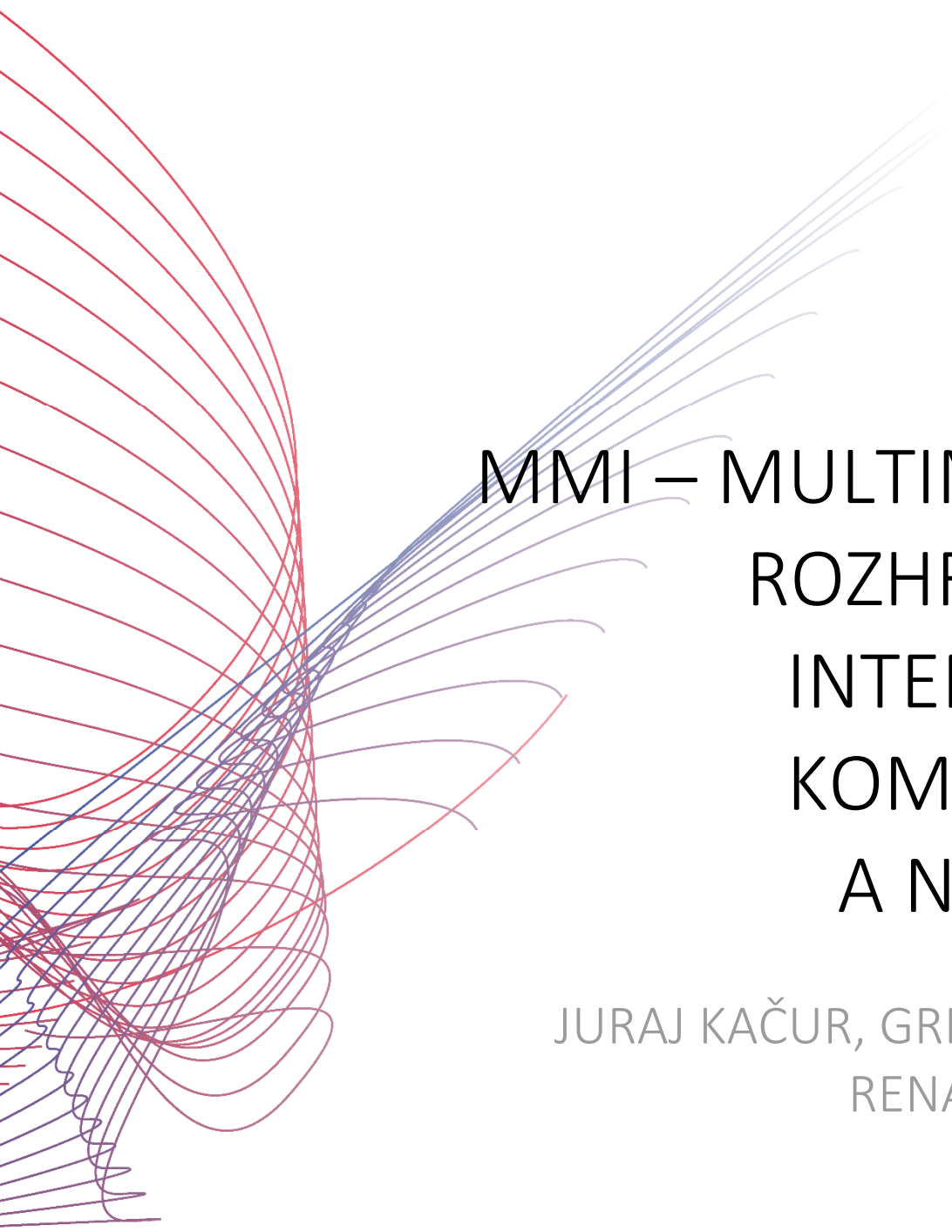
en cs de sk es

TECH pedia



MMI – MULTIMODÁLNE ROZHRANIE PRE INTERAKTÍVNU KOMUNIKÁCIU A NAVIGÁCIU

JURAJ KAČUR, GREGOR ROZINAJ,
RENATA RYBÁROVÁ

Názov: MMI – multimodálne rozhranie pre interaktívnu komunikáciu a navigáciu

Autor: Juraj Kačur, Gregor Rozinaj, Renata Rybárová

Vydalo: České vysoké učení technické v Praze

Fakulta elektrotechnická

Kontaktná adresa: Technická 2, Praha 6, Česká republika

Tel.: +420 224352084

Tlač: (iba elektronická)

Počet strán: 40

Edícia (vydanie): 1. vydanie, 2017

ISBN 978-80-01-06245-6

TechPedia
European Virtual Learning Platform for Electrical and Information Engineering

http://www.techpedia.eu

Erasmus+

Tento projekt bol financovaný s podporou Európskej Komisie.
Táto publikácia (dokument) reprezentuje výlučne názor autora a Komisia nezodpovedá za akékoľvek použitie informácií obsiahnutých v tejto publikácii (dokumente).

VYSVETLIVKY

	Definícia
	Zaujímavosť
	Poznámka
	Príklad
	Zhrnutie
	Výhody
	Nevýhody

ANOTÁCIA

Multimodálne rozhranie (angl. Multimodal interface - MMI) pozostáva z viacerých tém a modulov, ktoré slúžia pre prirodzenú a intuitívnu komunikáciu so systémom. Všetky dokopy predstavujú funkcionalitu MMI. Základný zoznam modulov, ktoré môžu byť súčasťou multimodálneho rozhrania môžu byť rozpoznanie rečníka, rozpoznanie reči a hlasových povelov, rozpoznanie tvárí, navigácie pomocou gest, navigácia pohybom očí a syntéza reči.

CIELE

Hlavným cieľom tohto výučbového kurzu je oboznámiť študentov so základmi multimodálneho rozhrania a multimodálnymi aplikáciami. Študentovi sú predstavené základné princípy rozpoznávania tváre, identifikácie hovoriaceho, navigácie pomocou hlasových povelov a gest. Koniec modulu je venovaný možnému vyžitiu opísaných modalít v multimodálnej aplikácií.

LITERATÚRA

[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

Obsah

1 MMI architektúra

V súčasnej dobe je najrozšírenejšie rozhranie slúžiace na komunikáciu človeka s počítačom klávesnica, myš alebo dotykový displej. Tieto zariadenia predstavujú adaptáciu ľudí na úroveň počítačov viac ako prirodzený spôsob komunikácie. V posledných niekoľkých rokoch sa začali objavovať požiadavky ľudí na potrebu komunikovať s počítačmi rovnakým spôsobom ako medzi sebou: rečou, mimikou a gestami. Takto dokážu ľudia preniesť oveľa viac informácií než použitím iba periférnych zariadení. To nás privádza k pojmu *multimodálne rozhranie* (**MMI**).

Multimodálne rozhranie sa skladá z niekoľkých oblastí a modulov, ktoré slúžia na prirodzenú a používateľsky prívetivú komunikáciu so systémom. Spolu tieto moduly predstavujú funkcie MMI. Súčasťou MMI môžu byť tieto moduly:

- Hlasová identifikácia
- Rozpoznanie reči a hlasových povelov
- Rozpoznanie viacerých tvárí
- Rozpoznanie gesta a navigácia pomocou gest
- Navigácia očami
- Syntéza reči
- Odporúčací modul (modul alebo systém na filtrovanie informácii, ktorý sa snaží predikovať ohodnotenie alebo preferenciu, ktorú by dal používateľ danej položke)

Všeobecne sa architektúra multimodálneho rozhrania skladá z niekoľkých vrstiev. Fyzická vrstva predstavuje hardvérové vstupné a výstupné zariadenia, ktoré umožňujú interakciu s okolím. Multimodálne údaje poskytnuté vstupnými zariadeniami (kamery, senzory, mikrofón, atď.) sú spracované paralelne každým modulom zvlášť (moduly sú nad fyzickou vrstvou, táto vrstva sa nazýva vrstvou modalít) . MMI kontrolér (na vrstve MMI) zbiera výstupné dáta zo všetkých modulov, vyhodnocuje a spája ich do jedného výstupného dátového toku. Výsledný tok informácií obsahuje informácie o používateľoch a autorizácii ich požadovaných akcií. Ten sa prenesie do aplikačnej vrstvy.

2 Identifikácia používateľa

V súčasnej dobe sa informačné technológie dostávajú stále viac a viac do popredia. Stretávame sa s nimi v našom každodennom živote, počínajúc malými smart telefónmi a končiac inteligentnými budovami.

Každé zariadenie pracujúce s našimi osobnými údajmi by malo byť preto chránené pomocou nejakého zabezpečenia. Najrozšírenejší typ overovania alebo autentifikácie je meno a heslo, ale s rozvojom informačných technológií a algoritmov pre ochranu bezpečnosti, systémy a aplikácie začínajú používať autentifikáciu na základe biometrických znakov. Napríklad autentifikácia na základe rozpoznania tváre je súčasť každého Windows 10 pod názvom Windows Hello. Pri použití biometrických údajov existuje niekoľko výhod: biometrické údaje sú jedinečné pre každú osobu, takže je postačujúca prítomnosť človeka pred senzorom a nie je potrebné pamätať si niekoľko hesiel alebo pinov. Poznáme identifikačné systémy založené na niekoľkých rôznych biometrických modalitách, ako sú očné dúhovky, odtlačky prstov, hlas, tvár, atď. V tejto časti sa všeobecne zameriame na systémy pre rozpoznávanie tvárí a identifikáciu hovoriaceho.

2.1 Rozpoznanie 2D a 3D tváre

Rozpoznanie tvárí je podmnožinou veľkej oblasti výskumu a technológií zameraných na rozpoznávanie vzorov. Rozpoznávanie ľudskej tváre je jednou z najdôležitejších metód autentifikácie pomocou biometrických údajov v posledných niekoľkých desaťročiach, a to najmä vzhľadom na veľký potenciál pre širokú škálu aplikácií a oblasti možného využitia (bezpečnostný dohľad, zabezpečenie domácnosti, hraničné kontroly, atď.). Biometrické systémy pre identifikáciu osôb, ktoré sú vyvinuté niekoľkými spoločnosťami, dosahujú vysokú presnosť v rozpoznávaní tvárí. Väčšina z týchto aplikácií musí spĺňať [1]:

- Dokáže rozpoznať niekoľko tvárí z jedného video záberu alebo jedného obrázku
- Vysokú úspešnosť rozpoznávania
- Nezávislosť od svetelných podmienok
- Stabilitu pri zmene výrazu tváre alebo pózy
- Rozpoznanie v reálnom čase, atď.

Hlavnou výhodou v porovnaní s inými prístupmi je to, že rozpoznávanie tvárí nevyžaduje explicitný súhlas používateľa, pretože obrazy tvárí možno získať kamerou aj z väčšej vzdialenosti. Ďalšou výhodou je, že zariadenia na zaznamenanie tvárí (napr. kamery) sú lacné a stávajú bežnou komoditou.

Hlavnou nevýhodou rozpoznávania tvárí je relatívna jednoduchosť, ktorá môže potencionálne spôsobiť jej nepresnosť.

V porovnaní s rozpoznávaním hovoriaceho, rozpoznávanie tvárí dosahuje oveľa lepšie výsledky. Všeobecne existujú tri hlavné prístupy založené na type dát, ktoré sa používajú pri procese rozpoznávania. Patria sem metódy založené na 2D obrazoch, na údajoch získaných z 3D tvárí a techniky, ktoré používajú obidva typy dát. Celý proces rozpoznávania sa skladá z 3 hlavných fáz. Prvá z nich je získanie a predspracovanie dát, druhá je registrácia dát a tretia fáza je rozpoznanie. Detailný popis 2D a 3D metód na rozpoznávanie tváre možno nájsť v module *Identifikácia používateľa*.

