VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

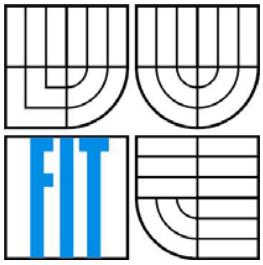
TEXTOVĚ ZÁVISLÉ ROZPOZNÁVÁNÍ MLUVČÍHO

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE JAN FUX
AUTHOR

BRNO 2013

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

TEXTOVĚ ZÁVISLÉ ROZPOZNÁVÁNÍ MLUVČÍHO

TEXT DEPENDENT SPEAKER VERIFICATION

BAKALÁŘSKÁ PRÁCE
BACHELOR´S THESIS

AUTOR PRÁCE — JAN FUX
AUTHOR

VEDOUCÍ PRÁCE — ING. PAVEL MATĚJKA, PH.D.
SUPERVISOR

BRNO 2013

Abstrakt

Cílem této bakalářské práce bylo navrhnout systém pro textově závislé rozpoznávání mluvčího. Bylo otestováno několik přístupů na databázi MIT, která obsahuje nahrávky průměrné délky 0,46s. Z otestovaných přístupů se jeví jako nejlepší kombinace systému DTW s využitím odhadu posteriorních pravděpodobností fonémů (posteriogramu) jako výstupu z Fonémového rozpoznávače, a akustického SID systému založeného na iVektorech a PLDA (Probabilistic Linear Component Analysis). Fúze těchto dvou systémů pomocí Neuronové sítě dosahuje nejlepších výsledků (EER) a to 17,84% pro ženy a 16,38% pro muže, což je relativní zlepšení 49,9% u žen a 54,2% u mužů oproti samostatnému akustickému rozpoznávání.

Klíčová slova

rozpoznávání mluvčího, DTW, foném, neuronová síť, GMM, rozpoznávání promluvy, fúze, DET křivka, EER, DCF

Abstract

The goal of this Bachelor's thesis was to design text dependent speaker recognition system. There were few systems tested for MIT database. This database contains recordings of 0.46s average length. Best case for recognition is to use a combination of DTW system using posterior probability estimation (posteriograms) as an output of Phoneme recognizer and acoustic SID system based on iVectors and PLDA (Probabilistic Linear Component Analysis). Fusion with Neural network gives the best results (EER). These are 17.84% EER for women and 16.38% for men. It's 49.9% relative improvement for women and 54.2% for men against acoustic recognition alone.

Keywords

speaker recognition, speaker verification, DTW, phoneme, neural network, GMM, speech recognition, fusion, DET curve, EER, DCF

Citace

FUX, Jan: Textově závislé rozpoznávání mluvčího, bakalářská práce, Brno, FIT VUT v Brně, 2013

Textově závislé rozpoznávání mluvčího

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením Ing. Pavla Matějky, Ph.D.
Další informace mi poskytli Bc. Miroslav Skácel a Ing. Petr Schwarz, Ph.D.
Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

……
Jan Fux
14. 5. 2013

Poděkování

Za pomoc při řešení této práce a za spoustu užitečných podnětů a informací velmi děkuji svému vedoucímu Ing. Pavlu Matějkovi, Ph.D.

Obsah

1 Úvod

Komunikace člověka s počítačem a dalšími výpočetními systémy je v dnešní době důležité téma, kterému se věnují mnohé vývojářské firmy. Doby, kdy například počítačová myš fungovala jako jediné ukazovátko na počítačové ploše, jsou už dávno pryč. Nyní je trend využívat plně všech lidských smyslů. Proto například vznikají dotykové displeje, které myš plně nahrazují. Odpadá tím externí zařízení, které bylo potřeba pro používání počítače. Ovládání prsty je také mnohem rychlejší a přesnější, než tomu bylo v případě myši. To je jeden z příkladů, proč se těmito alternativními způsoby komunikace s počítači a jinými výpočetními systémy zabývat.

Dalším příkladem nového přístupu ke komunikaci s počítačem je ovládání hlasem. Zde je využit člověku dobře známý způsob komunikace, a to mluvenou řečí. Tento způsob je člověku nejpřirozenější a také nejrychlejší, protože jej běžně používá v životě. V této práci se nebudeme zabývat přímo komunikací jako takovou, ale spíš tím, jak je možné realizovat, aby počítač určil osobu, která pronesla nějakou promluvu (*Rozpoznávání mluvčího*), a aby poznal, co bylo v dané promluvě řečeno (*Rozpoznávání promluvy*).

Systémy rozpoznávání mluvčího pomáhají při různých bezpečnostních opatřeních, kde je lidská síla nespolehlivá nebo drahá. Využití je zejména při řízení přístupu do budov nebo na webové stránky, kde je na místě ověření, že mluvčí, který požaduje přístup do daného objektu či na webovou stránku, patří skutečně do skupiny uživatelů s oprávněním vstupu. Pokud si představíme webové stránky nějaké banky, lze najít využití systémů rozpoznávání mluvčího také u provádění transakcí, kde uživatel bude před převodem peněz ověřen na základě hlasového příkazu k úhradě. Z policejního prostředí je možno využít rozpoznávání mluvčího také jako důkazný materiál. Policie monitoruje hovory a vyhledává, zda celostátně hledaná osoba někam nevolá. Rozpoznávač je tedy využit k odhalení totožnosti pachatele. Oproti tomu systémy rozpoznávání promluvy se dají využít například jako příkazy pro počítač. Jedná se například o listování složkou, vyhledání určitých typů souborů, zobrazení řečené webové stránky, procházení emailů pomocí řečených příkazů apod. U těchto systémů není podstatné, kdo promluvu pronesl, ale co bylo řečeno a co tento příkaz znamená (jaká akce se má vykonat). V praxi se často kombinují oba přístupy do jednoho systému. Například u hesla pro vstup do budovy systém nejprve ověří, zda bylo řečeno správné heslo, a následně provede kontrolu záznamu, jestli mluvčí, který heslo pronesl, je jedním z uživatelů, kteří mají povolen vstup. Následující podkapitoly jsou inspirovány zdrojem [16].

1.1 Návrh úlohy a potenciální rizika

Jak bylo zmíněno v kapitole 1, samotné rozpoznávání mluvčího není příliš bezpečné a obzvlášť u krátkých nahrávek může dávat špatné výsledky. Chceme-li aplikaci nasadit na takových místech, kde je kladen důraz na bezpečnost, musíme se zabývat i rozpoznáváním promluvy. Jako příklad si představme, že zaměstnanec X má zřízen přístup do budovy A. Do budovy vstupuje po ověření svého hlasu, podle kterého je identifikován, a systémem je ověřeno, že do této budovy

zaměstnanec přístup skutečně má. Hlas zaměstnance je však při obědě nahrán podvodníkem P, který se následně může dostat do budovy A tím způsobem, že po výzvě na ověření hlasu přehraje hlas zaměstnance X do mikrofonu. Systém hlas vyhodnotí jako hlas zaměstnance a podvodníka vpustí dovnitř. Při kombinaci rozpoznání hlasu a rozpoznání promluvy je však toto nebezpečí eliminováno, protože podvodníkovi nestačí pouze jakákoli nahrávka s hlasem zaměstnance, ale potřebuje mít nahrané přímo zaměstnancem pronesené heslo, které je mnohem těžší získat.

Další motivace ke kombinaci rozpoznávání mluvčího a textu je při použití krátkých nahrávek. U nich totiž není k dispozici dostatečné množství informací, které by jednoznačně mluvčího identifikovaly, a může být horší úspěšnost. Pokud však využijeme i rozpoznávání promluvy, skóre se zlepší. Takovému systému, u kterého ověřujeme hlasovou i obsahovou část předložené promluvy, se říká *Textově závislé rozpoznávání mluvčího*. Právě tímto přístupem se tato práce zabývá. Pokud si vystačíme pouze s hlasem mluvčího a nezajímá nás obsah promluvy, jedná se o *Textově nezávislé rozpoznávání mluvčího*.

