavky nezávislosti na mluvčím, stejně jako lehké spojení jednotlivých modelů (např. hlásek) do větších celků popisujících plynulou řeč. Navíc je tento princip poměrně snadno matematicky popsatelný.

Pro každou zvolenou hovorovou jednotku (fonému, slabiku, slovo,…) je vytvořen **HMM** model vhodné struktury. Obyčejně mají modely základních částí jazyka,

jako jsou fonémy, srovnatelnou strukturu a liší se jen nastavením svých volných parametrů. V procesu trénování se tyto volné parametry nastaví tak, aby co nejlépe popisovaly vzorky z tréninkové databáze řeči, která musí být dostatečně pestrá a velká, aby pokryla všechny statisticky významné variace řeči. Jednotlivé promluvy jsou popsané, tedy je zřejmé, jaké posloupnosti slov se v nich nachází. Nejčastěji se modely nastavují tak, aby tréninkové vzorky dosahovaly na příslušných modelech co nejvyšší pravděpodobnosti. Tento postup je nazýván jako trénování pomocí maximální věrohodnosti. Používají se i jiná kritéria jako např. co největší separace (odstup) mezi správnými a nesprávnými modely, minimalizace chyby, (takové postupy jsou označovány jako diskriminativní trénování) atd., které mohou dosahovat ještě lepších výsledků.

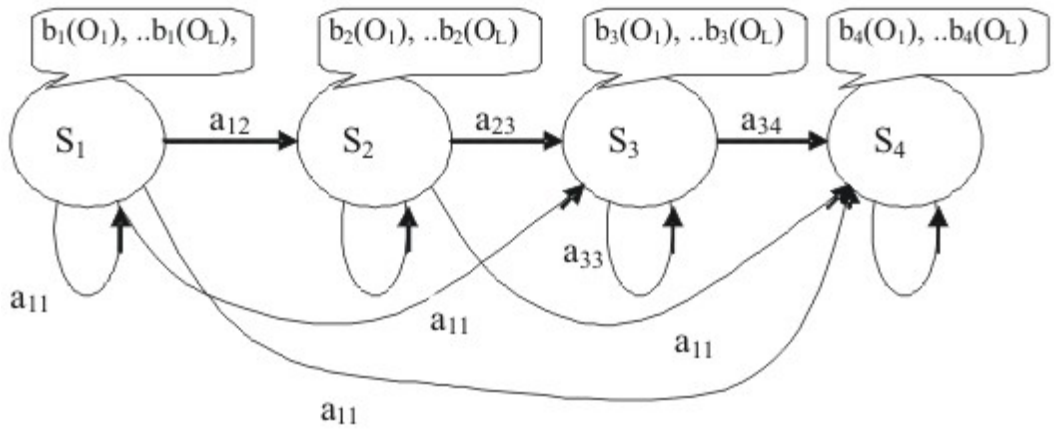
Každý model se skládá z několika stavů, které jsou navzájem vhodně propojeny. Každému spojení je přidělena tzv. přechodová pravděpodobnost *(p)*. Dále existuje tzv. počáteční pravděpodobnost *(π)*, kdy se model nachází na začátku právě v daném stavu. Potom lze tímto modelem popsat pravděpodobnost vzniku sekvence stavů $S_1, S_2, S_3, \ldots S_N$ pro daný model, a to následovně:

$$P(S_1, S_2, ..., S_{N-1}, S_N) = \pi(s_1) p_{12} ... p_{(N-1)N}$$

Nejvýše je daná dodatečná pravděpodobnost přináležející ke každému stavu, která popisuje pravděpodobnost výskytu vektoru příznaků X v daném stavu S, tedy $P(X/S)$. Touto pravděpodobností je dáno propojení mezi stavy a vektory řečových příznaků. Potom pravděpodobnost pozorování sekvence příznaků $X_1, \ldots, X_M$ a stavů $S_1, S_2, \ldots S_N$ lze jednoduše vyjádřit vztahem:

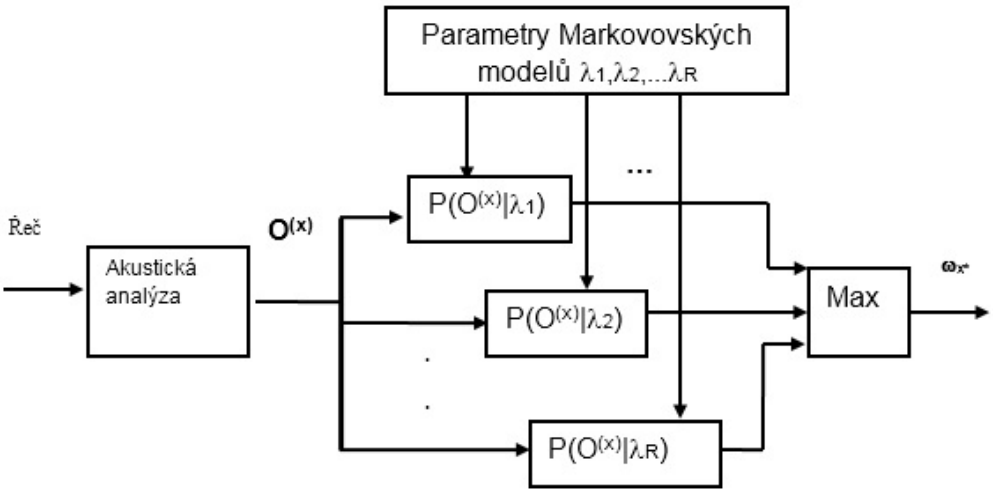
$$P(S_1, X_1, S_2, X_2, ..., S_N, X_N) =$$
$$= \pi_1 P(X_1 / S_1) p_{12} P(X_2 / S_2) p_{23} ... P(X_{N-1} / S_{N-1}) p_{N-1,N} P(X_N / S_N)$$

Příklad 4-stavového levo-pravého **HMM** modelu je zobrazen na obr. 3.7.