V moderných systémoch možno implementovať rozpoznávanie viacerých tvárí súčasne. V tomto prípade sú detekované všetky tváre na snímkach alebo obraze. Polohy detekovaných tvárí sú zapamätané pre daný obraz. Ďalej je obraz rozdelený do niekoľkých častí (vzoriek) na základe pozície detekovaných tvárí. Tieto generované vzorky reprezentujú všetky tváre v obraze a umožnia sledovať každú tvár samostatne. Rozdelenie obrazu je hlavným krokom vedúcim k rozpoznávaniu viacerých tvárí v systéme alebo aplikácií.

2.2 Identifikácia hovoriaceho

V stručnosti povedané, cieľom identifikácie hovoriaceho je na základe hlasu určiť totožnosť hovoriaceho. Rozhodnutie sa robí na základe množiny rečových vzoriek hovoriacich, ktoré sú uložené v databáze počas fázy trénovania. Systém teda rozpozná len používateľov z databázy, resp. rozpoznanie nevykoná ak je vierohodnosť výsledku príliš nízka.

Pretože počet používateľov v databáze je konečný, táto úloha sa tiež označuje ako problém uzavretej skupiny. Identifikácii na základe hlasu sa veda venuje už viac ako 40 rokov a vývoj v tejto oblasti stále intenzívne pokračuje. S príchodom nových a dostupných technológií si identifikácia nachádza veľa nových aplikácií, z ktorých najvýznamnejšie sú:

- **Kriminalistika**
- **Automatická indexácia rečových a audio nahrávok**
- **Neinvazívna kontrola prístupu ku citlivým dátam, službám, atď.**
- **Hrací priemysel**
- **Pomôcky pre postihnutých**

Identifikácia hovoriaceho je pomerne zložitá úloha, a to vďaka viacerým dôvodom. Tie najvýznamnejšie sú:

- **Akustická variabilita hovoriaceho**
- **Zmeny vplyvom zdravotného stavu, nálady, psychického stavu, atď.**
- **Prítomnosť šumu pozadia**
- **Vplyvy nahrávacích zariadení a prostredia**
- **Jednoduchá reprodukcia zvuku pomocou nahrávacích zariadení**

Na druhej strane použitie reči ako biometrického signálu má nasledovné výhody:

- **Netreba používať heslo**
- **Je to neinvazívna metóda**
- **Je to jednoduchá metóda získavania biometrického signálu**

Je veľa aplikácií pre identifikáciu používateľa, ktoré sa medzi sebou líšia v komplexnosti, použitých metódach, konceptoch, spôsoboch činnosti, požiadavkách, obmedzeniach, atď. Preto sa zaužívali isté delenia, a tie najvýznamnejšie sú:

- **Textovo nezávislé systémy** – tieto pracujú nezávisle na tom, čo bolo povedané, teda nevyžadujú predpísané slová alebo frázy okrem dostatočnej dĺžky prehovorenia. Úspešnosť týchto systémov sa v závislosti od počtu používateľov a prostredia pohybuje v rozmedzí 80 až 95%. Tieto systémy sú najbežnejšie i keď vykazujú nižšiu úspešnosť.

- **Textovo závislé systémy** – požadujú pre svoju správnu činnosť vyhovorenie predpísaných fráz alebo hesiel. To umožňuje postihnúť aj koartikulačné efekty a prípadne využitie neznámeho hesla. Tieto systémy dosahujú preto vyššiu úspešnosť v rozmedzí od 95 do 99%. V špeciálnych prípadoch tieto systémy môžu meniť očakávaný text pomocou rôznych otázok.

- **Systémy pracujúce na akustických príznakoch** – využívajú extrahované akustické informácie reči, ako sú rôzne modifikované krátkodobé spektrá reči vystihujúce formantové frekvencie hlasového traktu. Tieto príznaky v sebe odrážajú fyzické vlastnosti hlasových orgánov ako sú ich veľkosť tuhosť, tvar atď. Ľahko sa získavajú a sú základom skoro všetkých systémov.

- **Systémy pracujúce s prozodickými informáciami** – extrahujú a ďalej spracovávajú špecifické informácie ako dynamika reči, rýchlosť rozprávania, modulácia hlasu, pauzy, prízvuk, atď. Určujú sa z dlhších intervalov reči a sú ťažšie detekovateľné a kvantifikovateľné.

- **Systémy pracujúce s príznakmi vyššej úrovne** – používajú informácie ako dialekt, jazyk, slovná zásoba, atď., z ktorých sa dá určiť pôvod a socialno-vzdelanostné postavenie jedinca. Sú však málo diskriminujúce.

- **Systémy pracujúce v reálnom čase –** poskytujú skoro okamžitú informáciu o hovoriacom s oneskorením niekoľkých sekúnd.

- **Systémy nepracujúce v reálnom čase –** poskytujú informáciu až po vyhodnotení celého prehovorenia a preto môžu používať zložitejšie algoritmy, dlhšie časové intervaly a dosahovať tým lepšie výsledky.

Tieto základné delenia, koncepty a tvrdenia v sebe obsahujú množstvá metód so svojimi vlastnosťami a obmedzeniami. Preto je kvôli podrobnejšiemu výkladu danej problematiky možné využiť prehľadové články ako napr. [6].

2.3 Viacúrovňová identifikácia

K niektorým aplikáciám (ako napr. AppStore) môže pristupovať viac používateľov z jedného alebo z viacerých typov zariadení. Pre tento účel je zavedená viacúrovňová identifikácia.

Viacúrovňová identifikácia sa zaoberá rôznymi požiadavkami na prístupové práva alebo kontrolu prostredníctvom niekoľkých biometrických modalít v kombinácii s menom, heslom alebo pinom. Zo skupiny biometrických modalít slúži na identifikáciu používateľa najmä hlas a tvár. V závislosti na výsledku identifikačného procesu môže byť prístup k systému alebo aplikácií umožnený alebo sa môže požadovať ďalšie overenie pomocou prihlasovacích údajov - používateľským menom, heslom alebo pinom. V budúcnosti môže byť identifikácia v systéme rozšírená na použitie rozpoznávania dúhovky, a tak nahradí použitie pinu alebo hesla. Celkovo sa tým zlepší bezpečnosť v systéme.

3 Komunikácia pomocou hlasových príkazov

Komunikácia pomocou rečových príkazov spadá do rozsiahlej vednej oblasti nazvanej *automatické rozpoznávanie reči* (**ASR**).

Tento termín označuje automatický (strojový) prepis hovoreného slova. Vstupom je digitálny rečový signál, pričom na výstupe je textový reťazec označujúci rozpoznané slová daného jazyka, ktoré sa nachádzajú v strojom používanom slovníku. Keďže slovník celého jazyka je obrovský, zvyčajne sa používajú len obmedzené verzie pre konkrétnu oblasť nasadenia. Výstupný text vyhovuje predpísanej gramatike alebo bežne používaným slovným spojenia existujúcim v jazyku.

ASR sa odlišuje od úlohy porozumieť vyhovoreniu, ktoré pracuje na vyššej úrovni. Oblasť, ktorá sa týmto problémom zaoberá sa volá umelá inteligencia. Jej vstupom je text po aplikácií ASR systému.

Počas niekoľkých posledných desaťročí prebieha intenzívny výskum v oblasti konštrukcie ASR systémov, ktoré sa využívajú v nasledovných oblastiach: systémy poskytovania informácií, dialógové systémy, pomôcky pre hendikepovaných, prepisy hovoreného slova pre archiváciu atď. Napriek tomu je to len nedávno kedy sa ukázali prvé použiteľné systémy, ktoré sa dajú použiť aj mimo kontrolované prostredia laboratórií. Zväčšuje sa aj ich komplexnosť a aplikácie prepisu diktovaného slova alebo dokonca aj plynulej reči či bežnej konverzácie nie sú úplnou raritou. Táto úloha je veľmi zložitá a musí riešiť problémy spadajúce do viacerých vedných oblastí. Praktické systémy musia pracovať v nepriaznivých prostrediach (zmena pozadia a prostredia, nahrávacích zariadení, atď.), musia podporovať veľkú gramatickú pestrosť jazyka, variabilitu hovoriacich (akustickej podoby), veľké slovníky, atď. Navyše bežný používateľ vyžaduje okamžitú odpoveď systému, nie je ochotný meniť svoj spôsob rozprávania, obmedziť sa na danú slovnú zásobu a rýchlo stráca trpezlivosť pri možných chybách ASR systému.

Tak ako existuje veľké množstvo ASR aplikácií rôznych zložitostí, požiadaviek a obmedzení, tak sa zaužívali rôzne klasifikácie takýchto systémov podľa rôznych kritérií. To najhlavnejšie z nich je veľkosť podporovaného slovníka. Rozlišujú sa nasledovné delenia:

- **Malý slovník-** obsahuje niekoľko desiatok až stoviek slov
- **Stredný slovník -** obsahuje niekoľko stoviek až tisícok slov
- **Veľký slovník –** niekoľko desiatok až stoviek tisíc slov

Toto delenie sa prirodzene s rastom technológie mení.

Systémy môžu byť ďalej závislé alebo nezávislé od hovoriaceho. To znamená, že systém nezávislý od hovoriaceho musí pracovať nezávisle od toho aký hovoriaci rozpráva. To si vyžaduje existenciu všeobecných modelov získaných z veľkého množstva hovoriacich. Systémy závislé na hovoriacom sú prispôsobené pre konkrétneho používateľa, a preto zvyčajne dosahujú lepšie výsledky, ale len pre daného hovoriaceho. Systémy, ktoré poskytujú okamžitú odpoveď sa nazývajú systémy pracujúce v reálnom čase. Významnú úlohu hrá aj to, v akom tvare je reč očakávaná na vstupe. Podľa toho sa rozlišujú nasledovné scenáre:

- Rozpoznanie izolovaných slov– kde sa na vstupe očakáva jedno slovo zo slovníka
- Diktačné systémy – predpokladajú, že je vyhovorených za sebou veľa slov ale je medzi nimi dostatočná, dobre detekovateľná pauza.
- Plynulá reč – kde slova idú plynulo za sebou, ale sa kladú viac či menej obmedzujúce pravidlá na používanú gramatiku
- Prirodzená reč – je najzložitejšia úloha, ktorá obsahuje plynulú reč bez obmedzujúcej gramatiky v reálnom prostredí nasadenia.

Nakoniec sa delia systémy podľa toho, akú časť reči modelujú (fonémy, slabiky, slová, frázy, atď.) a ako ju modelujú, napr. štatistický prístup pomocou *skrytých Markovových modelov* (**HMM)**.

Rečový signál je produkovaný hlasovými orgánmi človeka a je pozorovaný ako vibrovanie zvuku. Okrem iného nesie lexikálnu informáciu (čo bolo povedané). Tá je reprezentovaná sekvenciou vhodných zvukov, ktoré majú špecifické akustické formy. Tie akustické jednotky, ktoré stačia na vytvorenie daného jazyka sa nazývajú fonémy. Ich počet sa môže meniť a pohybuje sa od 40 do 60, a to v závislosti od jazyka. Samozrejme ich aktuálna akustická podoba sa mení od hovoriaceho k hovoriacemu a je významne ovplyvnená okolitými hláskami, tzv. koartikulačný efekt.