Zajímavou variantou pro předchozí příklad je tzv. *Liveness test* systém. Tento systém předloží zaměstnanci X při vstupu do budovy několik vygenerovaných slov, které musí přečíst. Jedná se vlastně o ověření, že je zaměstnanec opravdu přítomen. Podvodník P, kterému se minulý den podařilo nepozorovaně nahrát zaměstnance A při pronášení promluvy, není schopen pomocí této nahrávky do budovy A proniknout, protože vstupní heslo je vygenerováno náhodně z mnoha natrénovaných. Každý ze zaměstnanců má natrénovánu svoji databázi hesel. Pro tyto účely je dobré volit generování hesel spíše slovních než číselných, protože u číslic hrozí nebezpečí, že podvodník může získat nahrávky číslic 0 až 9 a tyto čísla potom kombinovat podle požadavků. Trénování takového systému je možné například několika desítkami minut řeči. Z vět, které by měly být foneticky vyvážené, si systém následně vytvoří sadu fonémů pro daného uživatele. Při ověřování jeho hlasu potom systém předkládá pro přečtení slova, která je z daných fonémů schopen sestavit a porovnává takto sestavenou referenci se čteným slovem. Jednodušší a pro tento případ vhodnější variantou je databáze slov, která uživatel předem do systému nahraje. Měla by to být kombinace běžných slov s některými netypickými a měl by jich být dostatek. Systém potom poskládá některá z nich různě za sebe a uživatel je musí přečíst. Tedy v příkladu výše by museli všichni zaměstnanci s povolením vstupu do budovy A nejprve nahrát do systému svoji databázi slov. Slova si můžou zvolit sami, která se jim líbí, ale musí jich být alespoň 20 a více. Stačí i přečtení nějakého článku z novin. Systém by jim následně při každém vstupu do budovy zobrazil na obrazovce například 4 náhodná slova z jejich natrénovaného slovníku. Po jejich přečtení by systém mohl velmi spolehlivě určit, zda se nejedná o podvodníka.

1.2 Druhy systémů rozpoznávání mluvčího

Při rozpoznávání můžeme narazit na dvě rozdílné úlohy, a to *Identifikace* a *Verifikace*. Identifikace je případ systému, který má uložené přepisy promluv od více mluvčích. Tento systém se snaží napasovat vstupní promluvu na některou z uložených, u kterých má definováno, kterému řečníkovi patří. Výsledkem je tedy jméno daného mluvčího, který promluvu pronesl, nebo chyba, pokud se systému nepodařilo spárovat vstupní promluvu s žádným přepisem hlasu, který má

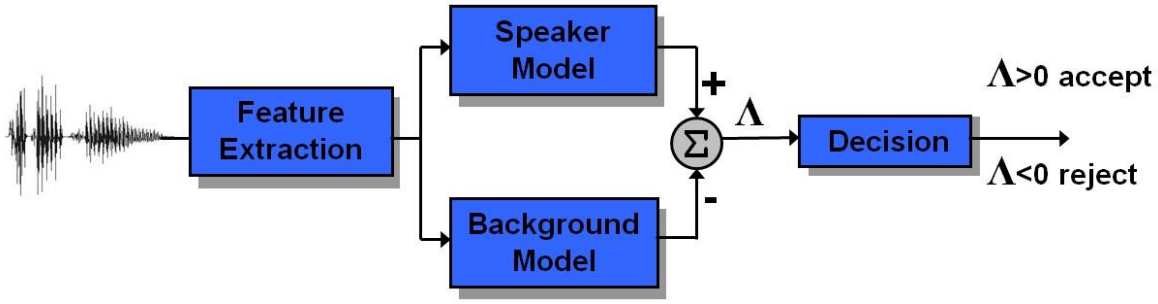
systém natrénován. Představme si jako příklad vstup do budovy, kde pracuje N zaměstnanců. Každý ze zaměstnanců má v centrální databázi uložen hlasový přepis, který je jednoznačně identifikuje. Zaměstnanec je po příchodu do práce vyzván k hlasovému zadání svého hesla a jeho hlas je porovnán s hlasovými přepisy v databázi. Pokud je v databázi jeho hlasový přepis nalezen, systém vyčte z databáze údaje o tomto člověku, pozdraví jej jménem, umožní mu vstup a zároveň mu do evidence docházky k jeho jménu zaznamená příchod.

Druhou úlohou v systémech rozpoznávání mluvčího je Verifikace. Rozumíme jí ověření, že daný mluvčí je tím, za koho se vydává. Systému je na vstup dána nahrávka obsahující ověřovanou promluvu a nějaký identifikátor mluvčího, za kterého se vydává. Identifikátorem může být například jeho celé jméno. Výsledkem systému je potom přijetí, nebo odmítnutí v závislosti na tom, zda mluvčí byl či nebyl tím, za koho se vydával. Jako příklad lze uvést volání do banky ze soukromého mobilního telefonu. Banka má zákazníkovo telefonní číslo uloženo, takže když z něj volá, banka podle něj zjistí, kdo je vlastníkem tohoto čísla. Banka tedy ví, kdo z jejích zákazníků by měl u telefonu být, ale protože mobilní telefon může být ukraden a může volat podvodník, který nemá žádné oprávnění zjišťovat informace o zákazníkově účtu, musí banka provést verifikaci, že do telefonu skutečně mluví oprávněná osoba. To provede na základě hlasu. Stěžejním úkolem je například při telefonní komunikaci odlišit ze záznamu, co který účastník telefonního hovoru říkal. Takový problém se řeší rozdělením textu na části – segmentace a identifikace mluvčího v těchto jednotlivých segmentech pomocí rozpoznávače.

1.3 Schéma rozpoznávání mluvčího

Obecně by se systém na rozpoznávání mluvčího dal znázornit pomocí schématu na obrázku [Obr. 1]. *Statistika* pro testovací promluvu *S* je vyjádřena jako pravděpodobnostní poměr:

$$\Lambda = \log \frac{\text{Pravděpodobnost,že S } \textbf{patří} \text{ do modelu mluvčího}}{\text{Pravděpodobnost,že S } \textbf{nepatří} \text{ do modelu mluvčího}} \quad (1)$$



[Obr. 1] Schéma rozpoznávání mluvčího

Na obrázku [Obr. 1] je vidět, že vstupem do systému je řečový signál. Z něj jsou následně vyčteny jeho charakteristické znaky a tyto znaky jsou porovnávány s *modelem mluvčího* a *modelem pozadí*. Model pozadí je referenční model vzniklý trénováním na mnoha mluvčích. Naproti tomu model mluvčího vznikl trénováním pouze promluvami od mluvčího, kterého se snažíme rozpoznat. Pokud tedy promluvu S řekl náš rozpoznávaný mluvčí, měl by na ni lépe padnout model mluvčího. Odečtením pravděpodobnosti modelu pozadí poté získáme hodnotu, dle které určíme, zda promluvu řekl náš mluvčí ($\Lambda > 0$), nebo zda se jedná o podvodníka ($\Lambda < 0$).

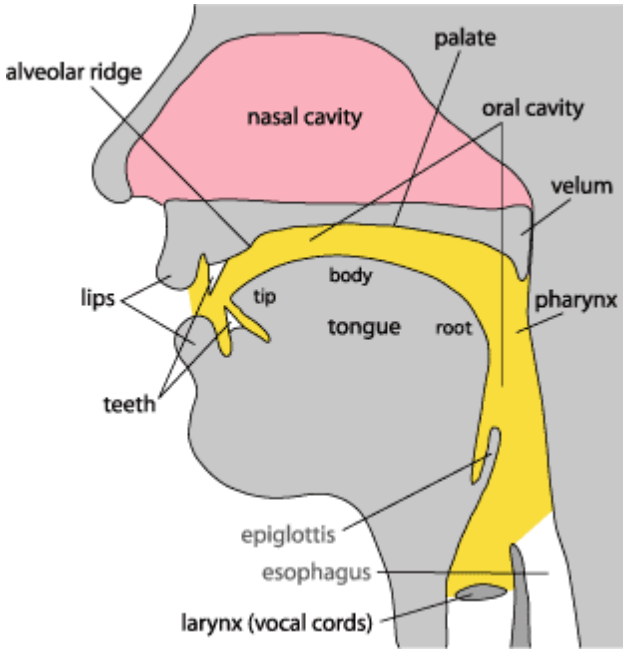
Fázi rozpoznávání předchází ještě fáze trénování modelu, ve které se také z řečového signálu nejprve extrahují potřebné znaky, a poté použijí k natrénování modelu. Rozdíl je však v tom, že k trénování se využívá více dat (promluv) než k rozpoznávání. Je to proto, že čím lépe model natrénujeme, tím máme lepší výsledky při rozpoznávání. Proto se k trénování také používají nahrávky z různých prostředí (tichá místnost, na ulici, v restauraci, v autě,…) a také různé nahrávací zařízení (profesionální mikrofon, telefonní mikrofon, handsfree,…).