Obrázek 3.7. 4-stavový levo-pravý HMM model

Následně se v procesu rozeznávání počítají pravděpodobnosti neznámé sekvence příznaků na všech **HMM** modelech slov ve slovníku (vzniknou zřetězením základních **HMM** modelů) a vybere se ten model (posloupnost modelů), který vykazoval největší pravděpodobnost. Tento proces je schematicky znázorněn na obr. 3.8.



Obrázek 3.8. Proces rozeznávání řeči pomocí HMM modelu

V současnosti jsou vyspělé systémy vykazující nejnižší slovní chybovost **WER** (*Word Error Rate*) založeny právě na principu metody **HMM**, a dále využívají diskriminativní strategie trénování, jako jsou maximální vzájemná informace **MMI** (*Maximum Mutual Information*), minimální klasifikační chyba **MCE** (*Minimum Classification Error*), atd. Nakonec se využívají hybridní spojení **HMM** systémů a dalších klasifikačních technik jakými jsou zejména systémy s podpůrnými vektory a neuronové sítě (hluboké). Ty v současnosti představují špičku v oblasti rozeznávání řeči a dosahují nejnižších hodnot **WER** i pro ty nejsložitější typy úloh.

Podrobnější informace ohledně rozeznávání řeči a používaných metodách najdete např. v [7].

4 Navigace pomocí gest

V současné době představují gesta velmi populární a oblíbený způsob ovládání aplikací a většina lidí je využívá každý den. Gesta používáme u mobilních telefonů, v počítačových aplikacích, na herních konzolích, atd. Lze očekávat, že za pár let bude interakce pomocí gest k dispozici téměř v každém zařízení. Vysoká popularita navigace pomocí gest motivuje výzkumníky ve snaze o další vylepšení a zdokonalení této technologie. Jde o vysoce populární trend, a jelikož výkon počítače již není překážkou v implementaci přirozenější navigace a ovládaní pomocí gest, věnuje se mu patřičně velká pozornost [2].

V oblasti rozeznávání gest existují dvě hlavní oblasti, které rozlišujeme podle vstupního hardwaru. První oblast je zaměřená na gesta určená pro dotyková zařízení, jako jsou tablety, touchpady nebo smartphony. Druhá oblast se zaměřuje na gesta realizovaná před dvourozměrnou nebo stereoskopickou kamerou. S menšími úpravami mohou být srovnatelné algoritmy aplikované na rozeznání gest v obou oblastech. Navíc algoritmy pro stereoskopické kamery mohou být upraveny i na využití 3D souřadnic pro monitorování ve všech třech směrech [2].

4.1 Klasifikace gest dle různých kritérií

Gesta je možné dle uživatelských zkušeností rozdělit do dvou základních kategorií. Vrozená gesta, která jsou založena na všeobecné zkušenosti všech uživatelů, jako je například přesun objektu vpravo pohybem ruky doprava, uchopení objektu znázorněný piktogramem s uzavřenými prsty apod. Samozřejmě, vrozená gesta mohou být ovlivněna zvyky a kulturou. To je potřeba zohlednit při vývoji pro konkrétní skupinu lidí. Použití těchto gest nevyžaduje potřebu uživatele se tato gesta učit, stačí mu na to jeho vrozené zkušenosti. Uživatel tedy může ovládat aplikaci přirozeně. Druhou skupinou gest jsou naučená gesta. V tomto případě je nutné, aby se je uživatel naučil a osvojil si je. Obecně platí, že gesta využívaná pro navigaci systémů a aplikací musí být jednoduchá, přirozená a měly by využívat jen minimální množství lidské energie.

Statická gesta nejsou závislá na pohybu. Statická gesta představují tvary končetin znázorňující gesta, která nesou smysluplnou informaci (viz obr. 4.1).

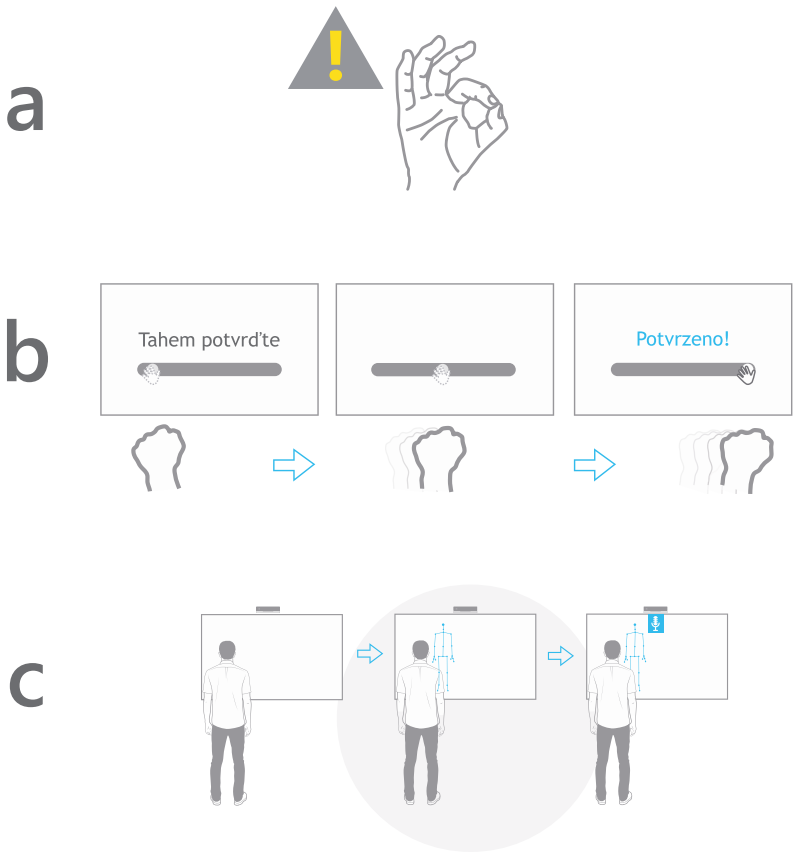
Jedná se o „nepohyblivý“ typ gest, kde nepotřebujeme informace o pohybu. Proto není potřeba zkoumat sekvenci snímků, ale skutečně jen jeden snímek. Sekvence několika snímků, které obsahují různá gesta, je důležitá jen ve vyšší úrovni porozumění.