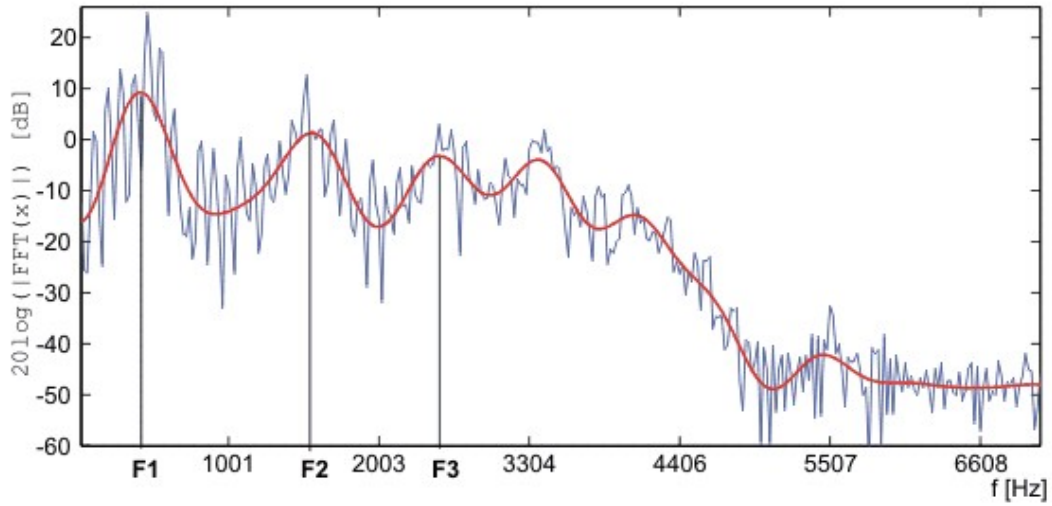
Na potlačenie tej časti informácie obsiahnutej v reči, ktorá neobsahuje lexikálnu informáciu je potrebné použiť metódy extrakcie vhodných príznakov. Zvyšná informácia totiž môže značne zhoršiť samotný spôsob rozpoznania slov. Teda informácia o hovoriacom (jeho nálade, zdravotnom a aktuálnom fyzickom a emočnom stave) je na prekážku. Je to teda primárna úloha systémov extrakcie príznakov vybrať len tie najlepšie, čo v skutočnosti nie je úplne dobre možné. Rečový signál obsahuje cca 100kb/s informácie, naproti tomu lexikálny obsah má len cca 10b/s informácie, preto algoritmy extrakcie príznakov pracujú ako účinné kompresory informácie.

Dobrou voľbou je teda simulovať sluchový systém človeka, matematicky ho opísať, zjednodušiť pre praktické použitie a adaptovať ho pre ľahkú kooperáciu s nadradenými systémami rozpoznania reči.

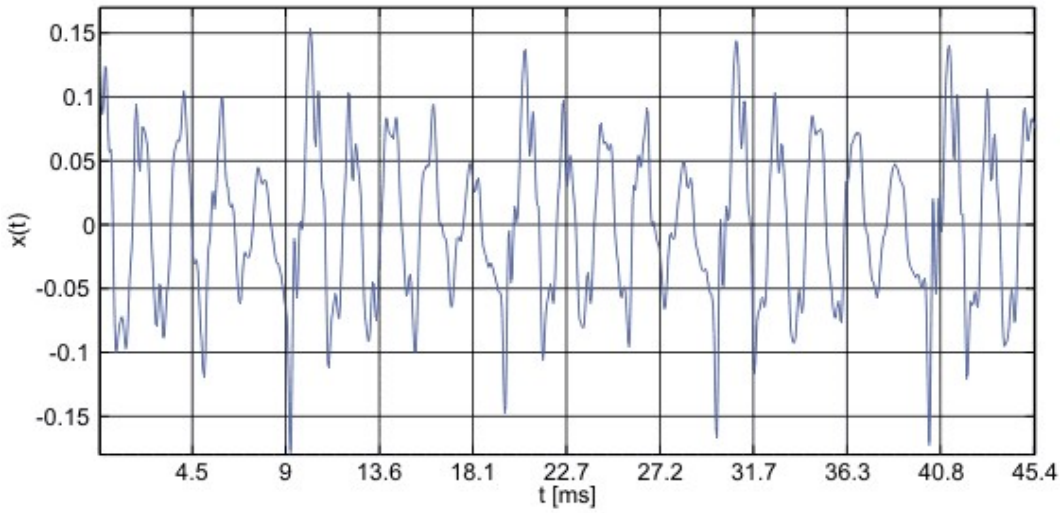
Existuje veľa úspešných metód simulujúcich buď sluchový systém človeka (kritické pásma, krivka hlasitosti, atď.) alebo proces produkcie reči (budenie, hlasový trakt, ...). Výhoda metód simulujúcich sluchový systém je však v tom, že majú počas evolúcie v sebe zakomponované vhodné mechanizmy na zvýraznenie

práve reči medzi ostatným bežnými nepriaznivými vplyvmi reálneho prostredia. Potláčajú teda bežné šumy pozadia a skreslenie prostredia.

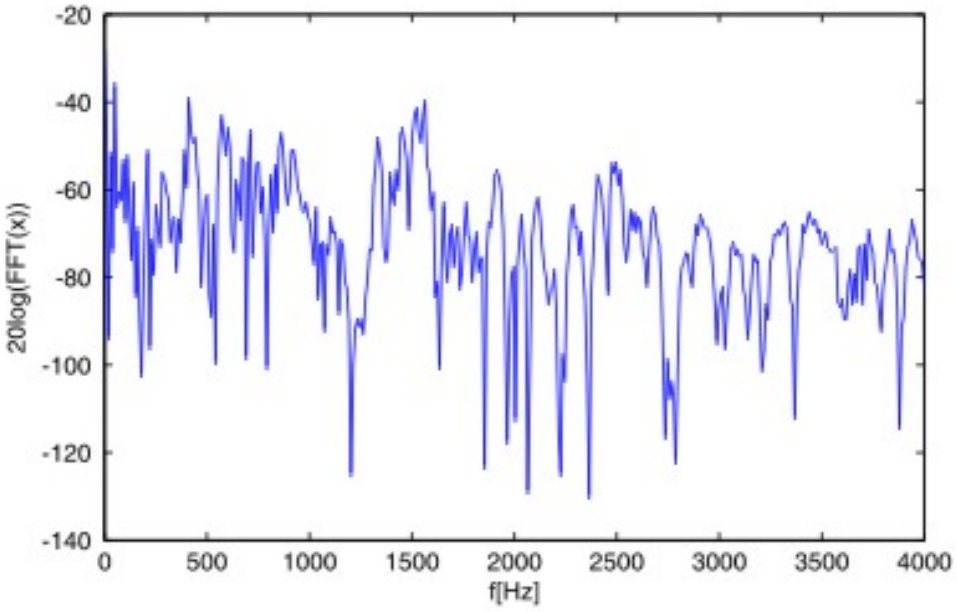
Počas výskumu sa zistilo, že podstatná lexikálna informácia je ukrytá v obálke frekvenčného spektra signálu. Tá napr. slúži na klasifikáciu jednotlivých foném medzi sebou. Presnejšie táto informácia je ukrytá v polohe, počte a tvare dominantných frekvenčných zložiek. Pre lepšiu demonštráciu tohto faktu je na obr. 3.1 zobrazené spektrum samohlásky "e", obálka spektra a polohy tzv. formantových frekvencií (dominantných frekvencií, ktoré súvisia s rezonanciami v hlasovom trakte). Na obr. 3.2 je zobrazený časový priebeh tej iste hlásky.



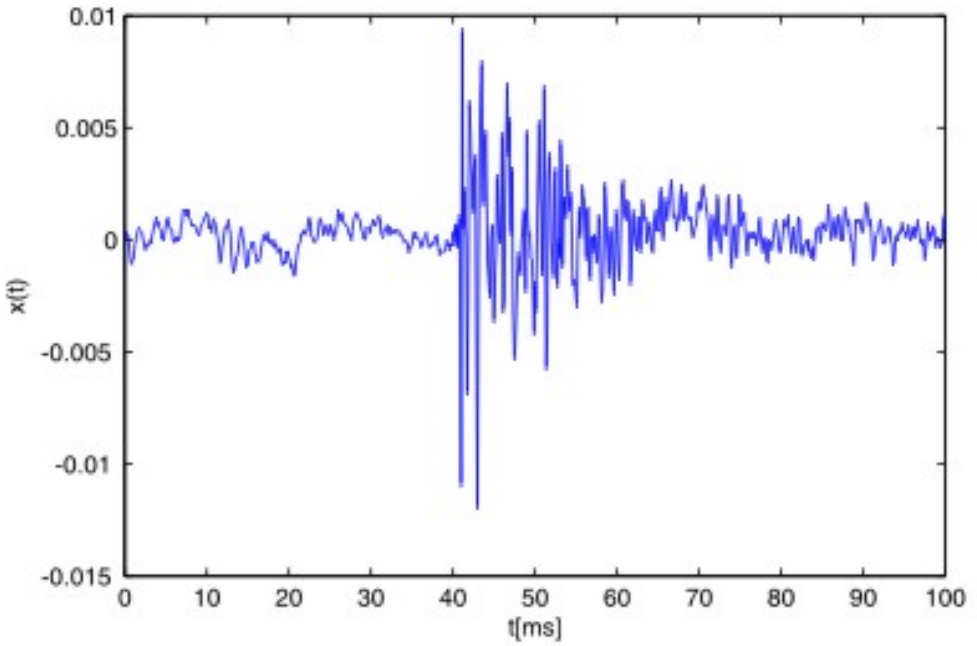
Obrázok 3.1. Spektrum formantové frekvencie a obálka spektra pre hlásku".



Obrázok 3.2. Signál hlásky "e".



Obrázok 3.3. Spektrum hlásky “t”.



Obrázok 3.4. Signál hlásky “t”.

Aby bol rozdiel v hláskach zrejmý je na obr. 3.3 zobrazené spektrum hlásky “t” a na obr. 3.4 jej časový priebeh.

Kvôli sumarizácií je v nasledujúcej tabuľke zobrazené rozloženie prvých 2 formantových frekvencií pre slovenské samohlásky, a to oddelene pre mužov a ženy získane ako priemer z populácie. Táto tabuľka poskytuje veľmi hrubý a jednoduchý spôsob ako klasifikovať spoluhlásky na základe ich frekvenčnej podoby.

Tabuľka 3.1. Prvé dve formantové frekvencie pre samohlásky, oddelene pre mužov a ženy

samohláska	Muži		Ženy	
	F1 [Hz]	F2 [Hz]	F1 [Hz]	F2 [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

Preto vybrané príznaky by mali byť schopné odhadnúť rozdiely v polohách formantových frekvencií. Na druhej strane musia byť „hluché“ voči zmenám, ktoré sú prirodzené a nepočuteľné. V nasledujúcej tabuľke sú uvedené najpočuteľnejšie a najmenej počuteľné modifikácie spektra ľudským uchom.

Tabuľka 3.2. Počuteľne a nepočuteľné zmeny spektra

Typy zmeny	
Počuteľné	**Nepočuteľné**
Počet formantových frekvencií	Celkový sklon spektra
Poloha formantových frekvencií	Frekvencie pod prvou formantovou frekvenciou
Šírka formantových frekvencií	Frekvencie nad treťou formantovou frekvenciou
-	Úzko-pásmové filtrovanie

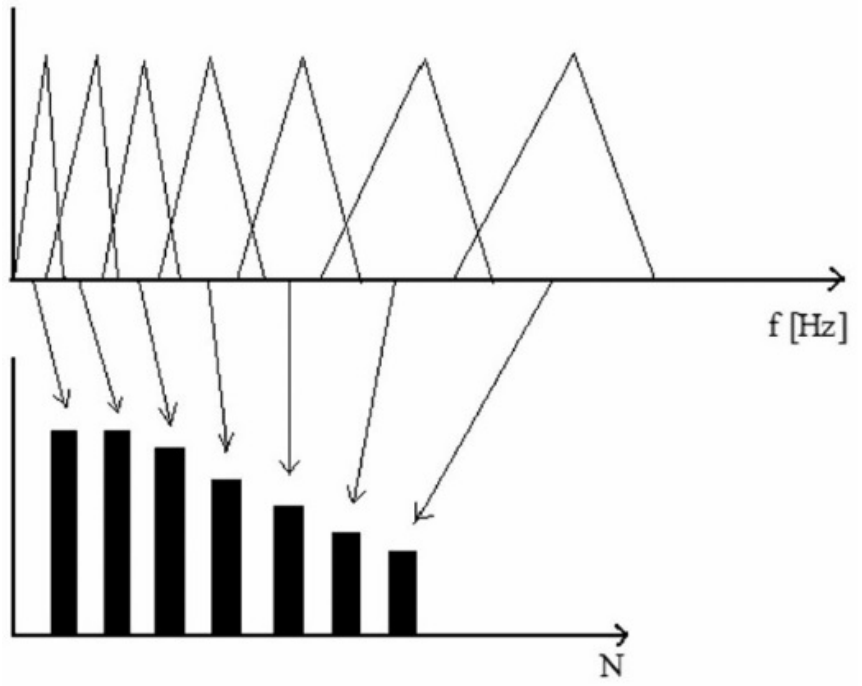
Navyše intenzita zvuku je vnímaná nelineárne, čo sa dá aproximvať pomocou funkcie logaritmu.