1.4 Důležité příznaky při rozpoznávání mluvčího

Neexistují dva mluvčí, kteří by řekli tutéž promluvu úplně stejně. Záznam jejich signálu bude mít vždy určité odlišnosti. Tyto odlišnosti můžeme rozdělit do tří kategorií podle toho, zda se s nimi člověk narodil (fyzické znaky), nebo k jejich vývoji došlo až v průběhu života (naučené znaky).

1.4.1 Vysoko-úrovňové a nízko-úrovňové znaky

Mezi *nízko-úrovňové znaky* zařadíme hlavně akustické znaky řeči jako tvar nosní dutiny, délka hrtanu a hlasivek, sípání a drsnost hlasu. Tyto znaky vycházejí z tělesné struktury konkrétního člověka a tvaru jeho hlasivkového ústrojí. Anatomie lidského hlasového ústrojí je zobrazena na obrázku [Obr. 2].



[Obr. 2] Anatomická struktura hlasového ústrojí

Na pomezí těchto dvou kategorií najdeme například prozodické znaky (zvukový projev) jako je rytmus a intonace, a také jak mluvčí pracuje s hlasitostí svého projevu. Zda mluví monotónně a špatně artikuluje, nebo naopak mluví zřetelně a zesiluje hlas u důležitých slov. Tyto znaky závisí na konkrétním osobnostním typu daného jedince a také například na vlivu rodičů.

Do *vysoko-úrovňových znaků* patří všechny naučené rysy a znaky, jako je sémantika, idiolekt (výrazové zvláštnosti), výslovnost a osobní projev. Jsou ovlivněné především sociálním postavením ve společnosti, zaměstnáním, místem narození apod.

1.4.2 Implementovatelné znaky

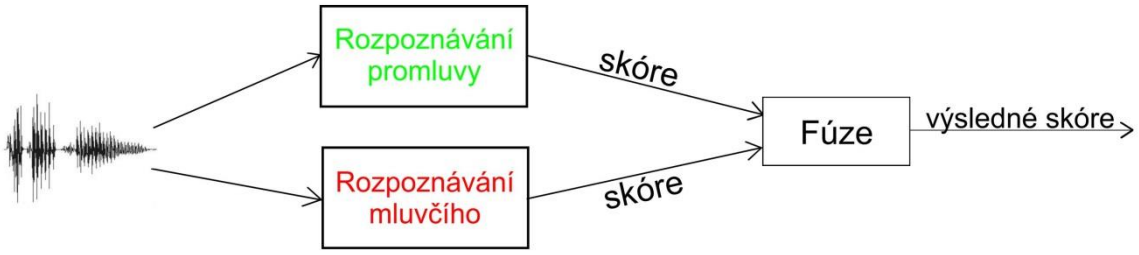
Je jasné, že není možné postihnout v implementaci rozpoznávače všechny zmíněné znaky charakterizující řeč mluvčího, a proto je třeba zaměřit se na ty významnější, jejichž kombinací jsme schopni docílit dobrého výsledku. Prakticky se jedná o takové znaky, které se často vyskytují v přirozené řeči, jsou lehce měřitelné, nejsou závislé na hluku v pozadí ani na přenosových charakteristikách. Protože nejsme schopni postihnout všechny tyto možnosti, je třeba volit podle zkušeností. Obecně platí, že přístupy vyvozené z řečového spektra bývají v automatických systémech nejefektivnější.

Některé znaky ani postihnout nemůžeme, protože se s časem mění, nebo k nim může dojít nenadále. Jedná se například o onemocnění mluvčího, kdy se mu dočasně změní struktura jeho hlasového ústrojí (např. rýma, vypadnutý zub, nebo popálený jazyk). Proto se jimi nebudeme zabývat.

2 Teorie

Textově závislé rozpoznávání mluvčího v sobě obsahuje dvě podúlohy, jak již bylo zmíněno v kapitole 1, které je dobré řešit odděleně. Jsou jimi *Rozpoznávání textu* a *Rozpoznávání mluvčího*. Chceme-li dosáhnout dobrých výsledků, je žádoucí, aby byl systém postaven na více různých přístupech.

Systém tedy bude vypadat podobně, jako blokové schéma na obrázku [Obr. 3]. Systém pro rozpoznávání promluvy i systém pro rozpoznávání mluvčího dostanou na vstup tutéž nahrávku. Provedou svou část rozpoznávání a na výstup vrátí vyhodnocené skóre pro danou nahrávku. Tato dvě skóre budou následně sloučena pomocí fúze, viz Fúze dvou systémů. Z fúze dostaneme výsledné skóre rozpoznávání, které by mělo být lepší, než jednotlivé dílčí skóre obou slučovaných systémů. Výsledky fúze budou následně vyhodnocovány.



[Obr. 3] Blokové schéma celkového systému

Pro rozpoznávání textu je možné použít více různých metod. Nejčastější je lokalizace jednotlivých fonémů v promluvě. Ty můžou být použity přímo k porovnávání s referencí, nebo mohou být následně zpracovány jako text - Detekce klíčových slov. Další variantou je vzít jenom pravděpodobnosti výskytů daných fonémů v určitém čase (posteriogram) a dále s ním pracovat například pomocí algoritmu DTW. Všechny tyto části budou v této práci prozkoumány, aby byla nalezena optimální metoda. Pro rozpoznávání mluvčího bude využit produkční systém firmy Phonexia, viz kapitola 2.1.

2.1 Rozpoznávání mluvčího – Speaker ID

Tato kapitola je věnována teoretickému popisu použitého systému pro rozpoznávání mluvčího. Je zde nastíněna základní funkčnost programu a jeho úspěšnost. Také je popsáno jisté úskalí, které plyne z jeho úspěšnosti pro nepříliš dlouhé nahrávky.

Speaker Identification System (dále jen SID systém) od společnosti Phonexia je profesionální rozpoznávací systém, který dokáže porovnat dvě zadané nahrávky mezi sebou a určit, zda se jedná o téhož mluvčího. Nebo jej lze použít pro ověření, že testovanou nahrávku namluvila osoba z nějaké větší množiny oprávněných žadatelů uložených v databázi. Systém je založen na přepisu řeči do hlasových stop - *voice-printů*. Tyto voice-printy jsou získány extrakcí ze zvukového záznamu. Pomocí voice-printů lze jednoznačně popsat mluvčího a odlišit jej od ostatních. Odlišnosti mezi mluvčími, které se při porovnávání voice-printů projeví, jsou dány především fyzickým uspořádáním hlasového traktu neboli nízko-úrovňovými příznaky. Závisí ale také například na použitém mikrofonu, prostředí, ve kterém je nahrávka pořizována, apod. Systém je také jazykově a textově nezávislý. [5]

Trénování \ Testování	10s	30s	60s	90s	120s	150s
10s	13.55	9.87	8.06	8.06	7.82	7.78
30s	9.56	6.31	5.10	4.71	4.67	4.55
60s	8.12	5.09	4.13	3.73	3.64	3.60
90s	7.63	4.72	3.82	3.60	3.48	3.45
120s	7.51	4.66	3.78	3.54	3.41	3.37
150s	7.49	4.60	3.77	3.49	3.39	3.34

[Obr. 4] Tabulka úspěšnosti systému SID v závislosti na délce nahrávek [17]

Úspěšnost SID systému u krátkých promluv není optimální. Doporučená délka nahrávek je totiž alespoň 10 sekund řeči pro trénovací i testovací nahrávky. V tabulce na obrázku [Obr. 4] jsou uvedeny hodnoty EER při použití různých délek nahrávek, tak jak je u svého systému udává společnost Phonexia. Je vidět, že u systému není příliš počítáno s krátkými nahrávkami. Využité nahrávky z databáze MIT jsou ovšem dlouhé průměrně 0,46 sekundy. Proto není systém samostatně příliš použitelný a je lepší jej využít jako součást textově závislého systému, což znamená ověřovat i obsah promluvy.

Program je k dispozici jak v klasické GUI (Graphical User Interface) verzi, tak i ve verzi pro příkazovou řádku. Jelikož je tato práce založena na experimentování s danými systémy, byla použita verze pro příkazový řádek z důvodu nutnosti spouštět program v dávkách a jako součást bash skriptů.

Program je složen ze tří částí [5]:

1. *vpextract* – nejprve je třeba zpracovat vstupní nahrávku a vytvořit z ní voice-print. To provede program vpextract spuštěný s parametry - konfigurační soubor s informacemi o převodu, soubor se vstupní nahrávkou a výstupní soubor, do kterého bude uložen voice-print.
2. *vpinfo* – tento program slouží hlavně pro zobrazení a kontrolu voice-print souborů.
3. *vpcompare* – hlavní část programu SID je skrytá zde. Jedná se o porovnávač voice-printů, jehož hlavním výsledkem je skóre rozpoznání, které je později využito pro zjištění efektivnosti metody. Vstupním parametrem je tedy opět konfigurační soubor a dva voice-printy porovnávaných nahrávek. Výstupním parametrem je potom soubor, kam bude uloženo skóre.