Kontinuální (spojitá) gesta slouží jako základ pro interakci s aplikací. Kontinuální gesto je prodloužené sledování pohybu, který nepředstavuje žádnou specifickou pózu, ale používá se pro interakci s aplikací nebo ve virtuálním prostředí (viz obr. 4.1).

Typickým příkladem kontinuálních gest je ovládání hry pomocí bezdotykové technologie, jakou je například Microsoft Kinect. Systém mapuje změny v postoji nebo póze ve vztahu ke změnám ve videohře, ale přitom zde není žádný specifický pohyb vyvolávající sled změn.

Dynamická gesta představují pohyb, který umožňuje uživatelům přímo manipulovat s objektem nebo ovládat aplikaci (viz obr. 4.1).

Dynamická gesta mohou být definovaná dvěma způsoby. Prvním způsobem je algoritmicky nastavit pravidla a podmínky, které musí vykonané gesto splňovat, aby bylo úspěšně rozeznáno. Druhým způsobem je použití šablon. Šablony jsou množiny bodů, které identifikují tvar gesta. Na vyhodnocení úspešnosti jsou využity sofistikované algoritmy, které porovnávají pohyby uživatele se šablonami.



Obrázek 4.1. Typy gest: a) statické gesto, b) dynamické gesto, c) kontinuální gesto

4.2 Získávání obrazových dat

Na začátku byl výzkum algoritmů rozeznávání gest založený na vstupu z **RGB** (*Red Green Blue*) kamery. Tento přístup přináší některé výhody, které však jsou převáženy řadou nevýhod. **RGB** obraz se v principu skládá z barevných kanálů a existuje mnoho faktorů, které negativně ovlivňují kvalitu **RGB** obrazu.

Např. změna osvětlení může radikálně ovlivnit kvalitu obrazu. Změna podmínek osvětlení v místnosti totiž zcela zásadně mění jas a kontrast obrazu. Tím se stává rozeznávání pomocí **RGB** obrazu nejisté a méně přesné [2].

V posledních několika letech začali vědci používat tzv. hloubkový obraz pro další zpracování v oblasti rozeznávání gest. Hloubkový obraz je dvojrozměrný obraz, který obsahuje další dodatečné informace – hloubku. Většina hloubkových senzorů je založena na infračerveném **IR** (*Infra Red*) záření (např. senzor Kinect – viz obr. 4.2). Senzor vrací data, která představují vzdálenost každého pixelu na snímku od snímače (senzoru). Získané vzdálenosti mohou být jednoduše přepočítány na odstíny šedé barvy, a tak jsou hloubková data reprezentována jako černobílá videosekvence. Získání dat nevyžaduje žádné zvláštní podmínky.

Hloubková data mají víc informací než data získaná **RGB** kamerami. Hlavní výhodou je, že hloubková kamera využívá infračervené světlo a v infračervené zóně se nevyskytuje tolik šumu a rušení. Takže infračervený snímač není tolik citlivý na kvalitu scény jako **RGB** snímače [2].

Obrázek 4.2. Obraz z hloubkové kamery

Vzdálenost každého pixelu ze snímače je definovaná v milimetrech.

Pokud chceme převést hloubkový obraz do obrazu v odstínech šedé barvy, musíme znát minimální a maximální možnou vzdálenost.

$$\text{int } enzita = 255 - \frac{255 - (d - d_{\min})}{d_{\max}}$$

kde d je aktuální vzdálenost daného pixelu, d_{min} je minimální možná vzdálenost od senzoru a d_{max} je maximální možná vzdálenost od senzoru [2].

Jako pomůcka při získávání vstupního obrazu mohou sloužit extra prvky, jako např. světélkující náramek, prsten anebo malý míček v ruce.

4.3 Metodika návrhu gest

Problematika rozeznávání gest není zcela triviální (jednoduchá). Všechna omezení a nepřesnosti, které se vyskytují při získávání vstupních obrázků pomocí různých zařízení, je nutné vzít v úvahu při procesu navrhování gest pro konkrétní aplikaci nebo kontrolu systému. Metodika pro návrh gest musí splňovat několik požadavků: přirozenost, lehkou zapamatovatelnost a komfort. Komfort je nejdůležitějším kritériem pro aplikace řízené pomocí gest.

Stejně tak jako v každé nové oblasti, tak i v oblasti ovládání gesty je jedním z hlavních problémů v procesu adaptace neexistující standardizace. Pokud vybereme zcela náhodně deset lidí a požádáme je, aby udělali gesto pro spuštění hudby, je vysoce pravděpodobné, že bude navrženo minimálně 5 různých gest. To je důvod, proč každý výrobce poskytující ovládání gesty používá odlišnou sadu gest. Z praktického hlediska je to velmi komplikovaný fakt, protože uživatel se musí učit novým gestům pro každý nový systém. Existence alespoň jedné standardní sady gest pro ovládání základních funkcí by mohla významně urychlit proces domestikace ovládání pomocí gest i v běžné praxi. V dnešní době jsou již k dispozici televizní přijímače využívající ovládání gesty. Jsou založeny na snímání obrazu 2D kamerou. Umožňují jednoduchou navigaci v menu televizoru. I přes své nedokonalosti má zavádění gest do praxe výchovný vliv na lidi, kteří si postupně zvykají na tento druh ovládání.