Z pohľadu rušenia by mali byť príznaky odolné voči aditívnym ako aj konvolutórnym šumom. V neposlednom rade sa musia dať príznaky jednoducho implementovať, matematicky opísať, mať kompaktné vyjadrenie atď. Zvyčajne je dobré, keď príznaky sú medzi sebou lineárne nezávislé, kvôli redukcií informácie a ľahšiemu následnému modelovaniu.

3.1 Parametre pre rozpoznanie reči

Na úvod treba poznamenať, že zatiaľ neexistuje parametrizácia reči, ktorá by spĺňala všetky vlastnosti potrebné pre optimálnu parametrizáciu uvedenú v predošlej časti. Preto je výskum v tejto oblasti stále aktívny a bolo nájdených veľa rôznych druhov. Tie najznámejšie a najlepšie sú *Melovo frekvenčné kepstráe koeficienty* (**MFCC**) a *perceptuálna lineárna predikcia* (**PLP**). PLP a MFCC sa snažia simulovať sluchový systém človeka, čo vedie k dobrým výsledkom v oblasti rozpoznávania reči. Obidve sú schopné vystihnúť polohu, počet a šírku formantových frekvencií, ktoré sú vnímateľné. Napriek podobnosti sa líšia v psychoakustických princípoch, ktoré obsahujú.

MFCC filtruje signál horno-priepustným filtrom (potlačenie vplyvu vyžarovania zvuku z úst), segmentuje reč Hamingovým oknom, z ktorého sa počíta spektrum pomocou FFT. Následne sa spektrum nelineárne upravuje, a to do Melovej škály (psychoakustická škála zodpovedá vnemu frekvencií človekom), ktorá sa rovnomerne rozdelí na prekrývajúce sa pásma s 50% prekrytím, s ktorých sa vypočíta výkon. Tým sa simuluje banka filtrov (psychoakustická), viď obr. 3.5. V poslednej fáze sa aplikuje logaritmus a **diskrétna kosínusová transformácia** (*DCT*). Logaritmus upravuje intenzitu zvuku tak, ako ju približne vníma človek a DCT zmenšuje závislosti koeficientov medzi sebou- kompresia.



Obrázok 3.5. Rozdelenie spektra do bánk filtrov v Melovej škále s použitím trojuholníkových okien.

PLP príznaky sa líšia v niekoľkých aspektoch: používajú Barkovu frekvenčnú mierku, vyhladzujú spektrum a vzorkujú ho v intervaloch 1 Bark, čím simulujú kritické pásma pre ľudský vnem (rozlíšenie), váhovanie krivkou rovnomernej

hlasitosti, transformácia energie do hlasitosti, výpočet koeficientov modelu reči a ich transformácia na kepstum.

Ako je vidno PLP je zložitejšie a obsahuje viac psychoakustických princípov ako MFCC, ale v priaznivých a laboratórnych podmienkach obidve metódy dosahujú podobné výsledky.

Dynamické príznaky a energia

Reč je postupnosť rôznych zvukov, preto je vhodne vyhodnocovať ich vývoj v čase a zmeny, ktoré nastávajú. Najbežnejšia metóda ako to robiť je pomocou rozdielových a akceleračných koeficientov, ktoré sa počítajú v čase zo základných príznakov. Môže sa to vykonať ako jednoduchý rozdiel susedných blokov, ale častejšie je to dané ako lineárna kombinácia rozdielov pokrývajúca širší časový úsek. Ďalej sa ukázalo, že jednoduchá energia môže dobre lokalizovať polohu samohlások vďaka ich veľkej energií, čo zvyšuje celkovú diskriminačnú schopnosť príznakov. Preto sa energia (normalizovaná) často pridáva do výsledného vektora príznakov.

3.2 Metódy rozpoznávania reči

Po parametrizácií je reč daná postupnosťou vektorov príznakov, napr. MFCC. Hrubo povedané proces rozpoznávania potom zoberie vzorky alebo modely známej reči (jej časti: fonémy, slabiky atď.) a porovnáva ich s neznámym prehovorením. Potom sú vzorky alebo modely s najväčšou zhodou prehlásené za rozpoznané slová. Kvôli špecifickým vlastnostiam reči ako napr.: rôzne prehovorenia rovnakého slova sa líšia v dĺžke ako aj vo frekvencií, atď. Navyše táto dĺžková variabilita sa nerovnomerne mení v rámci slova, preto niektoré časti môžu byť kratšie alebo dlhšie. Jednoduché lineárne natiahnutie (lineárna interpolácia) alebo skrátenie neprináša požadovaný efekt. Výsledné prehovorenie je dané ako reťazec jednotlivých častí (modelov) reči, ktoré musia na seba nadväzovať. Tieto špecifické vlastnosti dali vznik metódam rozpoznávania (porovnávania vzorov) vhodných práve pre reč. Najznámejšie z nich sú *Dynamické ohýbanie času* (**DTW**) a *Skryté Markovové Modely* (**HMM**). Samozrejme existuje veľa modifikácií, či dokonca kombinácií vhodných pre tú ktorú aplikáciu. V ďalšej časti budú stručne opísané práve tieto dve základné metódy, t.j. DTW a HMM

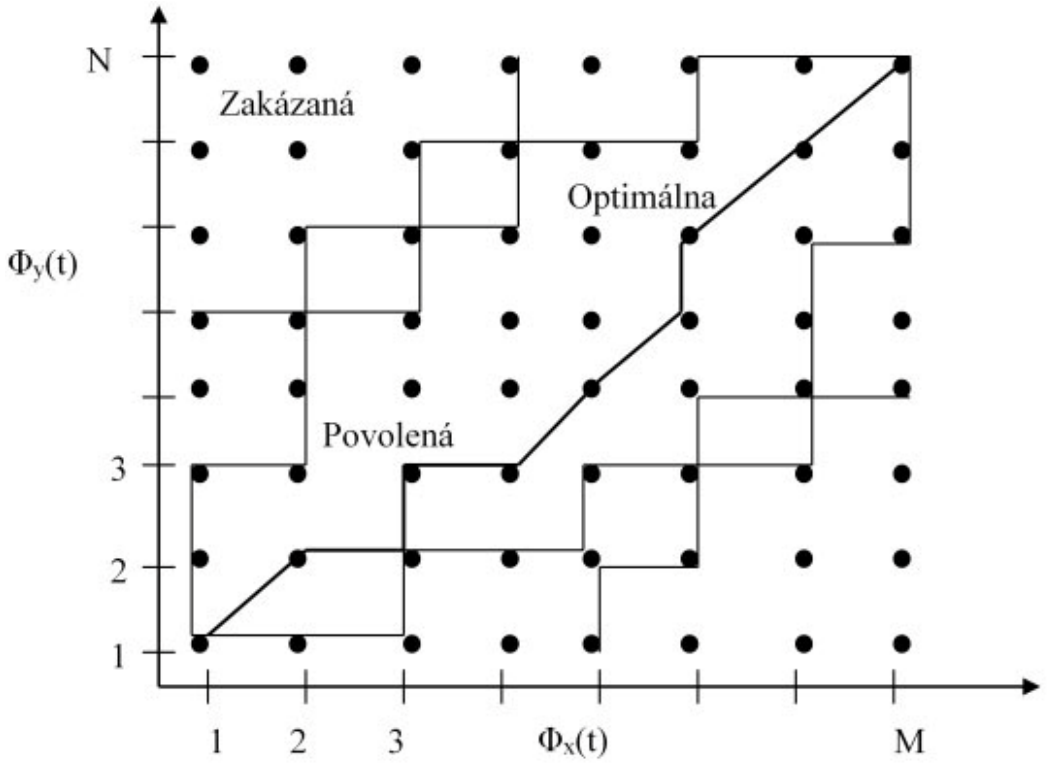
DTW

DTW je skratka pre dynamické ohýbanie času, ktoré slúži na akustické porovnanie dvoch vo všeobecnosti rôzne dlhých sekvencií príznakov. Je založené na nelineárnom ohýbaní časových indexov referenčnej ako aj neznámej postupnosti tak, aby vyhovovali štruktúre reči a dosahovali čo najmenšiu vzdialenosť medzi sebou. Tým dochádza k nelineárnej kompenzácií rozdielov medzi dĺžkami oboch postupností.

Logickou podmienkou je to, aby začiatočné a koncové vektory obidvoch postupností boli zarovnané na seba. To si vyžaduje použitie algoritmu detekcie začiatku a konca prehovorenia, čo môže byť samo o sebe pomerne zložitou úlohou. Samozrejme existujú modifikácie, ktoré túto striktnú podmienku presne nevyžadujú.

V stručnosti, táto metóda sa snaží nájsť mapovanie medzi dvoma sekvenciami tak, aby každý vektor z jednej postupnosti mal priradený jeden vektor z druhej postupnosti. To vedie k záveru, že niektoré vektory jednej postupnosti môžu byť namapované viackrát na rôzne vektory druhej postupnosti a naopak. Samozrejme tento proces sa musí riadiť istými základnými logickými obmedzeniami, najmä: začiatočné a koncové vektory jednej postupnosti musia byť namapované na svoje partnerské vektory druhej postupnosti, ohýbacia funkcia nesmie byť klesajúca (nedôjde k zmene poradia vektorov) a je stanovená maximálna rozdielnosť v polohe, ktorá ešte môže byť preklenutá. Pri výpočte DTW sa používajú dve matice (matica lokálnych a matica globálnych vzdialeností). V matici lokálnych vzdialeností sa nachádzajú akustické vzdialenosti medzi jednotlivými vektormi príznakov referenčného a neznámeho prehovorenia. Globálna matica slúži na výpočet akumulovanej minimálnej vzdialenosti a nájdenie cesty medzi obidvomi postupnosťami. Teda pre každý element globálnej matice existuje minimálna vzdialenosť a cesta spájajúca tento bod so začiatkom matice (začiatočné vektory

jednej a druhej postupnosti), ktorý sa nachádza vľavo dole. Táto situácia je objasnená na obr. 3.6. Samozrejme sú stanovené prirodzené pravidlá ako sa dá dostať z jedného bodu matice globálnych vzdialenosti do druhého, napr. cesta nesmie klesať. Keď sa proces dostane do konca (pravý horný roh), porovnanie skončilo a najmenšia vzdialenosť bola nájdená. Spätným sledovaním sa dá zistiť ako vyzerá minimálna cesta. Tento proces sa vykonáva pre každé slovo zo slovníka a to, ktoré má najmenšiu vzdialenosť s neznámym je prehlásené za rozpoznané prehovorenie. Ako je z výkladu zrejmé, táto metóda je priamo predurčená pre úlohy rozpoznávania izolovaných slov najlepšie závislých na hovoriacom.



Obrázok 3.6. Matica globálnych vzdialeností, optimálna cesta a globálne obmedzenia na cestu.