2.2 Rozpoznávání textu

Tato kapitola v sobě shrnuje teoretický popis systémů a metod využitých pro rozpoznávání promluvy. Ty budou následně aplikovány pro rozpoznávání mluvčího, nebo zlepšení skóre stávajícího SID systému. Nejprve je popsán Fonémový rozpoznávač, který se jevil jako nejuniverzálnější nástroj. Je také uveden příklad rozpoznání promluvy tímto systémem a doplněno je i teoretické vysvětlení některých pojmů. Také zde byl zařazen krátký popis programu HResults. Dále následuje popis systému Key Word Spotting s příkladem rozpoznání klíčového slova. Jako poslední je uveden algoritmus DTW jako ideální nástroj pro srovnávání nahrávek, které jsou různé délky, nebo uvnitř promluvy skrývají nějakou variabilitu. Kapitolu uzavírá příklad srovnání dvou různě dlouhých promluv se stejným obsahem a od stejného mluvčího.

2.2.1 Fonémový rozpoznávač

Jedná se o systém vyvinutý na Fakultě Informačních technologií VUT v Brně skupinou BUT Speech@FIT[1]. Je vyvinut především za účelem výzkumu a lze aplikovat na rozmanité úlohy, jako např. identifikace jazyka, indexování a vyhledávání ve zvukových nahrávkách a také pro vyhledávání klíčových slov.

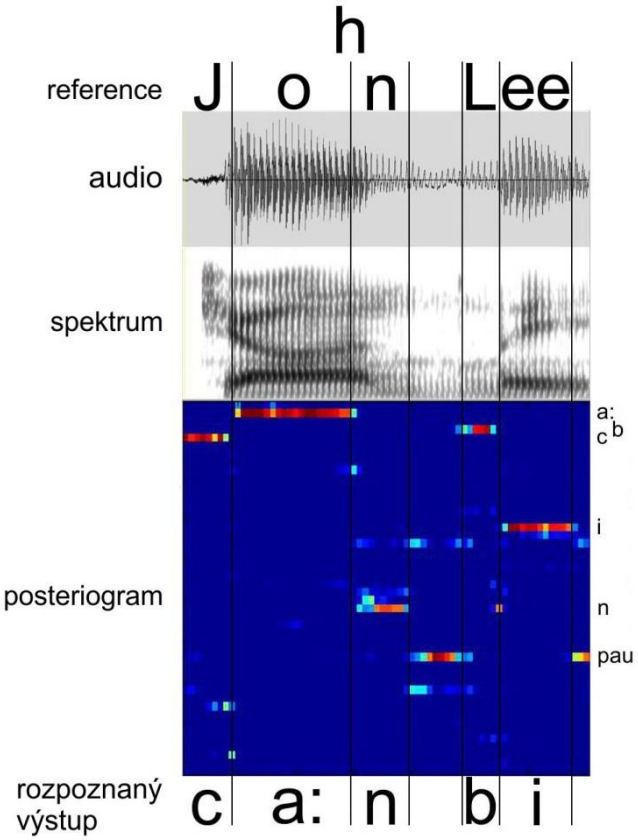
Ke klasifikaci využívá *Neuronové sítě*. Kontext extrahovaný z řeči je rozdělen na levý a pravý, což umožní přesnější modelování při poměrně malé velikosti modelu. Na tyto části je aplikována diskrétní kosinová transformace a jsou pro ně natrénovány dvě neuronové sítě tak, aby produkovaly pravděpodobnosti jednotlivých fonémů v promluvě (*posteriogramy*). Třetí neuronová síť je spojením předchozích a produkuje konečnou množinu pravděpodobností (výsledný posteriogram).

[1] Ke stažení zde: http://speech.fit.vutbr.cz/files/software/phnrec_v2_21.tgz

Pro dekódování fonémového řetězce je použit *Viterbiho algoritmus* bez použití jazykového modelu. [3]

Posteriogram

Posteriogram je matice pravděpodobností výskytu fonémů daného jazyka v čase. Příklad posteriogramu vygenerovaného fonémovým rozpoznávačem pro promluvu „John Lee“ je na obrázku [Obr. 5]. Tento obrázek se skládá z pěti částí. V horní části je referenční zápis promluvy. Následuje zvukový záznam řeči, spektrogram a posteriogram vygenerovaný fonémovým rozpoznávačem. Na konci je fonémový přepis referenční promluvy, který také vygeneroval fonémový rozpoznávač. Svislé čáry určují místo, kde došlo ke změně fonému v promluvě. Na vodorovné ose posteriogramu jsou vyneseny vzorky z promluvy po deseti milisekundách a hodnoty jeho svislé osy odpovídají jednotlivým fonémům. Místa, kde je zjištěna největší pravděpodobnost výskytu daného fonému jsou v posteriogramu zobrazena červeně. Čím výraznější je červená barva bodu v obrázku, tím si je systém více jist, že na daném místě promluvy je pronesen daný foném. Obrázek také obsahuje dvě části neoznačené žádným fonémem. Tyto části reprezentují pauzu mezi slovy „John“ a „Lee“ a šum na konci promluvy.



[Obr. 5] Příklad výstupu fonémového rozpoznávače pro promluvu „John Lee“

Fonémový rozpoznávač však negeneruje jenom posteriogram. Jeho častěji používaný výstup je tzv. *Master Label File* (dále jen MLF soubor), který obsahuje výpis fonémů v daném zvukovém souboru. Pro každý foném je také zaznamenána časová značka pro počátek a konec výskytu a také je mu přiřazeno skóre, kterým systém vyjadřuje, jak moc si je daným fonémem jist. Příklad MLF souboru pro promluvu „John Lee“ je na obrázku [Obr. 6]. MLF soubory se dále vyhodnocují pomocí dalších programů. V této práci byl využit program *HResults*.

```
000000 800000 c -7.532771
800000 3000000 a: -22.954567
3000000 3900000 n -13.576040
3900000 4800000 pau -14.812462
4800000 5400000 b -8.259445
5400000 6700000 i -18.154503
6700000 8200000 pau -23.145821
```

[Obr. 6] Příklad MLF souboru vygenerovaného fonémovým rozpoznávačem

HResults

HResults je jedním ze souboru HTK nástrojů (Hidden Markov Models Toolkit). Je určený pro analýzu výstupu z rozpoznávačů. Je to dobrý nástroj, když je třeba zkontrolovat, jak dobře rozpoznávač funguje a jaké dává pro vstupní data výsledky. Program HResults čte MLF soubory a porovnává je mezi sebou. Výsledkem je potom statistika, která obsahuje zarovnané fonémy získané z obou vstupních souborů, informaci o procentuální *přesnosti* (Acc) a *správnosti* (Corr) a také speciální hodnoty *Insertion, Deletion* a *Substitution*. Tyto hodnoty udávají, kolik bylo v porovnávaném fonémovému přepisu *vloženo, smazáno* a *nahrazeno* fonémů oproti přepisu referenčnímu. HTK Toolkit je po registraci dostupný ke stažení ze zdroje [9] se všemi potřebnými nástroji. [10]

V této práci byl fonémový rozpoznávač využit hned dvěma způsoby. Jednak jako systém pro nalezení fonémů v promluvě a vygenerování příslušných MLF souborů, které byly následně zpracovány a vyhodnoceny programem HResults a dále byl systém využit jako nástroj pro vygenerování posteriogramů pro dané vstupní promluvy. Tyto posteriogramy dále sloužily jako vstup pro systém DTW. MLF soubory i posteriogramy byly sestrojeny pro referenční i testovací promluvu a pomocí programu HResults, nebo algoritmu DTW byly mezi sebou srovnávány.

2.2.2 Detekce klíčových slov v řeči

Další z prozkoumaných systémů byla *Detekce klíčových slov v řeči (Key Word Spotting,* dále jen KWS*)*. Použil jsem KWS od společnosti Phonexia. Tento systém je založen na technologii *umělých neuronových sítí*. Využívá akustický model pro přístup k řečovému signálu. Tento přístup

zajišťuje dobrou rychlost zpracování nahrávek při zachování optimální úspěšnosti detekce klíčových slov. Systém KWS podporuje jazyky: čeština, slovenština, angličtina, ruština, maďarština a polština a umožňuje nastavovat práh pro detekci a úpravy variant výslovnosti klíčového slova [1].