Současné sady gest jsou založeny na vstupních periferních zařízeních (např. klávesnice, myš,…) používaných u počítačů. Gesta se ve své podstatě snaží odstranit vstupní zařízení, ale zachovávají srovnatelné způsoby využití. Cílem návrhu gest od základu, tedy bez ohledu na existující uživatelská rozhraní, je co nejvíc se přiblížit přirozenému bezdotykovému ovládání tak, jakoby gesta byla jediným vstupním zařízením.

4.4 Techniky pro rozeznávání základních gest

Hlavním cílem při rozeznávání gesta ruky je určit následující: umístění ruky, držení nebo póza ruky nebo vyjádření gesta.

Rozeznávání gest může být realizováno dvěma způsoby. Buď použitím datové rukavice, která převádí informaci o pohybu končetiny do elektronické podoby, nebo je možné využít přístup založený na pozorování, tj. kamera slouží jako lidské oko zaznamenávající polohy těla, které se následně extrahují při zpracování obrazu [5].

Je zřejmé, že první způsob přináší přesnější výsledky, jelikož se jedná o přímé a přesné pozorování skutečného stavu příslušné části lidského těla, na které se nachází oblečení typu datová rukavice.

Na druhé straně, přístup založený na pozorování neklade žádné další požadavky na zařízení pro koncového uživatele (s výjimkou kamery), takže je tento přístup vhodný pro všeobecné použití [5].

Použití datové rukavice je poměrně nepohodlné, pokud jde o komfort uživatele. Vybavení potřebné pro využívání této metody by bylo nepřijatelně nákladné pro většinu běžných zákazníků, takže je uplatnitelná pouze pro speciální účely (např. digitalizace pohybu pro animované filmy, tvorba počítačových her apod.).

Nevýhodou metody založené na pozorování spočívá v jeho algoritmické složitosti, kdy je potřebné značné množství času a výpočetního výkonu pro extrakci a extrapolaci pohybu těla [5].

Existují však rozličné algoritmy, které se zaměřují na různé projevy gestikulující osoby (a vycházejí tím z určitých předpokladů). Všeobecně platí, že mohou být rozděleny do dvou kategorií, a to na metody založené na vzhledu a na 3D metody založené na principu analýzy modelu. 3D metody založené na principu analýzy modelu porovnávají vstupní parametry končetiny s 2D projekcí trojrozměrného modelu končetiny. Přístup založený na vzhledu využívá obrazové příznaky pro modelování vizuálního vzhledu končetiny a porovnává jej s extrahovanými příznaky ze vstupního videa [5].

V kapitole 4.1 byly definovány tři typy gest. Pro detekci statických gest (tj. pózy) se používá všeobecný klasifikátor. Pod pojmem klasifikátor rozumějme – prvek nebo vlastnost, na základě kterého rozhodneme o zařazení do skupiny nebo kategorie. Takže klasifikátor nás informuje o tom, které gesto (nebo póza) bylo rozpoznané. Avšak u dynamických gest je třeba zohlednit i časové hledisko a proto rozeznání vyžaduje techniky, které zpracují i tuto dimenzi, např. skryté Markovovské modely **HMM**. Jiný způsob je pomocí pohybových modelů.

Některé z technik používaných pro statické (a dynamické) rozeznání gest jsou shluková analýza, algoritmus **KNN** (*k-Nearest Neighbors*), metoda strojového učení **SVM** (*Support Vector Machines*), a už vzpomínaný **HMM** algoritmus, **DTW** algoritmus či neuronové sítě [3].

Techniky používané pro rozeznání statických gest

Pro rychlejší zpracování dat se v procesu rozeznávání gest nepracuje s celým obrazem.

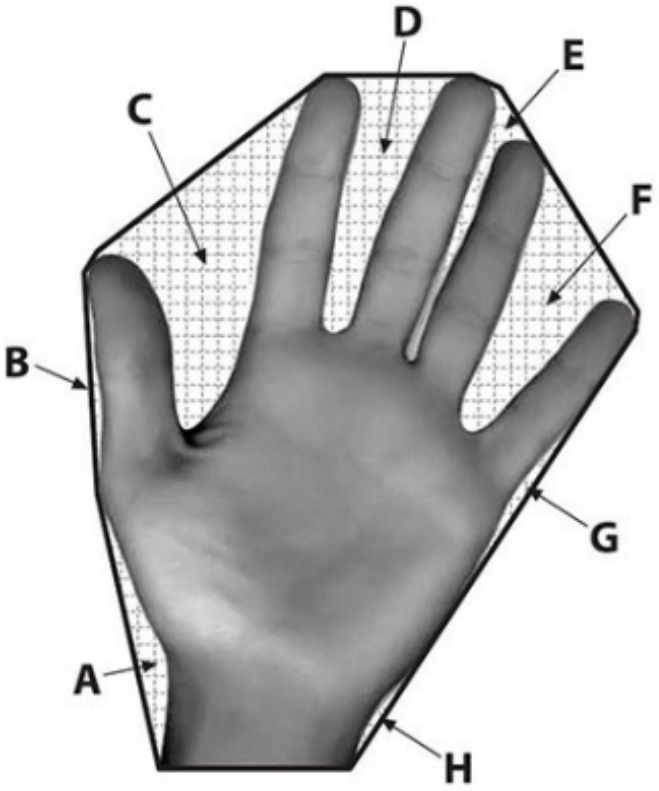
Poté, co je automaticky detekována ruka, označí se oblast okolo ní a dále se zpracovává jen ona. Tím se sníží výpočetní náročnost. Získané vzdálenosti se převedou na odstíny šedé a získáme tak obrys ruky.

Následuje popis dvou metod používaných v procesu rozeznávání.