DTW malo významné postavenie v oblasti rozpoznávania reči najmä pre problémy izolovaných slov nezávisle od hovoriaceho. Ako sa však požiadavky zväčšovali (nezávislosť od hovoriaceho, plynulá reč) strácala svoje významné postavenie a bola nahradená metódou HMM.

Skryté Markovové Modely (HMM)

Skryté Markovové modely patria do štatistických modelovacích techník zväčša používaných v oblasti rozpoznávania reči. Tie z princípu ich konštrukcie jednoducho zahŕňajú požiadavku nezávislosti na hovoriacom, ako aj ľahké spájanie jednotlivých modelov (napr. hlások) do väčších celkov opisujúcich plynulú reč. Navyše tento princíp je pomerne dobre matematicky opísateľný.

Pre každú zvolenú rečovú jednotku (fonému, slabiku, slovo,...) je vytvorený HMM model vhodnej štruktúry. Zvyčajne modely základných časti jazyka ako sú fonémy majú rovnakú štruktúru a líšia sa len nastavením svojich voľných parametrov. V procese trénovania sa tieto voľné parametre nastavia tak, aby čo najlepšie opisovali vzorky z trénovacej databázy reči, ktorá musí byť dostatočne pestrá a veľká, aby pokryla všetky štatisticky významné variácie reči. Jednotlivé vyhovorenia sú popísané, teda je zrejmé aké postupnosti slov sa v nich nachádzajú. Najčastejšie sa modely nastavujú tak, aby trénovacie vzorky dosahovali na príslušných modeloch čo najvyššiu pravdepodobnosť. Tento postup sa nazýva aj trénovanie pomocou maximálnej vierohodnosti. Používajú sa aj iné kritéria ako: čo najväčšia separácia medzi správnymi a nesprávnymi modelmi, minimalizácia chyby, (takéto postupy sa označujú ako diskriminatívne trénovanie) atď., ktoré môžu dosahovať ešte lepšie výsledky.

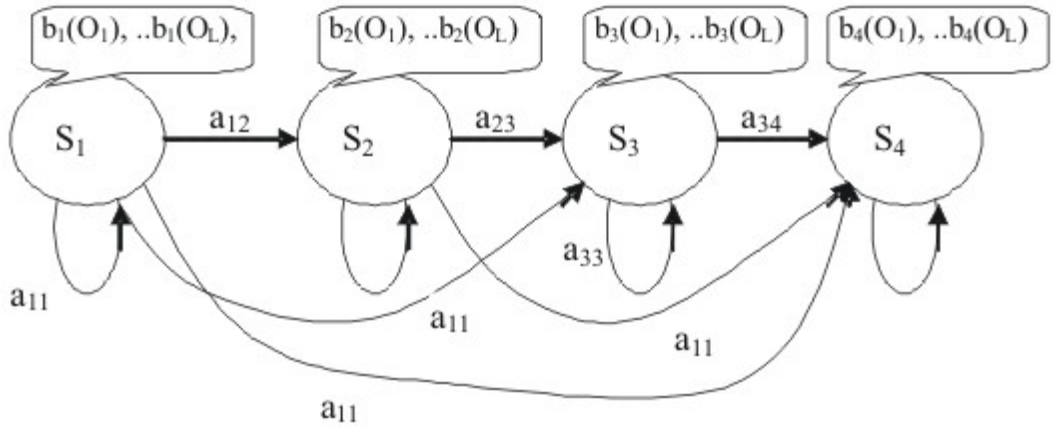
Každý model sa skladá z niekoľkých stavov, ktoré sú navzájom vhodne prepojené. Každé spojenie má pridelenú prechodovú pravdepodobnosť (p). Ďalej existuje počiatočná pravdepodobnosť (π) a to, že sa model nachádza na začiatku práve v danom stave. Potom sa takýmto modelom dá opísať pravdepodobnosť nastania sekvencie stavov $S_1, S_2, S_3,.. S_N$ na danom modeli, a to nasledovne:

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

Navyše je daná dodatočná pravdepodobnosť prislúchajúca ku každému stavu, ktorá opisuje pravdepodobnosť výskytu vektora príznakov X v danom stave S, teda $P(X/S)$. Tou je dané prepojenie medzi stavmi a vektormi rečových príznakov. Potom pravdepodobnosť pozorovania sekvencie príznakov $X_1, ..., X_M$ a stavov $S_1, S_2, .. S_N$ sa dá jednoducho vyjadriť ako:

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

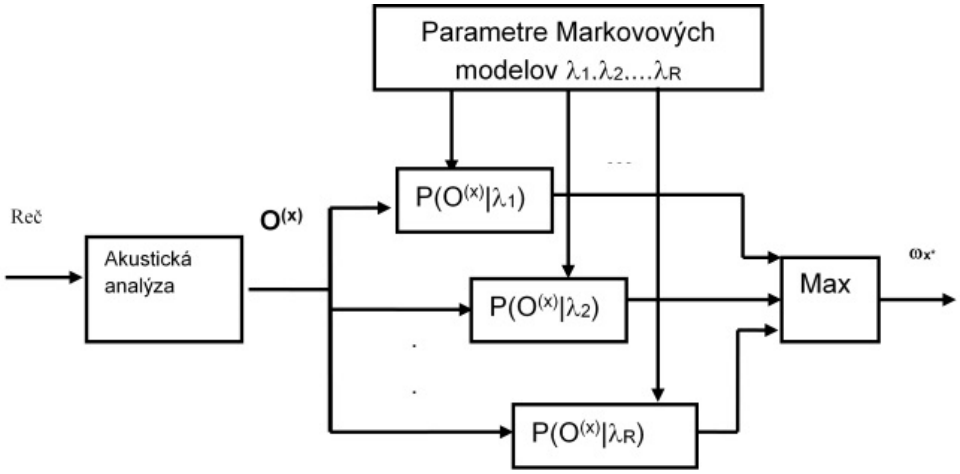
Príklad 4 stavového ľavo-pravého HMM modelu je zobrazený na obr. 3.7.



Obrázok 3.7. 4 stavový ľavo-pravý HMM model.

Potom sa v procese rozpoznávania počítajú pravdepodobnosti neznámej sekvencie príznakov na všetkých HMM modeloch slov v slovníku (vzniknú zreťazením základných HMM modelov) a vyberie sa ten model (postupnosť modelov), ktorý

dosahoval najväčšiu pravdepodobnosť. Tento proces je schematicky zobrazený na obr. 3.8.



Obrázok 3.8. Proces rozpoznávanie reči pomocou HMM.

V súčasnosti vyspelé systémy dosahujúce najmenšiu *slovnú chybu* (**WER**), založené na HMM, používajú diskriminatívne stratégie trénovania ako: *maximálna vzájomná informácia* (**MMI**), *Minimálna klasifikačná chyba* (**MCE**), atď. Nakoniec sa používajú hybridné spojenia HMM systémov a iných klasifikačných techník ako sú najmä systémy s podpornými vektormi a neurónové siete (hlboké). Tie v súčasnosti predstavujú špičku v oblasti rozpoznávania reči a dosahujú najnižšie WER pre najzložitejšie typy úloh.

Viac podrobnejších informácií ohľadom rozpoznávania reči a používaných metódach sa dá nájsť napr. v [7].

4 Navigácia pomocou gest

V súčasnej dobe predstavujú gestá veľmi populárny spôsob ovládania aplikácie a väčšina ľudí ich používa každý deň. Gestá používame v mobilných telefónoch, počítačových aplikáciách, herných konzolách, atď. Očakáva sa, že o pár rokov bude interakcia pomocou gest k dispozícii takmer v každom zariadení. Vysoká popularita navigácie pomocou gest poháňa výskumníkov v snahe zlepšiť tieto technológie. Je to populárny trend a keďže výkon počítača už nie je prekážkou v prirodzenejšej navigácií a ovládaní pomocou gest venuje sa mu veľká pozornosť [2].

V oblasti rozpoznávania gest sú dve hlavné oblasti, ktoré rozlišujeme podľa vstupného hardvéru. Prvá oblasť je zameraná na gestá určené pre dotykové zariadenia ako sú tablety, touchpady alebo smart telefóny. Druhá oblasť sa zaoberá gestami vykonanými pred 2-rozmernou alebo stereoskopickou kamerou. S menšími úpravami môžu byť rovnaké algoritmy aplikované na rozpoznanie gest z obidvoch oblastí. Naviac algoritmy pre stereoskopické kamery môžu byť upravené na použitie 3D súradníc pre monitorovanie vo všetkých troch smeroch [2].

4.1 Klasifikácia gest podľa rôznych kritérií

Gestá možno rozdeliť do dvoch základných kategórií podľa skúseností používateľov. Vrodené gestá, ktoré sú založené na všeobecnej skúsenosti všetkých používateľov, ako je napríklad presunúť objekt vpravo pohybom ruky doprava, uchopenie objektu znázornené uzavretými prstami a pod. Samozrejme, vrodené gestá môžu byť ovplyvnené zvykmi a kultúrou. To je potrebné zohľadniť pri vývoji pre príslušnú skupinu ľudí. Použitie týchto gest nevyžaduje potrebu používateľa učiť sa gestá, stačia mu jeho skúsenosti pre ovládanie gestami. Používateľ teda môže ovládať aplikáciu prirodzene. Druhá skupina gest sú naučené gestá. V tomto prípade je nutné, aby sa ich používateľ naučil. Gestá používané na navigáciu systémov a aplikácií musia byť jednoduché, prirodzené a mali by využívať minimálnu ľudskú energiu.

Statické gestá

Statické gestá nie sú závislé na pohybe. Statické gestá predstavujú tvary končatín znázorňujúcich gestá, ktoré nesú zmysluplné informácie (obr. 4.1).

Sú "nepohyblivým" typom gest, kde nepotrebujeme informácie o pohybe. Preto nie je potrebné skúmať sekvencie snímok, ale len jeden snímok. Sekvencia niekoľkých snímkov, ktoré obsahujú rôzne gestá, je dôležitá len vo vyššej úrovni porozumenia.

Kontinuálne gestá

Kontinuálne gestá slúžia ako základ pre interakciu s aplikáciou. Kontinuálne gesto je predĺžené sledovanie pohybu, ktorý nepredstavuje žiadnu špecifickú pózu, ale používa sa na interakciu s aplikáciou alebo vo virtuálnom prostredí (obr. 4.1).

Typickým príkladom kontinuálnych gest je ovládanie hry pomocou bezdotykovej technológie, akou je napríklad Microsoft Kinect. Systém mapuje zmeny v postoji alebo póze k zmenám vo videohre, ale nie je tam žiadny špecifický pohyb vyvolávajúci sled zmien.

Dynamické gestá

Dynamické gestá predstavujú pohyb, ktorý umožňuje používateľom priamo manipulovať s objektom alebo ovládať aplikáciu (obr. 4.1).

Dynamické gestá môžu byť definované dvoma spôsobmi. Prvý prístup je algoritmicky nastaviť pravidlá a podmienky, ktoré musí vykonané gesto spĺňať aby bolo úspešne rozpoznané. Druhým spôsobom je použitie šablón. Šablóny sú množiny bodov, ktoré identifikujú tvar gesta. Na vyhodnotenie úspešnosti sú použité sofistikované algoritmy, ktoré porovnávajú pohyby používateľa so šablónami.