Samotné rozpoznávání probíhá tak, že při trénovací fázi je systému zadán soubor klíčových slov, která budou v nahrávkách použita. Tato klíčová slova jsou následně rozdělena do jednotlivých *fonémů* – elementárních částí slov, které mají v daném řečovém signálu daného jazyka ještě rozlišovací schopnost. Ve vstupních nahrávkách je potom vyhledáváno požadované klíčové slovo. Výstupem je potom seznam rozpoznaných klíčových slov s časy výskytu v nahrávce a také pravděpodobnost, s jakou bylo klíčové slovo rozpoznáno.

Jako první krok se provede v každé nahrávce rozpoznání klíčových slov, které jsou v ní obsaženy, a následně je vyhodnocovala *hodnota důvěry* (*confidence*) každého z klíčových slov. Dále se spočítá výsledná důvěra pro celou promluvu. Předložené promluvy systém klasifikuje do samostatných složek *detected, not_detected* a *rejected* podle toho, zda klíčová slova úspěšně rozpoznal.

```
5900000 8600000 mint -0.328512 0.001322 1 0
```

[Obr. 7] Výstupní soubor pro promluvu „Mint chocotale chip“

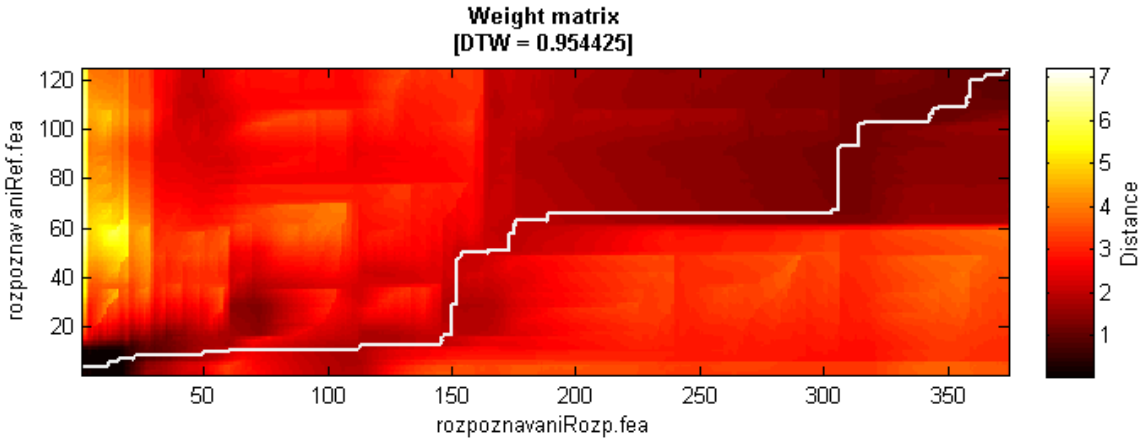
Na obrázku [Obr. 7] je zobrazen výstupní soubor pro jednu z promluv „Mint chocolate chip“ obsažených v MIT databázi. Výstupní soubor má pouze jeden řádek, což znamená, že systém rozpoznal pouze jedno klíčové slovo. Z obrázku je důležitý hlavně první a druhý sloupec, který značí místo, kde bylo v promluvě klíčové slovo nalezeno, tedy *počáteční a koncový čas*. Třetí sloupec je ono rozpoznané *klíčové slovo* a v pátém je uvedena hodnota confidence, tedy jak moc si je systém jist, že rozpoznané slovo je správně. Tato hodnota se může pohybovat v rozsahu od 0 do 1, přičemž 1 znamená největší jistotu, že dané slovo je správně rozpoznáno. Když se podíváme na hodnotu na obrázku, které se velmi blíží nule, je jasné, že systém si příliš jist není. Další klíčová slova systém v nahrávce neobjevil vůbec.

2.2.3 Dynamické borcení času (Dynamic Time Warping)

Princip *Dynamického borcení času* (dále jen DTW) umožňuje měřit podobnost mezi jednotlivými promluvami, i když nejsou promluvy časově zarovnané. Redukování této variability je dosaženo *Dynamickým programováním*. Podobnost těchto dvou promluv se určuje jako jejich *vzdálenost*. [4]

Promluvy jsou reprezentované maticemi parametrů, mezi kterými se hledá optimální srovnávací cesta. Vstupem DTW algoritmu jsou tedy tyto dvě matice, vstupní a referenční, které mezi sebou mají určitou variabilitu. Ta může vzniknout posunutím nahrávky v čase oproti referenční (ticho na začátku), nebo například různou rychlostí promluvy. Na obrázku [Obr. 8] je

příklad srovnání dvou nahrávek. Na svislé ose jsou vyneseny vektory referenční matice a na vodorovné ose jsou vektory vstupní matice. Obě promluvy obsahují slovo „rozpoznávání“. Referenční promluva je však pronesena rychle a promluva, která je rozpoznávána je pronesena velmi pomalu a táhle („rooozpoooznáávání“). Z obrázku vyplývá, že obě promluvy začínaly i končily stejně. Najdeme zde dvě dlouhé, téměř vodorovné čáry, které reprezentují písmeno „o“. U referenční promluvy bylo písmeno řečeno tak rychle, že pro jeho reprezentaci stačil jeden, nebo jen několik málo vektorů, kdežto pro rozpoznávanou táhlou promluvu je vektorů spousta. Algoritmus DTW přesto srovnal promluvy a několik málo referenčních vektorů písmene „o“ se přiřadilo ke spoustě vektorů v promluvě rozpoznávané. Křivka zobrazená mezi vektory udává cestu minimální vzdálenosti.



[Obr. 8] Srovnání referenční a vstupní promluvy pomocí DTW [4]

Minimální vzdálenost a průběh cesty mezi oběma srovnávanými promluvami se určí pomocí *částečných kumulovaných vzdáleností*. Tyto vzdálenosti jsou vypočítány pro každý element srovnávané matice vektorů a pro každý je také určena minimální vzdálenost z předchozích uzlů. Nakonec je vypočítána *minimální normovaná vzdálenost*, která udává, jak moc jsou si referenční a testovací promluva podobné. Čím menší je výsledná vzdálenost, tím podobnější si jsou.

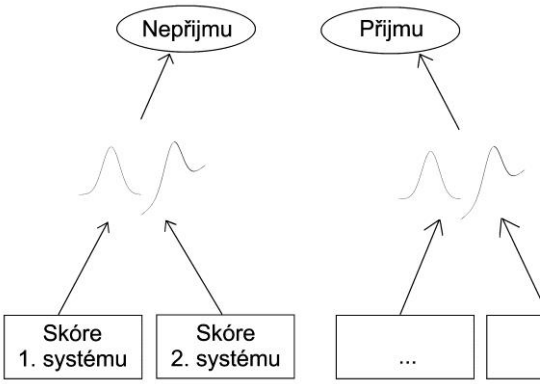
2.3 Fúze dvou systémů

Možností, jak spojit výsledky více systémů dohromady, je spousta. V této práci jsem si vyzkoušel jednak jednoduché spojení pomocí *Mean and variance normalizace*, ale také pokročilejší metody pomocí *Gaussovských modelů* a *Neuronové sítě*. Cílem bylo zjistit, jakou metodou je možné nejlépe spojit výsledky rozpoznávání dvou různých systémů. Pro lepší pochopení využitých metod věnuji kapitoly 2.3.1 a 2.3.2 krátkému popisu Gaussovských modelů a Neuronových sítí. Pro pochopení principu fúze s využitím Mean and variance normalizace postačí

nástin v kapitole 4.2.1. Odkazy na zdroje využité pro inspiraci k popisu metod jsou uvedeny v hranatých závorkách.

2.3.1 Směs Gaussových rozdělení

Fúze dvou systémů znamená spojení výsledků ze dvou využitých systémů tak, že výstupem bude jen jedno skóre popisující konečnou úspěšnost rozpoznávání, podle které se rozhodne, zda bude nahrávka akceptována či nikoliv. Zavedeme si tedy dvě třídy, do kterých se pokusíme výsledky vždy napasovat. Budou to třídy *Přijmu nahrávku* a *Nepřijmu nahrávku*. Tyto třídy budou modelovány pomocí směsí Gaussových křivek, neboli v angličtině *Gaussian Mixture Models* (dále jen GMM).