Metoda konvexních nedostatků

V prvním kroku je nutné oddělit ruku od pozadí obrazu. Oddělení je možné udělat z hloubkového obrazu, kde lze určit, který pixel patří ruce a který ne. Druhým krokem je určení obrysu (tzv. kontury) ruky.

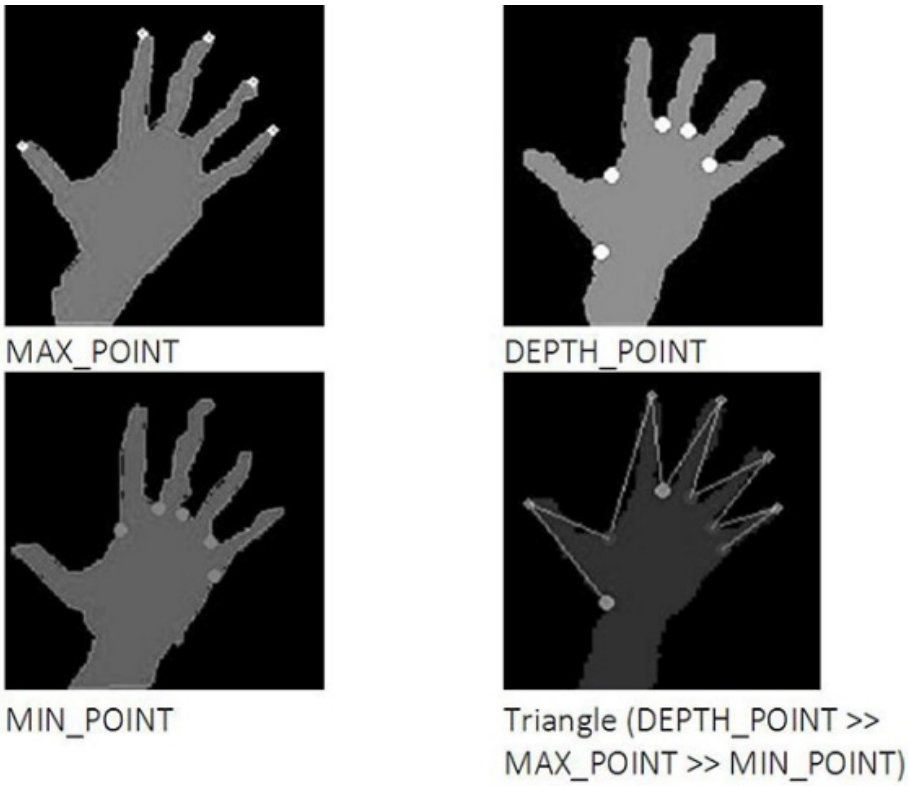
Tvary mnohých složitých objektů jsou dobře charakterizované pomocí tzv. konvexních nedostatků. Obrázek 4.3 znázorňuje koncept konvexních defektů na obrázku ruky. Konvexní obal je zobrazen jako tmavá čára okolo ruky, oblasti označené A až H jsou „nedostatky“ patřící k tomuto obalu. Jak je vidět, tyto konvexní nedostatky charakterizují též stav ruky. Algoritmus vrátí souřadnice tří bodů, *počáteční bod*, *nejhlubší bod* a *koncový bod* (viz Obr. 4.4). Nejhlubší bod je možné chápat jako maximální vzdálenost mezi obalem a obrysem ruky.



Obrázek 4.3. Konvexní obal vykreslený jako tmavá čára okolo ruky

Obrázek 4.4 Označená oblast jsou konvexní nedostatky a tři body: počáteční bod (vlevo), nejhlubší bod (dole) a koncový bod (nahoře)

Cílem algoritmu je nalezení bodu na prstě, který je nejdále od středu ruky. Prvním krokem je odstranit všechny nedostatky, jejichž výška je menší než zadaná hodnota. Potom se odstraní nedostatky, které mají vzdálenost mezi počátečním bodem a maximálním bodem větší než zadaná hodnota, která je definována dynamicky (pokud jsou totiž dva body příliš daleko, znamená to, že nemohou představovat prst). Dále jsou také odstraněny nedostatky, které mají vzdálenost mezi počátečním a hloubkovým bodem menší, než je stanovená hodnota. Tato hodnota se mění dynamicky podle velikosti oblasti, kde je detekována ruka (viz Obr. 4.5). Posledním krokem je odstranění všech nedostatků, které se vyskytují pod zápěstím.



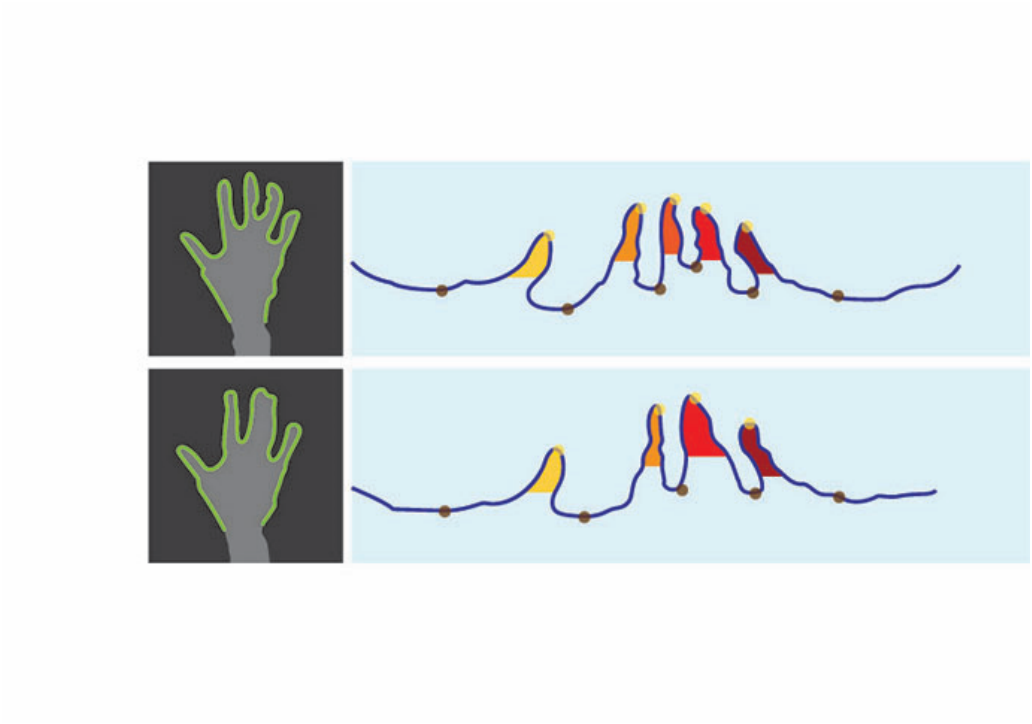
Obrázek 4.5. Znázornění logiky používající trojúhelníky na minimalizování potřebných nedostatků