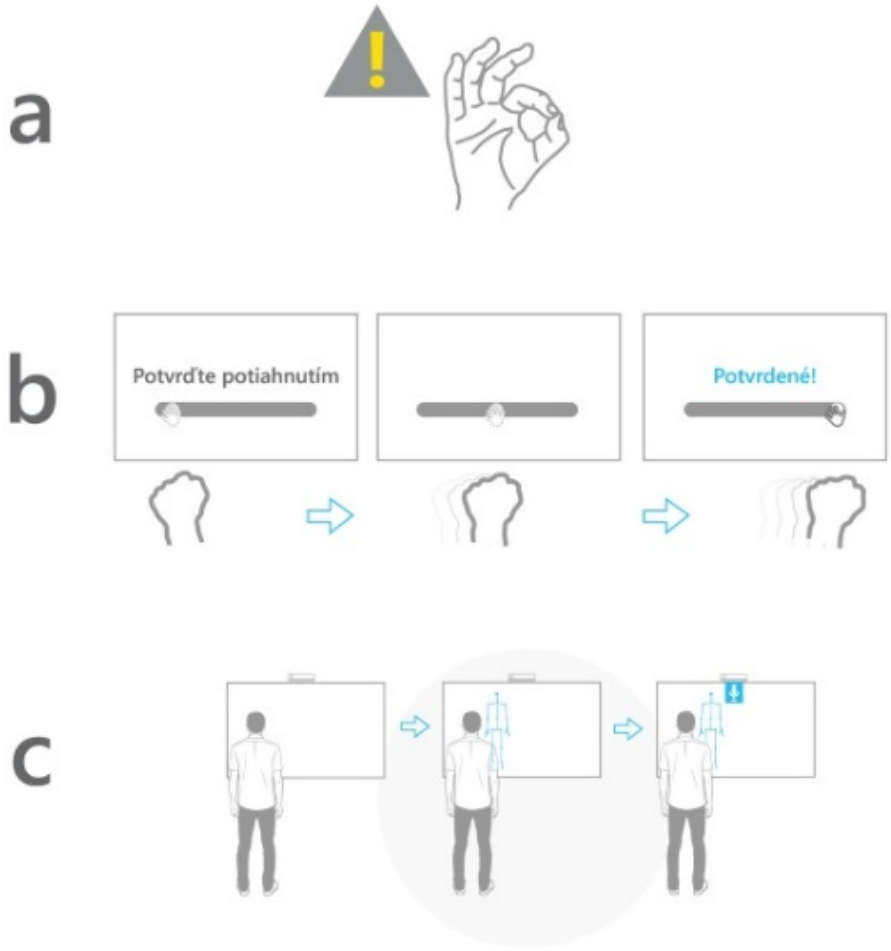


Fig. 4.1 – Typy gest: a) statické gesto, b) dynamické statické gesto, c) kontinuálne gesto

4.2 Získavanie obrazových dát

Na začiatku bol výskum algoritmov rozpoznávania gest založený na vstupe z **RGB** (*Red Green Blue*) kamery. Tento prístup prináša niektoré výhody, ktoré sú však prevážené nevýhodami. RGB obraz sa skladá z farebných kanálov a existuje mnoho faktorov, ktoré negatívne ovplyvňujú kvalitu RGB obrazu.

Zmena osvetlenia môže radikálne ovplyvniť kvalitu obrazu. Zmena podmienok osvetlenia v miestnosti úplne mení jas a kontrast obrazu. Tým sa stáva rozpoznávanie pomocou RGB obrazu neisté a menej presné [2].

V posledných niekoľkých rokoch začali vedci používať hĺbkový obraz pre ďalšie spracovanie v oblasti rozpoznávania gest. Hĺbkový obraz je dvojrozmerný obraz, ktorý obsahuje ďalšie informácie - hĺbku. Väčšina hĺbkových senzorov je založená na infračervenom (*infrared* **IR**) žiarení (napr. Kinect senzor na Fig. 4.2). Senzor vracia dáta, ktoré predstavujú vzdialenosť každého pixelu v snímke od snímača (senzora). Získané vzdialenosti môžu byť jednoducho prepočítané na odtiene sivej, a tak sú hĺbkové dáta reprezentované ako čiernobiela video sekvencia. Získanie dát nevyžaduje žiadne zvláštne podmienky.

Hĺbkové dáta majú viac informácií ako dáta získané RGB kamerami. Hlavnou výhodou je, že hĺbková kamera využíva infračervené svetlo a v infračervenej zóne nie je toľko šumu a rušení. Takže infračervený snímač nie je tak citlivý na kvalitu scény ako RGB snímače [2].

Fig. 4.2 – Obraz z hĺbkovej kamery

Vzdialenosť každého pixelu zo snímača je definovaná v milimetroch.

Ak chceme previesť hĺbkový obraz do obrazu v odtieňoch sivej, musíme vedieť minimálnu a maximálnu možnú vzdialenosť.

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

kde d je aktuálna vzdialenosť daného pixelu, d_{min} je minimálna možná vzdialenosť od senzora a d_{max} je maximálna možná vzdialenosť od senzora [2].

Ako pomôcka pri získavaní vstupného obrazu môžu slúžiť extra prvky ako svetielkujúci náramok, prsteň alebo malá loptička v ruke.

4.3 Metodológia návrhu gest

Problematika rozpoznávania gest nie je triviálna. Všetky obmedzenia a nepresnosti, ktoré sa vyskytujú pri získavaní vstupných obrázkov pomocou rôznych zariadení, je nutné vziať do úvahy pri procese navrhovania gest pre aplikáciu alebo kontrolu systému. Metodika pre návrh gest musí splniť niekoľko požiadaviek: prirodzenosť, ľahkú zapamätateľnosť a komfort. Komfort je najdôležitejším kritériom pre aplikácie riadené gestami .

Rovnako ako v každej novej oblasti, tak aj v oblasti ovládania gestami je jedným z hlavných problémov v procese adaptácie neexistujúca štandardizácia. Keď náhodne vyberieme desať ľudí a požiadame ich, aby urobili gesto pre spustenie hudby, je vysoko pravdepodobné, že bude navrhnutých aspoň 5 rôznych gest. To je dôvod, prečo každý výrobca poskytujúci ovládanie gestami používa inú sadu gest. Z praktického hľadiska je to veľmi komplikujúci fakt, pretože používateľ sa musí učiť nové gestá pre každý nový systém. Existencia aspoň jednej štandardnej sady gest pre ovládanie základných funkcií by mohla významne urýchliť proces udomácnenia ovládania gestami v bežnej praxi. V dnešnej dobe sú už k dispozícií televízne prijímače vyžívajúce ovládanie gestami. Sú založené na snímaní obrazu 2D kamerou. Umožňujú jednoduchú navigáciu v menu televízora. Aj napriek svojej nedokonalosti má zavádzanie gest do praxe výchovný vplyv na ľudí, ktorí si postupne zvykajú na tento druh ovládania.

Súčasné sady gest sú založené na vstupných periférnych zariadenia (klávesnica, myš,..) používaných s počítačmi. Gestá sa v podstate snažia odstrániť vstupné zariadenia, ale zachovávajú rovnaké spôsoby používania. Cieľom návrhu gest od základov, teda bez ohľadu na existujúce používateľské rozhranie, je čo najviac sa priblížiť prirodzenému bezdotykovému ovládaniu tak, akoby gestá boli jediným vstupným zariadením.

4.4 Techniky pre rozpoznávanie základných gest

Celkovým cieľom pri rozpoznaní gesta rukou je určiť nasledovné: umiestnenie ruky, držanie alebo póza ruky alebo vyjadrenie gesta.

Rozpoznávanie gest môže byť vykonané dvoma spôsobmi. Buď použitím dátovej rukavice, ktorá prevádza informáciu o ohybe končatiny do elektronickej formy alebo sa použije prístup založený na pozorovaní, čiže kamera slúži ako ľudské oko zaznamenávajúce polohy tela, ktoré sa potom extrahujú pri spracovaní obrazu [5].

Je zrejmé, že prvý spôsob prináša presnejšie výsledky, keďže sa jedná o priame a presné pozorovanie skutočného stavu príslušnej časti ľudského tela, na ktorej sa nachádza oblečenie typu dátová rukavica.

Na druhej strane prístup založený na pozorovaní nemá žiadne ďalšie požiadavky na zariadenie pre koncového používateľa (s výnimkou kamery), takže je vhodný pre všeobecné použitie [5].

Použitie dátovej rukavice je pomerne nepohodlné, pokiaľ ide o komfort používateľa. Tiež vybavenie potrebné na použitie tejto metódy by bolo neprijateľne nákladné pre väčšinu štandardných zákazníkov, takže je vhodný iba pre špeciálne použitie (napr. digitalízácii pohybu pre animované filmy, tvorbe počítačových hier a pod.).

Nevýhoda metódy založenej na pozorovaní spočíva v algoritmickej zložitosti, kde je potrebné značné množstvo času a výpočtového výkonu pre extrakciu a extrapoláciu pohybu tela [5].

Existujú viaceré algoritmy, ktoré sa zameriavajú na rôzne aspekty gestikulujúcej osoby a vychádzajú z rôznych predpokladov. Všeobecne platí, že môžu byť rozdelené do dvoch kategórií, a to metódy založené na vzhľade a 3D metódy založené na princípe analýzy modelu. 3D metódy založené na princípe analýzy modelu porovnávajú vstupné parametre končatiny s 2D projekciou trojrozmerného modelu končatiny. Prístup založený na vzhľade využíva obrazové príznaky na modelovanie vizuálneho vzhľadu končatiny a porovnáva ho s extrahovanými príznakmi zo vstupného videa [5].

V sekcii 4.1 boli definované tri typy gest. Na detekciu statických gest (t.j. pózy) sa používa všeobecný klasifikátor. Pod klasifikátorom rozumieme – prvok alebo vlastnosť, na základe ktorého rozhodneme o zaradení do skupiny alebo kategórie. Takže klasifikátor nám hovorí o tom, ktoré gesto (alebo póza) bolo rozpoznané. Avšak pri dynamických gestách treba zohľadniť aj časové hľadisko a preto rozpoznanie vyžaduje techniky, ktoré spracúvajú aj túto dimenziu, napr. skryté Markovove modely (HMM). Iný spôsob je pomocou pohybových modelov.

Niektoré z techník používaných pre statické (a dynamické) rozpoznanie gest sú K-means algoritmus, KNN, SVM, už spomínaný HMM algoritmus, DTW algoritmus alebo neurónové siete [3].

Techniky použité pre rozpoznanie statických gest

Pre rýchlejšie spracovanie dát sa v procese rozpoznávania gest nepracuje s celým obrazom.

Potom, čo je automaticky detekovaná ruka, označí sa oblasť okolo nej a ďalej sa spracováva iba táto oblasť . Tým sa zníži výpočtová náročnosť. Získané vzdialenosti sa prevedú do odtieňov sivej a získame obrys ruky.

Nasleduje popis dvoch metód používaných v procese rozpoznávania.

Metóda konvexných nedostatkov

V prvom kroku sa musí oddeliť ruka od pozadia obrazu. Oddelenie sa môže spraviť z hĺbkového obrazu kde sa určí, ktorý pixel patrí ruke a ktorý nie. Druhým krokom je určenie obrysu (kontúry) ruky.

Tvary mnohých zložitých objektov sú dobre charakterizované pomocou konvexných nedostatkov. Fig. 4.3 znázorňuje koncept konvexných nedostatkov na obrázku ruky. Konvexný obal je zobrazený ako tmavá čiara okolo ruky, oblasti označené A až H sú "nedostatky" patriace k tomuto obalu. Ako je vidieť, tieto konvexné nedostatky charakterizujú tiež stav ruky. Algoritmus vráti súradnice troch bodov, *počiatočný bod*, *najhlbší bod* a *koncový bod* (Fig. 4.4). Najhlbší bod možno chápať ako maximálnu vzdialenosť medzi obalom a obrysom ruky.