[Obr. 9] Diagram fúze

V diagramu na obrázku [Obr. 9] vidíme sestavení fúze. Nejprve budou získány skóre dílčích dvou systémů pro všechny trénovací nahrávky. Z těchto skóre bude vytvořen požadovaný počet Gaussových křivek (v diagramu zakresleny dvě) pro každou třídu. Tyto třídy určí předlohu pro modely rozpoznávaných nahrávek. Čím blíže má model do určité třídy, tím větší nebo menší skóre bude mít.

Většinou se GMM využívají v biometrických systémech jako parametrické modely. V úlohách rozpoznávání mluvčího můžou být těmito parametry například charakteristické prvky hlasového traktu mluvčího. Parametry GMM jsou odhadovány z trénovacích dat pomocí speciálních algoritmů.

Pro odhad parametrů modelu, λ, využívá tato fúze tzv. *Metodu maximální věrohodnosti*, anglicky *Maximum likelihood* (dále jen ML). [15]

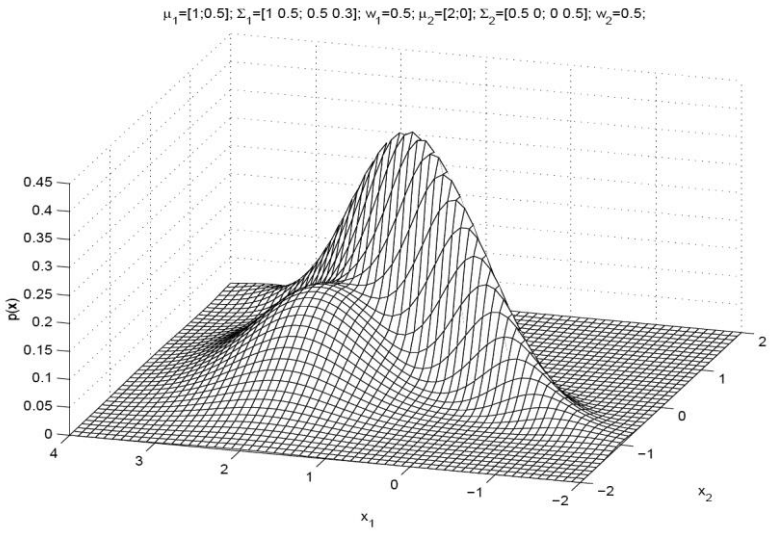
$$p(X|\lambda) = \prod_{t=1}^{T} p(x_t|\lambda) \qquad (2)$$

Proměnná $\boldsymbol{x_t}$ značí trénovací vektory, které jsou v intervalu $X = \{x_1, \ldots, x_t\}$. Rovnice pochází ze zdroje [13].

Model je tedy váženou sumou hustot Gaussovských rozložení,

$$p(X|\lambda) = \sum_{i=1}^{M} w_i \, g(x|\mu_i, \Sigma_i) \qquad (3)$$

kde $\boldsymbol{w_i}$ jsou váhy a $\boldsymbol{g(x|\mu_i, \Sigma_i)}$ jsou části hustot. Rovnice pochází ze zdroje [13]. Příklad GMM složeného ze dvou Gaussovských rozložení je na [Obr. 10].



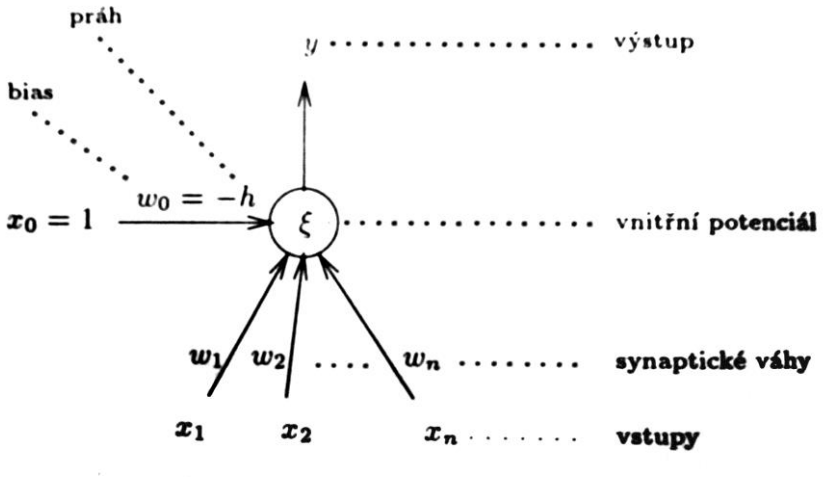
[Obr. 10] Příklad směsi Gaussových křivek [14]

2.3.2 Neuronové sítě

Neuronové sítě jsou velkým fenoménem při modelování učících se systémů. Neuronová síť je vlastně jakýmsi modelem lidského nervového systému plného neuronů, které jsou schopny přenášet a zpracovat vzruchy putující živým organismem. Tento způsob předávání vzruchů v mozku byl předlohou pro vytvoření něčeho podobného pro umělé systémy. Jako první se tato myšlenka objevila v práci Warrena McCullocha a Waltera Pittse v roce 1943 [6]. Praktický význam však zatím neměla, protože zde ukázali pouze možnosti, jak lze pomocí neuronové sítě počítat funkce (logické i aritmetické). S praktickým využitím nepočítali. Velmi to však inspirovalo další vědce k bádání nad možným využitím těchto neurnových sítí.

Matematický model neuronu [7]

Funkce nervového systému člověka a neuronu samotného posloužila jako inspirace pro modelování. Analogicky k biologickému neuronu byl zaveden pojem *formální neuron*, který má stejné vlastnosti jako biologický. Má vstupní terminály *x*, které přijímají na vstup tzv. *váhové hodnoty w*, jeden terminál výstupní, *y*, a uchovává si vnitřní stav tzv. *vnitřní potenciál neuronu*, značený ξ. Struktura formálního neuronu je znázorněna na obrázku [Obr. 11].



[Obr. 11] Formální neuron [7]

Výstup *y* je potom určen velikostí vnitřního potenciálu neuronu při dosažení tzv. *prahu, h.*

Neuronová síť [7]

Neuronová síť využívaná ve výpočetních systémech je modelem nervové soustavy člověka a její struktura z ní tedy vychází. Jako byly v nervové soustavě receptory, efektory a mezi nimi projekční dráhy, zde budeme hovořit o *vstupních, výstupních* a *skrytých (mezilehlých)* neuronech, které jsou pospojovány stejným způsobem jako biologické. Výstup jednoho neuronu je vstupem pro dalších několik neuronů. Celkový počet neuronů určuje tzv. *architekturu* neuronové sítě, hodnoty vah určují tzv. *konfiguraci* sítě. Síť se během přenosu informací vyvíjí, mění se hodnoty vah i vnitřní potenciál jednotlivých neuronů. Tím probíhá učení.

Největší využití neuronových sítí je při úloze rozpoznávání obrazců. Pro účely fúze dvou systémů se však jeví také jako velmi vhodné, neboť jejich schopnost učení by měla zajistit, že síť bude na nerozhodné vstupy reagovat správným způsobem, neboť bude mít toto chování naučené z předchozího natrénování. Nejpoužívanějším typem je *vícevrstvá síť* s *backpropagation algoritmem* (algoritmus zpětného šíření chyby), která byla využita i v této práci.

3 Experimentální setup

V první části následující kapitoly je popsána databáze nahrávek, která je použita pro trénování a testování všech využitých systémů. Dále je popsán způsob, jakým jsou jednotlivé experimenty vyhodnocovány a jaký konkrétní nástroj je pro to využit. Poslední část této kapitoly je věnována popisu způsobu, jímž jsou trénovány systémy rozpoznávání mluvčího, promluvy a jak je potřeba trénovací data set upravit, když se chystáme zabývat fúzí systémů.

3.1 Databáze – MIT

Testovací databáze *MIT Mobile Device Speaker Verification Corpus* (dále jen MIT databáze) pochází z university *Massachusetts Institute of Technology*. Ke stažení je dostupná po registraci[2]. Tato databáze obsahuje nahrávky několikaslovných spojení, jako například „Carol Owens“, nebo „Mint Chocolate Chip“ nahrané od různých mluvčích různým nahrávacím zařízením a v různě charakteristickém prostředí (hlučné, tiché,…). Postihne tudíž všechny možné případy, které mohou při ověřování mluvčího nastat.