Rozeznávání gesta ruky založené na částech

V prvním kroku tohoto algoritmu, se extrahují části údajů (datové segmenty) obsahující ruku a převedou se na binární obraz. Střed dlaně (Obr. 4.6) se vypočítá z vnitřního kruhu a přidáním bodu na obal kontury, která patří ruce a má maximální vzdálenost od zjištěného nedostatku.

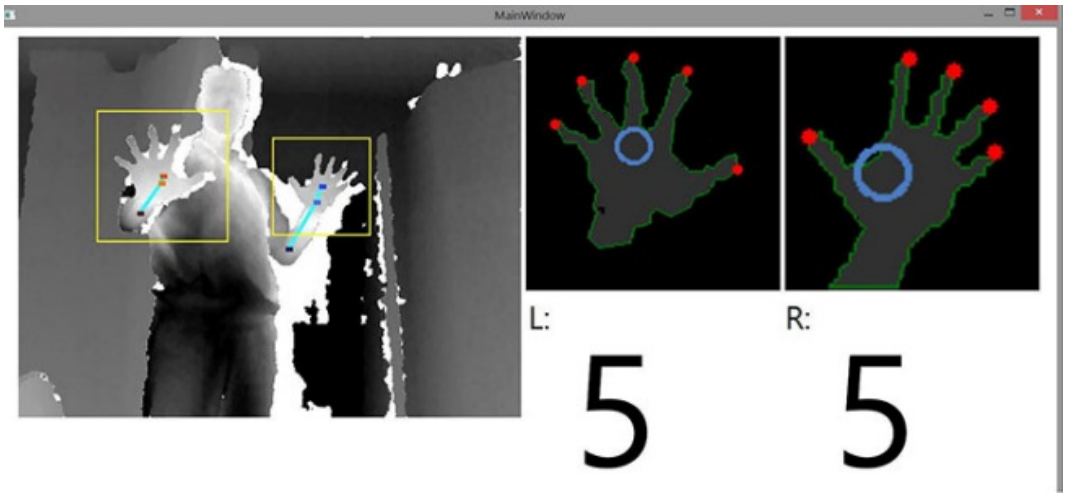
Obrázek 4.6. Hledání středu dlaně

Všechny body z kontury ruky jsou mapované na osu X. Y-ova osa potom opisuje relativní vzdálenost jednotlivých bodů od středu dlaně. Všechny mapované body vytvoří křivku (Obr. 4.7).



Obrázek 4.7. Křivková reprezentace

Dalším krokem algoritmu je analýza křivky s cílem najít její lokální maximum. Potom, co je dokončena analýza křivky a nalezeno maximum, začnou se hledat prsty. Vzdálenost každého maxima je porovnaná s nastavenou prahovou hodnotou. Každý prst má specifickou váhu a na jejím základě a na základě vztahů mezi prsty jsou prsty rozeznané (vztah je porovnání váhy s danými hodnotami a na základě toho je znám počet prstů, tedy pokud váha < 1,5× průměrná váha – segment obsahuje jeden prst, atd.). Tento algoritmus funguje, i pokud jsou prsty spojené (Obr. 4.8).



Obrázek 4.8. Detekované prsty

Techniky používané pro rozeznání dynamických gest

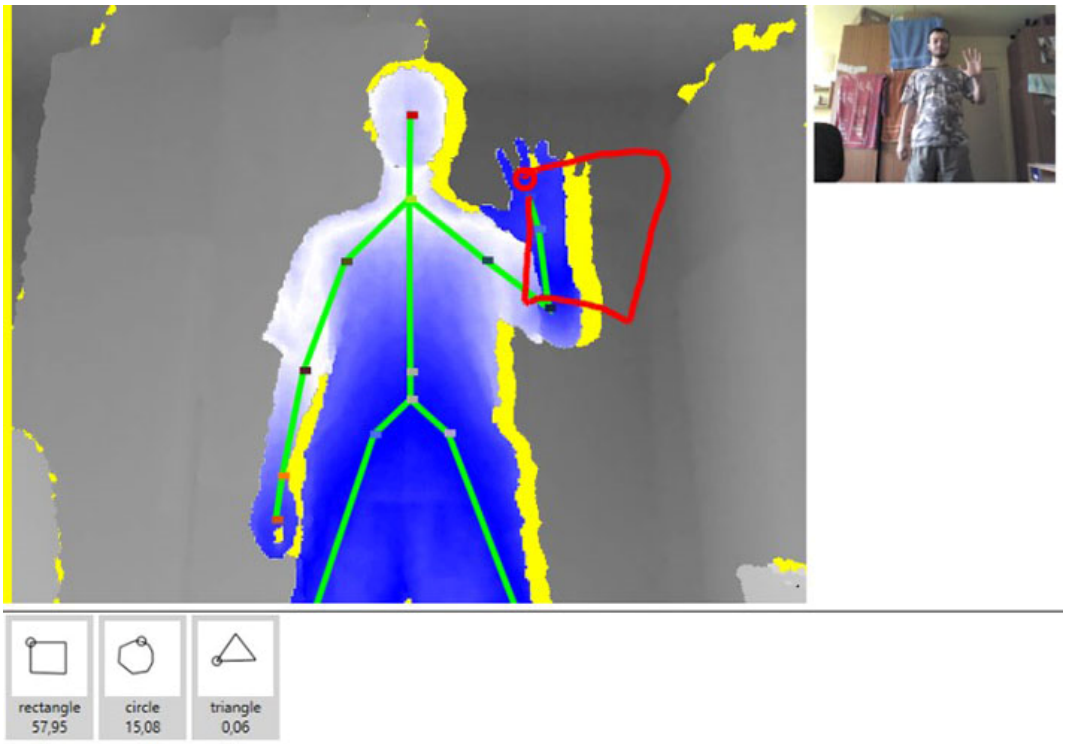
Dynamická gesta zajišťují přístup uživatelů k soukromému obsahu nebo mu umožňují ovládání systému nebo aplikace. Je možné je použít i namísto hesla. V počátcích rozeznávání gest byly většinou používány neuronové sítě a genetické algoritmy.