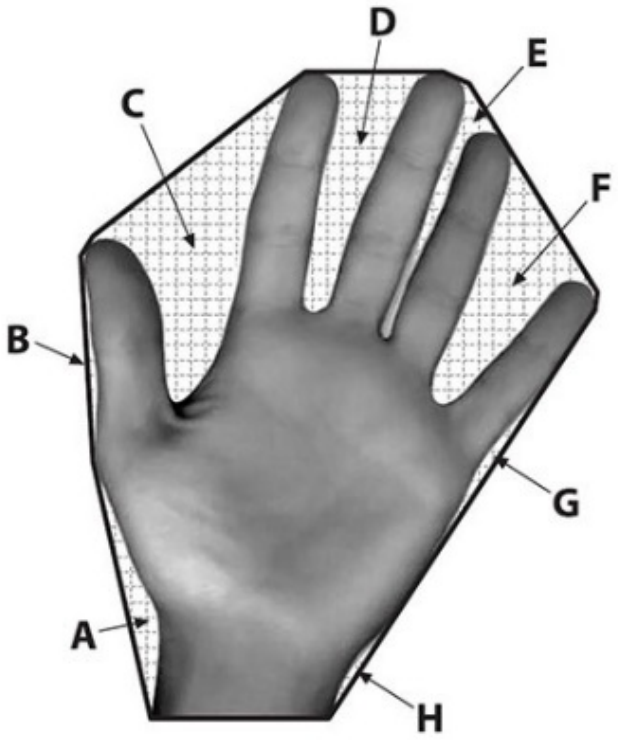


Fig. 4.3 – Konvexný obal vykreslený ako tmavá čiara okolo ruky

Fig. 4.4 – Označená oblasť sú konvexné nedostatky a tri body: počiatočný bod (vľavo), najhlbší bod (dolu) a koncový bod (hore)

Cieľom algoritmu je nájdenie bodu na prste, ktorý je najďalej od stredu ruky. Prvým krokom je odstrániť všetky nedostatky, ktorých výška je menšia ako zadaná hodnota. Potom sa odstránia nedostatky, ktoré majú vzdialenosť medzi počiatočným bodom a maximálnym bodom viac ako je zadaná hodnota, ktorá je definovaná dynamicky (ak sú dva body príliš ďaleko znamená to, že nemôžu predstavovať prst). Ďalej sú tiež odstránené nedostatky, ktoré majú vzdialenosť medzi počiatočným a hĺbkovým bodom menej, než je stanovená hodnota. Táto hodnota sa mení dynamicky podľa veľkosti oblasti, kde je detekovaná ruka (Fig. 4.5). Posledným krokom je odstrániť všetky nedostatky, ktoré sa vyskytujú pod zápästím.

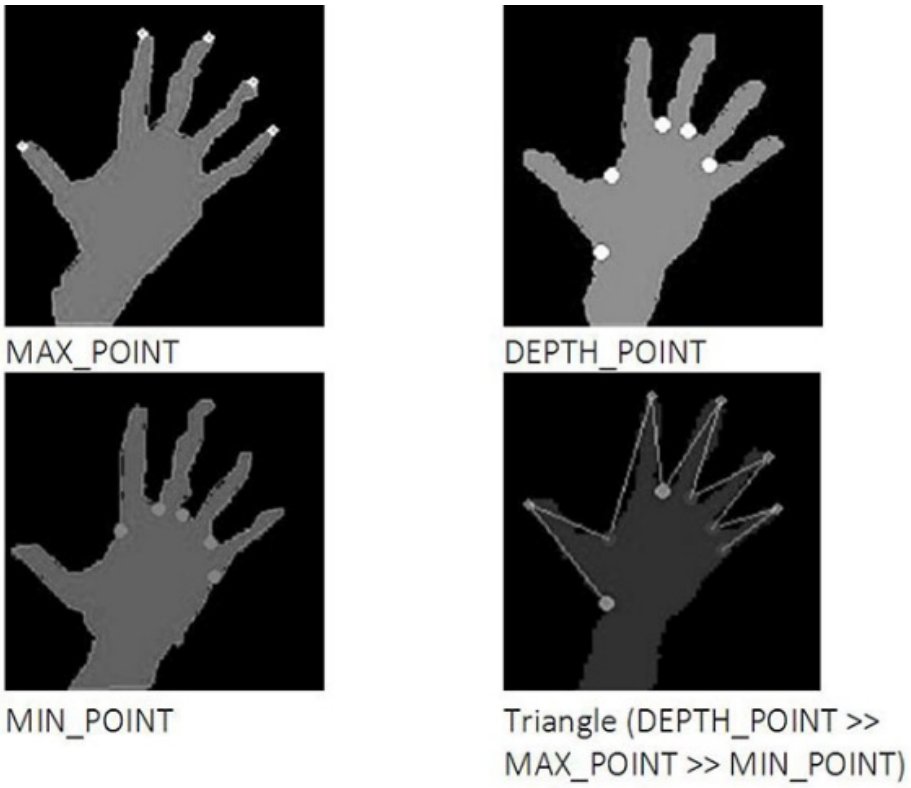


Fig. 4.5 – Znázornenie logiky používajúcej trojuholníky na minimalizovanie nedostatkov (obrázok vľavo hore znázorňuje maximálne body, obrázok vpravo hore znázorňuje hĺbkové body, obrázok vľavo dolu znázorňuje minimálne body a obrázok vpravo dolu znázorňuje výsledný trojuholník)

Rozpoznávanie gesta ruky založené na častiach

V prvom kroku tohto algoritmu sa extrahujú časti údajov (dátové segmenty) obsahujúce ruku a prevedú sa na binárny obraz. Stred dlane (Fig. 4.6) sa vypočíta z vnútorného kruhu a pridaním bodu na obal kontúry, ktorá patrí ruke a má maximálnu vzdialenosť od zisteného nedostatku.

Fig. 4.6 – Hľadanie stredu dlane

Všetky body z kontúry ruky sú mapované na osi X. Y-ová os potom opisuje relatívnu vzdialenosť jednotlivých bodov od stredu dlane. Všetky mapované body vytvoria krivku (Fig. 4.7).

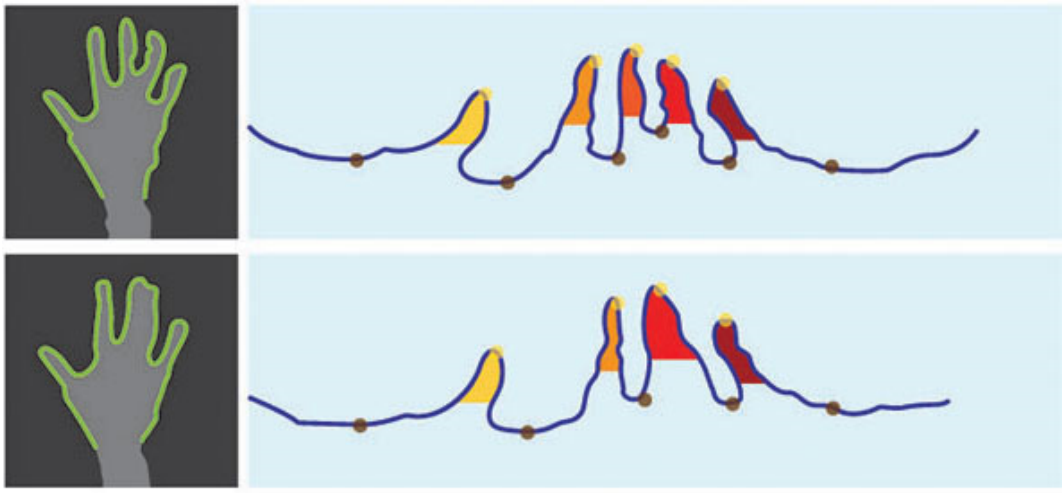


Fig. 4.7 – Krivková reprezentácia

Ďalším krokom algoritmu je analýza krivky s cieľom nájsť lokálne maximum. Potom, čo je dokončená analýza krivky, nájdené maximum, začnú sa hľadať prsty. Vzdialenosť každého maxima je porovnaná s nastavenou prahovou hodnotou. Každý prst má špecifickú váhu a na jej základe a na základe vzťahov medzi prstami sú prsty rozpoznané (vzťah je porovnanie váhy s danými hodnotami a na základe toho je známy počet prstov, teda ak váha <1,5x priemerná váha - segment obsahuje jeden prst, atď.). Tento algoritmus funguje aj keď sú prsty spojené (Fig. 4.8).

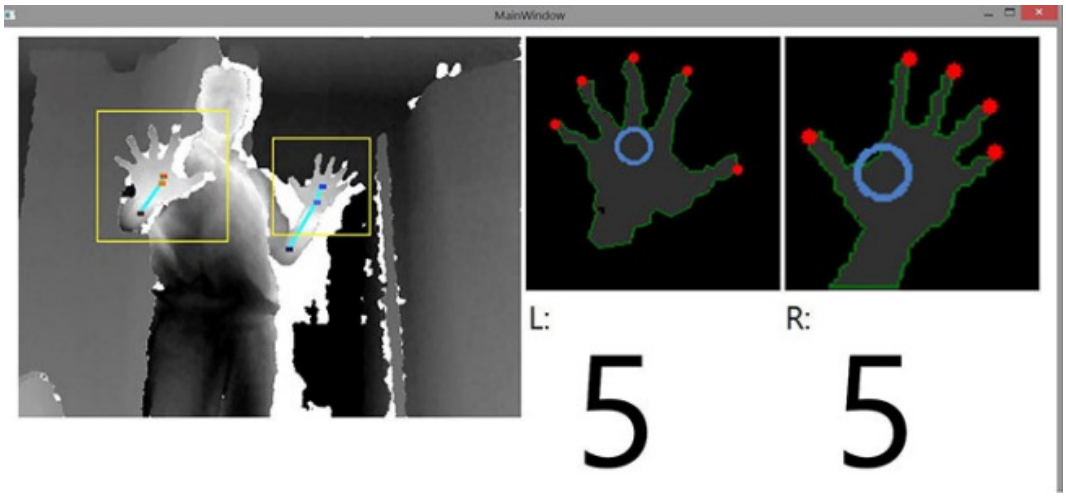


Fig. 4.8 –Detekované prsty

Techniky použité pre rozpoznanie dynamických gest

Dynamické gestá zaisťujú prístup používateľov k súkromnému obsahu alebo mu umožnia ovládanie systému alebo aplikácie. Možno ich použiť aj namiesto hesla. V začiatkoch rozpoznávania gest boli väčšinou používané neurónové siete a genetické algoritmy.

Tieto metódy mali prijateľnú úspešnosť rozpoznávania, ale najväčšou nevýhodou bol požadovaný výpočtový výkon a čas potrebný na trénovanie neurónových sietí. Obidva tieto aspekty boli neprijateľne vysoké pre praktické aplikácie.

V dnešnej dobe sa používajú na rozpoznanie gest nové techniky. Boli vyvinuté algoritmy, ktoré nevyžadujú neurónové siete, napríklad Vyhľadávanie v zlatom reze (Golden Section Search), Algoritmus prírastkového rozpoznávania (Incremental Recognition Algorithm) a pravdepodobnostné modely ako skryté Markovove modely (HMM). Na zvýšenie úspešnosti spomínaných algoritmov sa môže použiť strojové učenie (tzv. Machine learning). V posledných rokoch sú veľmi populárne HMM metódy. Hlavným dôvodom je to, že HMM prístup je veľmi dobre známy a používaný v mnohých oblastiach.

Algoritmus navrhnutý Kristensson a Denby [4], pôvodne navrhnutý pre rozpoznávanie ťahov digitálnym perom a na dotykovom displeji, je možné rozšíriť aj pre dynamické gestá.

V tomto algoritme je definovaná šablóna ako súbor segmentov popisujúcich gesto. Je zrejmé, že rozpoznanie gesta z danej sady gest, kde sú gestá dostatočne vzájomne odlíšiteľné, môže byť úspešné iba ak bola vykonaná dostatočne veľká časť z gesta (Fig. 4.9).