Databáze nahrávek obsahuje dvě *enroll* sezení. Jsou to části databáze se stejným obsahem nahrávek od stejných mluvčích, ale nahraných v jiný čas. Proto jsou promluvy vždy řečeny trochu jinak. První sezení je použito k trénování zkoumaného systému a druhé sezení je použito jako testovací množina, která je po natrénování systému ověřována. Každé sezení sestává z 22 mluvčích ženského pohlaví a z 26 mluvčích mužského pohlaví. Každý mluvčí má nahráno 54 nahrávek. Některé obsahují různé promluvy, některé jsou stejné a liší se pouze použitým nahrávacím zařízením. Databáze je vytvořena tak, aby žádná promluva nebyla obsažena pouze jednou. To by pro trénování ani testování systémů nebylo příliš směrodatné. U každého souboru s nahrávkou je přiložen textový soubor, který obsahuje popis promluvy, mluvčího, v jakém prostředí byla pořízena a jakým nahrávacím zařízením.

Databáze ještě obsahuje soubor nahrávek podvodníků (*Impostors*), ty však v systému nebyly využity, neboť obsahují nahrávky jiných mluvčích a promluv a celkově mají jiné rozložení nahrávek. Je jich také méně.

3.2 Evaluace – DET křivka

Detection Error Tradeoff křivka (dále jen DET křivka) se v úlohách rozpoznávání mluvčího používá k ohodnocení detekčního systému. V podstatě podle něj vyhodnocujeme, jak dobře systém funguje. Každý rozpoznávací systém má dva chybové výstupy. *Miss detection* (nesprávné přijetí) a

[2] http://groups.csail.mit.edu/sls/downloads/mdsvc/downloads.cgi

False alarm detection (nesprávné odmítnutí). Případ nesprávného přijetí nastává tehdy, pokud mluvčím je podvodník, který se snaží do systému dostat, ale systém jej vyhodnotí jako oprávněnou osobu a přijme jej. U případu nesprávného odmítnutí je naopak ověřovaným mluvčím osoba, která skutečně je tím, za koho se prohlašuje, systém ji však vyhodnotí jako podvodníka. V DET křivce je využito uniformní rozložení pro oba typy chyb. [2]

Křivka zobrazuje škálu možných operačních charakteristik daného systému. Na křivce lze nalézt jeden významný bod, pomocí kterého lze určit úspěšnost rozpoznávání daného systému. Tento bod se nazývá *Equal Error Rate* (dále jen EER) a nachází se v takovém místě křivky, kde se míra nesprávného přijetí rovná míře nesprávného odmítnutí. V praxi se tato hodnota využívá, pokud chceme systém jednoduše ohodnotit jedním číslem. Čím nižší je tato hodnota, tím lépe pracuje daný systém. Nejlepším porovnáním systémů je ale křivka sama o sobě, protože je na první pohled vidět, jak se systém chová pro různé prahy (operační body). Čím více je křivka posunuta směrem k levému dolnímu rohu grafu, tím lépe systém funguje. Dalším významným bodem křivky je *Decision cost function* (dále jen DCF). Je to rozhodovací funkce definovaná podle rovnice (4),

$$DCF = C_{Miss} * P_{Miss|Target} * P_{Target} + C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * (1 - P_{Target}) \quad (4)$$

kde C_{Miss} je cena nesprávného přijetí, $C_{FalseAlarm}$ je cena nesprávného odmítnutí a P_{Target} je předem daná pravděpodobnost cíle. Pro výpočty hodnoty DCF při vyhodnocování systémů byla hodnota parametrů C_{Miss} a $C_{FalseAlarm}$ volena „1“ a hodnota parametru P_{Target} „0,5“. Rovnice (4) byla převzata z pramenu [8].

Pro vykreslování výsledků systémů do DET křivek jsem použil nástroj DETware verze 2.1 od NIST (National Institute of Standards and Technology). Nástroj je dostupný pro použití s programem MATLAB[3].

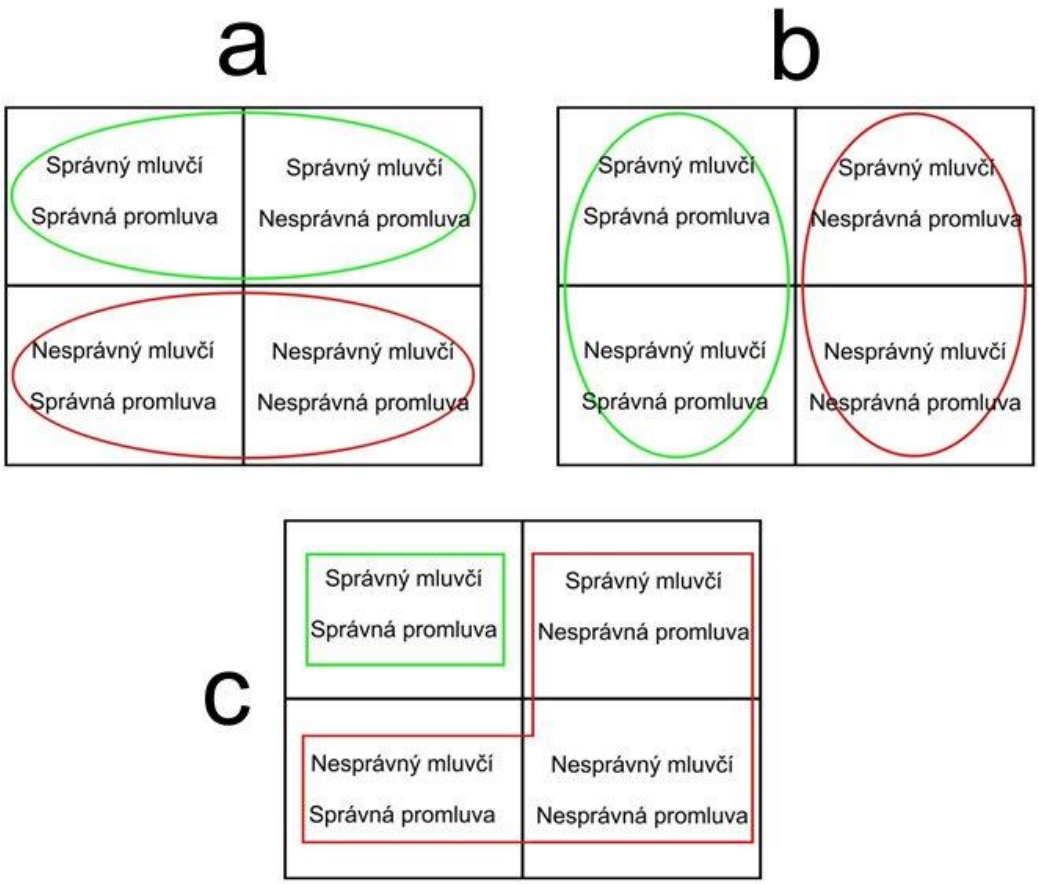
3.3 Definice problému textově závislého rozpoznávání mluvčího

Při ověřování, zda daná osoba má, nebo nemá získat přístup do dané budovy, můžeme vstupní podmínky rozdělit do čtyř kvadrantů podle toho, zda heslo řekl člověk, který má pro vstup oprávnění, a zda řekl správnou promluvu. Na následujících obrázcích značí zelená barva, že systém akceptuje osobu a povolí jí přístup. Červeně jsou zase vyznačeny případy, kdy systém žadatele o přístup odmítne. Při testování úspěšnosti systému je třeba rozdělit testovací nahrávky podle druhu systému, abychom mohli porovnat, zda systém skutečně uděluje dostatečně vysoké skóre nahrávkám, které mají být přijaty, a dostatečně nízké skóre těm, které mají být odmítnuty. Díky tomuto rozdělení je také možné vypozorovat tzv. *práh rozpoznávání*. To je hodnota, která udává

[3] http://www.itl.nist.gov/iad/mig/tools/DETware_v2.1.targz.htm

mez. Když dostane porovnávaná nahrávka skóre nižší než je práh, bude žadatel o přístup odmítnut. Pokud bude hodnota vyšší než práh, bude přijat.

Systémy verifikace mluvčího budou úspěšné, pokud výsledné skóre padne do horních dvou kvadrantů [Obr. 12a] a systémy verifikace promluvy zase tehdy, pokud jejich skóre padne do levých dvou kvadrantů [Obr. 12b]. Ani jedno ale nestačí. Pro úspěšné rozpoznání potřebujeme spojit výhody obou tak, aby byl systém úspěšný jenom tehdy, pokud se o vstup do systému hlásí správný mluvčí po pronesení správné promluvy, jinými slovy potřebujeme vytvořit průnik obou systémů. Proto se v dalším textu budu zabývat fúzí systémů. Fúze tedy bude úspěšná pouze v případech, kdy její výsledné skóre padne do levého horního kvadrantu [Obr. 12c].