Tyto metody měly přijatelnou úspěšnost rozeznávání, ale největší nevýhodou byl požadovaný výpočetní výkon a čas potřebný na trénování neuronových sítí. Oba tyto aspekty byly nepřijatelně vysoké pro užití v praktických aplikacích.

V dnešní době se používají na rozeznání gest nové techniky. Byly vyvinuty algoritmy, které nevyžadují neuronové sítě, například vyhledávání v zlatém řezu (*Golden Section Search*), algoritmus přírůstkového rozeznávání (*Incremental Recognition Algorithm*) a pravděpodobnostní modely jakými jsou skryté Markovovské modely **HMM**. Pro zvýšení úspěšnosti již zmíněných algoritmů se může použít strojové učení (tzv. *Machine Learning*). V posledních letech jsou velmi populární **HMM** metody. Hlavním důvodem je to, že **HMM** přístup je velmi dobře známý a užívaný v mnohých oblastech.

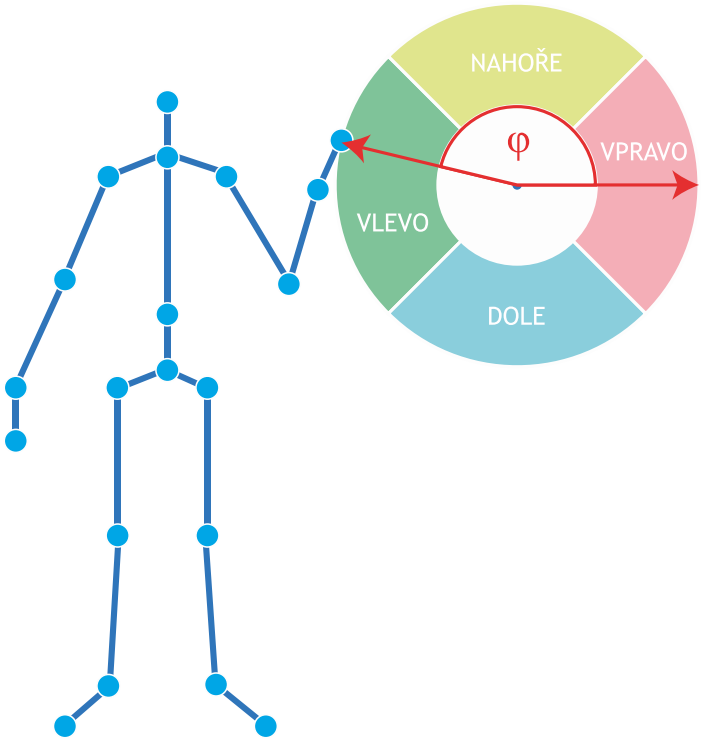
Algoritmus navržený Kristenssonem a Denbym [4], původně navržený pro rozeznávání tahů digitálním perem a na dotykovém displeji je možné rozšířit i na dynamická gesta.

V tomto algoritmu je definována šablona jako soubor segmentů popisujících dané gesto. Je zřejmé, že rozeznání gesta z dané sady gest, kde jsou gesta dostatečně vzájemně odlišitelná, může být úspěšné pouze v případě, pokud byla vykonána dostatečně velká část z daného gesta (Obr. 4.9).



Obrázek 4.9. Náhled testované aplikace. Červená čára znázorňuje vykonané gesto. Šablony pod obrazem představují nejpravděpodobněji vykonaná gesta (zleva doprava)

Používání gest se postupně rozšířilo o tahová gesta. Tento typ gest přináší velmi přirozený a pohodlný způsob ovládání především pro koncového uživatele. Tahová gesta jsou určena pro rychlé a zběžné procházení menu, programů a galerií a umožňují 4 směry pohybu pro každou ruku a několik kombinací obouruč. Metoda kruhového rozeznávání dynamických gest **CDGR** (*Circle Dynamic Gesture Recognition*) publikovaná v [1], je založena na detekci ruky, rychlosti pohybu a vzdálenosti. Pokud ruka vykoná rychlejší pohyb a vnitřní kruh projde do vnějšího kruhu, systém zpracuje tento pohyb a určí typ gesta. Gesto je dáno úhlem vykonaného pohybu od středu do zóny vnějšího kruhu. Možná gesta jsou: přejít tahem vlevo, vpravo, nahoru a dolů. To stejné gesto může být vykonáno i oběma rukama. Uživatel může též vykonat gesto přiblížení a oddálení (Obr. 4.10).