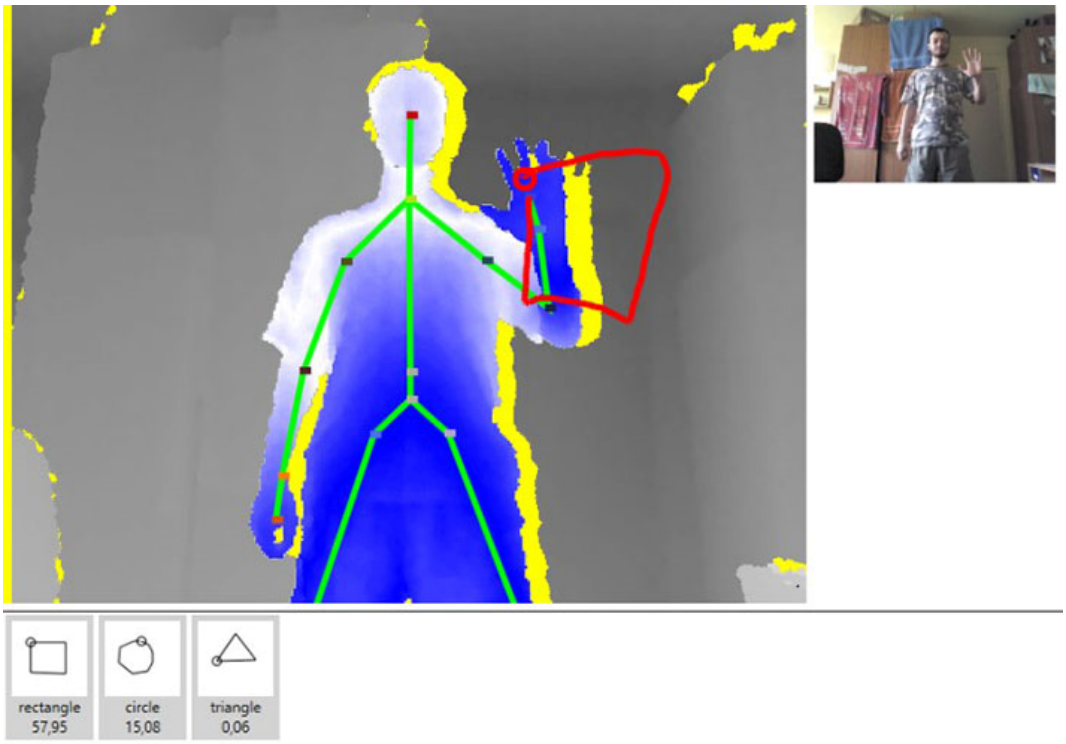


Fig. 4.9 – Náhľad testovanej aplikácie. Červená čiara znázorňuje vykonané gesto. Šablóny pod obrazom predstavujú najpravdepodobnejšie vykonané gestá (zľava doprava možno vidieť tvary štvorec, kruh a trojuholník. Pri každom útvare je vidieť aj hodnotu pravdepodobnosti s akou bolo gesto vykonané).

Používanie gest sa postupne rozšírilo o ťahové gestá. Tento typ gest prináša veľmi prirodzený a pohodlný spôsob najmä pre koncového používateľa. Ťahové gestá sú určené pre rýchle a bežné prezeranie v ponuke, v programoch a galériách a obsahujú 4 smery pre každú ruku a niekoľko kombinácií vľavo-vpravo. Tzv. metóda Kruhové rozpoznávanie dynamických gest (Circle Dynamic Gesture Recognition) - CDGR publikovaná v [1], je založená na detekcii ruky, rýchlosti pohybu a vzdialenosti. Ak ruka vykoná rýchlejší pohyb a vnútorný kruh prejde do vonkajšieho kruhu, systém spracuje tento pohyb a určí typ gesta. Gesto je dané uhlom vykonaného pohybu od polovice do vonkajšieho kruhu. Možné gestá sú: prejsť ťahom vľavo, vpravo, hore a dole. To iste gesto môže byť vykonané obidvoma rukami. Používateľ môže tiež vykonať gesto priblíženia a oddialenia (Fig. 4.10).

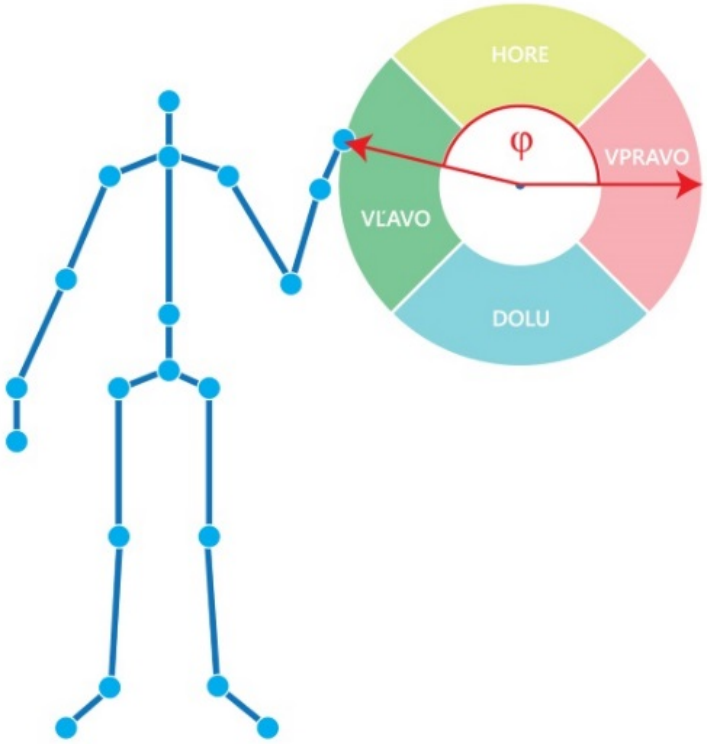


Fig. 4.10 – Kruhové rozpoznávanie dynamických gest (CDGR circle dynamic gesture recognition)

4.5 Použitie gest v praxi

Rozpoznanie gest môže byť interpretované ako jeden z prvých možných spôsobov ako počítače porozumejú človeku alebo reči ľudského tela. To môže viesť k veľkému kroku, keď vstupné zariadenia ako klávesnica a myš sa stanú nepotrebnými (však sa s nimi neráta ani v žiadnych sci-fi filmoch). Zavedením rozpoznávania gest bude možné napr. ukázať prstom na obrazovku počítača tak, že sa kurzor bude podľa toho pohybovať.

Navigácia gestami je stále viac a viac populárna predovšetkým v smart telefónoch, tabletoch a moderných televízoroch. Každý pozná ťahové gesto pre posun hore/dole alebo presunúť doľava/doprava, zväčšovanie alebo zmenšovanie. Existuje bohatá ponuka aplikácií ponúkajúcich sadu gest pre navigáciu v dotykových zariadeniach, napr. na ovládanie hlasitosti, uzamknutie obrazovky, ovládanie prehrávača hudby, kopírovanie obsahu obrazovky, návratu na domovskú obrazovku, krok späť, návrat k posledným aplikáciám či menu.

Oblasť pre použitie gest nie je obmedzená. Ako príklad môžeme spomenúť aplikáciu na trénovanie tenisu s využitím 3D rozpoznávania gest, ktorú publikoval Cristian García Bauza a jeho tím.

Napríklad Samsung predstavil rozpoznávanie gest vo svojej „vlajkovej lodi“, svojom najvyššom modeli - TV v roku 2013. Rozpoznávanie pozostáva z algoritmu sledovania ruky s rozpoznaním gesta "kliknite na tlačidlo", ktoré sa vykonáva zatvorením dlane. Aby používateľ mohol prepínať kanály alebo nastaviť hlasitosť, musel najprv zdvihnúť ruku k televízoru a po zjavení kurzora pohybovať rukou tak, aby kurzor prišiel k tlačidlám (analógia počítačovej myši). Na zvýšenie hlasitosti musel používateľ "klikať na tlačidlo", t.j. zatvoriť dlaň ruky toľko krát, kým sa dosiahla požadovaná hlasitosť. Tento prístup by mohol byť považovaný za intuitívny aspoň pre počítačovo zdatných používateľov, ale skôr to môže slúžiť ako príklad neprirodzených gest.

5 Grafické rozhranie - návrh

Grafické používateľské rozhranie (*graphical user interface* (**GUI**) predstavuje spôsob interakcie medzi ľuďmi a strojmi/počítačmi. Používateľské rozhranie zahŕňa hardvérové (fyzické) a softvérové (logické) komponenty.

Používateľ odošle príkaz alebo úlohu do počítača/systému a počítač reaguje späť plnením požadovaného príkazu alebo úlohy. V minulosti GUI prijímalo vstup prostredníctvom zariadení ako sú napríklad počítačová klávesnica a myš. Ako je vidieť s príchodom dotykových displejov, GUI sa prispôsobilo a akceptuje aj vstup daný dotykom prstov alebo pera.

Vytvoriť vhodné GUI pre multimodálne (alebo iné) aplikácie vyžaduje dobrú prax, skúsenosti a spätnú väzbu používateľa (v súčasnej dobe sa táto oblasť výskumu nazýva miera spokojnosti používateľa). GUI pre multimodálne aplikácie by malo umožniť ich jednoduché, efektívne a používateľsky komfortné ovládanie. V ideálnom prípade používateľ potrebuje minimálne úsilie na nastavenie vstupu a zariadenie vyžaduje minimálny prevádzkový čas na poskytnutie výstupu. Blízka téma k miere spokojnosti používateľa (*user experience* **UX**) je proces zlepšovania miery spokojnosti používateľa (*User experience design* **UXD**).

6 MMI applikácie

Moderný aplikačný výskum je zameraný na čo najprirodzenejšie multimodálne rozhranie a jeho začlenenie do multimediálneho systému už používaného na dennej báze. Multimodálne rozhranie nie je obmedzené iba na televíziu, ale má využitie v rôznych aplikáciach. Vízia budúcich aplikácií a systémov nie je len ovládať TV a prístup k multimediálnemu obsahu, ale budovať systém, ktorý pozná zvyky svojich používateľov, ich preferencie a záujmy. Takýto systém sa môže stať inteligentnou súčasťou domácnosti a môže poskytovať pokročilé prepojenie pre rôzne služby (nákupy popri sledovaní TV, e-voľby alebo digitálny vrátnik).

Pre čo najefektívnejšie použitie multimodálneho rozhrania je dôležité vedieť, ktorá modalita je zviazaná s ktorým krokom a čo sa má vykonať. Pre rôzne druhy krokov, resp. akcií sú vhodné rôzne druhy modalít. Napríklad zadávanie textu by bolo zbytočne obtiažne, časovo náročné a nepohodlné cez gestá, oveľa lepšie by sa tu uplatnilo rozpoznávanie reči. S týmto vedomím je nevyhnutné pre multimodálny systém použiť integračnú platformu, ktorá bude poskytovať požadované údaje čiastkovým aplikáciám. Tieto aplikácie potom nepotrebujú poznať zdroj modality.

Jednoduchý modelový príklad: MMI aplikácia pre ovládania TV zobrazí používateľov, ktorí boli rozpoznaní prostredníctvom hlasu alebo rozpoznaním tváre pomocou príslušného modulu. Iba rozpoznaní používatelia v tomto zozname majú povolené ovládať TV pomocou gest, hlasom alebo inou modalitou. Keď používateľ opustí miestnosť, je zo zoznamu automaticky odstránený.

Ostatné časti aplikácie zobrazujú odporúčané kanály. V závislosti od spoločných preferencií rozpoznaných ľudí v miestnosti systém dokáže prispôsobiť ponuku týchto kanálov. Pomocou ťahových gest je možné prehrať položku zo zoznamu, prípadne prehrávanie zastaviť. Niektoré z odporúčaných kanálov môžu byť aj zamknuté, čo znamená, že používatelia bez oprávnenia si ich nemôžu prehrať bez zadania tajného vzoru. Tajný vzor predstavuje dynamické gesto.