Obrázek 4.10. Kruhové rozeznávání dynamických gest CDGR

4.5 Použití gest v praxi

Rozeznávání gest může být interpretováno jako jeden z prvních možných způsobů, jak počítače porozumějí člověku nebo řeči lidského těla. To může onen pověstný velký krok, kdy se vstupní zařízení jako klávesnice a myš stanou nepotřebnými (však se s nimi také nepočítá ani v žádných sci-fi filmech). Zavedením rozeznávání gest bude možné např. ukázat prstem na obrazovku počítače tak, že se kurzor bude podle toho pohybovat.

Navigace pomocí gest je stále víc a víc populární především ve smartphonech, tabletech a moderních televizorech. Každý zná tahové gesto pro posun nahoru/dolu nebo přesunout doleva/doprava, zvětšování nebo zmenšování (*Zoom*). Existuje bohatá nabídka aplikací poskytujících sadu gest pro navigaci na dotykových zařízeních, např. pro ovládání hlasitosti, uzamknutí obrazovky, ovládání přehrávače hudby, kopírování obsahu obrazovky, návratu na domovskou obrazovku, krok zpět, návrat k posledním aplikacím či vstup do menu.

Oblast pro využití gest není prozatím omezená. Jako příklad můžeme uvést aplikaci na trénování tenisu využitím 3D rozeznávání gest, kterou publikoval *Cristian García Bauza* a jeho tým.

Například fa. Samsung představila rozeznávání gest u svého nejvyššího **TV** modelu, tj. u své „vlajkové lodi“, již v roce 2013. Rozeznávání je založeno na algoritmu sledování ruky s rozlišením gesta „*klikněte na tlačítko*“, které je realizováno zavřením dlaně. Aby uživatel mohl přepínat kanály nebo nastavit hlasitost, musel nejprve zdvihnout ruku k televizoru a po objevení kurzoru musel pohybovat rukou tak, aby se kurzor posunul k tlačítkům (analogie k počítačové myši). Pro zvýšení hlasitosti musel uživatel „*klikat na tlačítko*“, tj. zavírat dlaň s rukou tolikrát, než byla dosažena požadovaná hlasitost. Tento přístup by mohl být považován za intuitivní aspoň u počítačově gramotných uživatelů, ale spíše to zde slouží jako příklad nepřirozených gest.

5 Grafické rozhraní – návrh

Grafické uživatelské rozhraní **GUI** (*Graphical User Interface*) představuje způsob interakce mezi lidmi a stroji/počítači. Uživatelské rozhraní zahrnuje hardwarové (fyzické) a softwarové (logické) komponenty.

Uživatel odešle příkaz nebo úlohu do počítače/systému a počítač reaguje zpět plněním požadovaného příkazu nebo úlohy. V minulosti **GUI** přijímalo vstup prostřednictvím zařízení, jako jsou například počítačová klávesnice a myš. Jak je vidět s příchodem dotykových displejů, **GUI** se přizpůsobilo a akceptuje i vstup realizovaný dotykem prstů nebo pera (stylus).

Vytvořit vhodné **GUI** pro multimodální (ale i jiné) aplikace vyžaduje dobrou praxi, zkušenosti a zpětnou vazbu uživatele (v současné době se tato oblast výzkumu nazývá míra spokojenosti uživatele **UX** (*User eXperience*)). **GUI** pro multimodální aplikace by mělo umožnit jejich jednoduché, efektivní a uživatelsky komfortní ovládání. V ideálním případě uživatel potřebuje minimální úsilí pro nastavení vstupu a zařízení vyžaduje minimální provozní čas na zpracování výstupu. Blízkým tématem k míře spokojenosti uživatele **UX** je proces zlepšování míry spokojenosti uživatele **UXD** (*User eXperience Design*).

6 MMI aplikace

Moderní aplikační výzkum je zaměřen na vytvoření co nejpřirozenějšího multimodálního rozhraní a jeho začlenění do multimediálního systému již využívaného v každodenním životě. Multimodální rozhraní není omezeno pouze na televizi, ale má využití i v různých aplikacích. Vize budoucích aplikací a systémů není jen ovládat TV a mít přístup k multimediálnímu obsahu, ale budovat systém, který rozezná zvyky svých uživatelů, jejich preference a zájmy. Takový systém se pak může stát inteligentní součástí domácnosti, a může poskytovat pokročilé nástroje pro různé služby (např. nákupy při sledování TV, e-volby nebo digitální vrátný).

Z hlediska co nejefektivnějšího využití multimodálního rozhraní je důležité vědět, která modalita je svázána s konkrétní aktivitou, která se má vykonat. Pro různé druhy kroků, resp. aktivit jsou vhodné různé druhy modalit, například zadávání textu by bylo zbytečně obtížné, časově náročné a nepohodlné pomocí gest. Mnohem lépe by se zde uplatnilo rozeznávání řeči. S tímto vědomím je nevyhnutelné pro multimodální systém využít integrovanou platformu, která bude poskytovat požadované údaje dílčím aplikacím. Tyto aplikace potom nepotřebují znát zdroj modality.

Jednoduchý modelový příklad:

MMI aplikace pro ovládání **TV** zobrazí uživatele, kteří byli identifikováni prostřednictvím hlasu nebo rozeznáním tváře pomocí příslušného modulu. Pouze rozeznaní uživatelé v tomto seznamu mají povoleno ovládat **TV** pomocí gest, hlasem nebo jinou modalitou. Pokud uživatel opustí místnost, je ze seznamu automaticky odstraněn.

Ostatní části aplikace zobrazují dostupné **TV** kanály. V závislosti na společných preferencích rozeznaných lidí v místnosti dokáže systém přizpůsobit nabídku těchto kanálů. Pomocí tahových gest je také možné přehrát položku ze seznamu, případně přehrávání zastavit. Některé z dostupných kanálů mohou být i zamknuté, což znamená, že uživatelé bez oprávnění si je nemohou přehrát bez zadání tajného vzoru. Tajný vzor představuje dynamické gesto.