[Obr. 12] Rozdělení úlohy textově závislé rozpoznávání mluvčího podle metody

4 Experimenty

Tato kapitola se věnuje vyhodnocení jednotlivých systémů popsaných v kapitole 1. Nejprve jsou v části 4.1 rozebrány výsledky využitých metod a jsou předloženy grafy s DET křivkami, které reprezentují, jak systém funguje. Systémy byly testovány více způsoby. V prvé řadě byli odděleni muži a ženy, aby bylo možno u systémů vypozorovat, zda například není některý z přístupů nekonzistentní pro různé pohlaví. Pro tento případ byly mužské i ženské nahrávky testovány systémem zvlášť. Dále byly všechny systémy otestovány zvlášť s nahrávkami rozdělenými podle správnosti mluvčích (oprávněný žadatel/podvodník) a podle správnosti promluv (správná promluva/chybná promluva). To je z důvodu, abychom byli schopni vyhodnotit zvlášť rozpoznávání mluvčího a rozpoznávání promluvy. Každý výsledek je reprezentován vlastním grafem. Jsou také spočítány body DCF a EER, viz kapitola 3.2, jejichž hodnota je zaznamenána v tabulce na konci každé podkapitoly. Podle těchto dvou hodnot se nejčastěji posuzuje úspěšnost dané metody. Důležitý je však i samotný průběh grafů, proto je nelze vynechat. DET křivky v grafech jsou uváděny většinou po čtveřicích, kde plná čára znázorňuje křivky pro rozpoznávání mluvčího a přerušovaná pak rozpoznávání promluvy. Barevné odlišení křivek se týká pohlaví mluvčích, kde červená barva je využita pro výsledky systému pouze s ženami a modrá pouze s muži.

Další část této kapitoly (část 4.2) je věnována fúzím jednotlivých systémů pro rozpoznávání mluvčího a promluvy. Jsou zde prezentovány tři typy fúzí, a pokud je to možné, je daná fúze vyzkoušena s různým nastavením, aby bylo dosaženo co nejlepších výsledků. Jako výsledek fúze opět slouží grafy s DET křivkami a tabulky s konkrétními hodnotami. Grafy zobrazují rozdíl mezi původními systémy a fúzí. Pokud byla fúze vylepšována, nebo bylo experimentováno s různým nastavením, je připojen i srovnávací graf pro posouzení, zda vylepšení přineslo viditelné výsledky.

Poslední částí je porovnání přístupů, které byly pro fúzi systémů použity, a vyhodnocení, která z nich dávala pro toto zadání nejlepší výsledky.

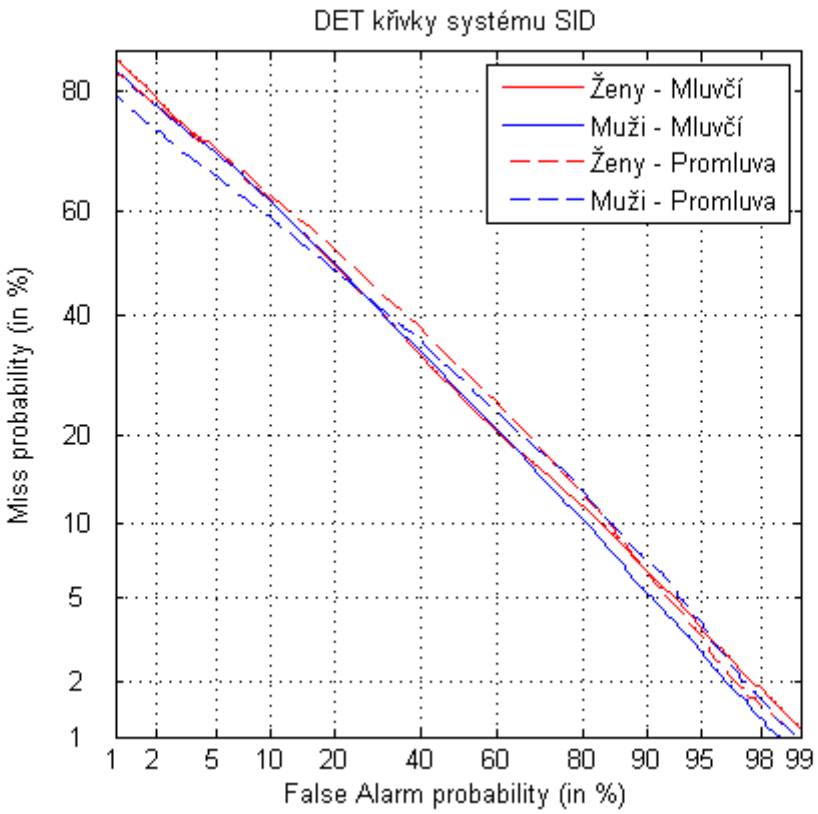
4.1 Vyhodnocení samostatných přístupů

Nyní následují popisy experimentů s jednotlivými systémy, které byly v rámci práce prováděny, a jejich výsledky. Následně jsou podle výsledků vybrány nejlepší systémy pro rozpoznávání mluvčího a promluvy. Tyto jsou pak v kapitole 4.2 využity pro fúzi.

4.1.1 Speaker Identification System – SID

Systém SID od společnosti Phonexia byl jediným testovaným systémem zaměřeným na rozpoznávání mluvčího. Je již úspěšně a s dobrými výsledky používán, takže jsem upustil od zkoumání dalších metod, neboť je velmi nepravděpodobné, že by bylo dosaženo lepších výsledků.

Výsledky vyhodnocení systému jsou viditelné v grafu na obrázku [Obr. 13]. Využití systému pro rozpoznávání mluvčího je zobrazeno plnou čarou, pro rozpoznávání promluvy naopak přerušovanou. Křivky jsou si hodně podobné a není příliš velký rozdíl, zda se jedná o muže, nebo ženy. Také je vidět, že pro rozpoznávání mluvčího má o něco lepší výsledky než pro rozpoznávání promluvy. To vychází i ze samotné podstaty tohoto systému, který je vyvinut především pro úlohy identifikace mluvčího. Systém tedy funguje podle očekávání. V tabulce [Tab. 1] jsou zobrazeny hodnoty DCF a EER. Hodnota EER je skutečně nižší u úlohy rozpoznávání mluvčího než u promluvy.



[Obr. 13] Výsledky systému SID

Systém	Rozpoznávání mluvčího		Rozpoznávání promluvy	
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**
DCF	0,3452	0,3474	0,3608	0,3396
EER	0,3564	0,3578	0,3839	0,3681

[Tab. 1] Hodnoty EER a DCF pro systém SID

4.1.2 Detekce klíčových slov v řeči

Systém *Key Word Spotting* od společnosti Phonexia se jevil jako ideální nástroj pro rozpoznávání krátkých promluv. Bohužel ale nefungoval při experimentech s MIT databází příliš spolehlivě. Systém promluvu rozdělil na fonémy, a následně vyhledával klíčová slova. V hledání však moc úspěšný nebyl. Větší část promluv systémem neprošla, protože v nich systém neobjevil žádné ze zadaných klíčových slov a řadil je jako odmítnuté (rejected). V takovém případě systém nemohl do výstupního souboru nic uložit, a tedy nevygeneroval žádná data, podle kterých by bylo možné systém porovnávat. Když některá slova nalezl, přiřadil jim tak malé skóre pravděpodobnosti, že by stejně nebylo možné považovat tento výsledek za úspěšné rozpoznávání. Protože použitelných výsledků ze systému bylo velmi málo a ještě se špatným skórem, byl nakonec z experimentů vyřazen, neboť by stejně nemohl být prohlášen za dobrou metodu pro tento typ úlohy.

Důvodem špatné funkce tohoto systému jsou pravděpodobně nahrávky, které nejsou příliš kvalitně nahrané, nebo jsou velmi tiché. Některé jsou zašuměné, nahrávané v různě rušných prostředích, a také nekvalitními mikrofony, a proto zřejmě systém nedetekuje takové fonémy, které by odpovídaly referenčním pro dané klíčové slovo. Záměrem, proč byla využita právě MIT databáze, je ovšem postihnutí všech možných neideálních případů. Pokud tedy systém při horších nahrávkách selhává, nehodí se pro aplikaci v případech sledovaných v této práci. Dalším důvodem pro špatnou funkci tohoto programu v této práci může být i délka nahrávek databáze. Nahrávky jsou totiž extrémně krátké.

4.1.3 Fonémový rozpoznávač

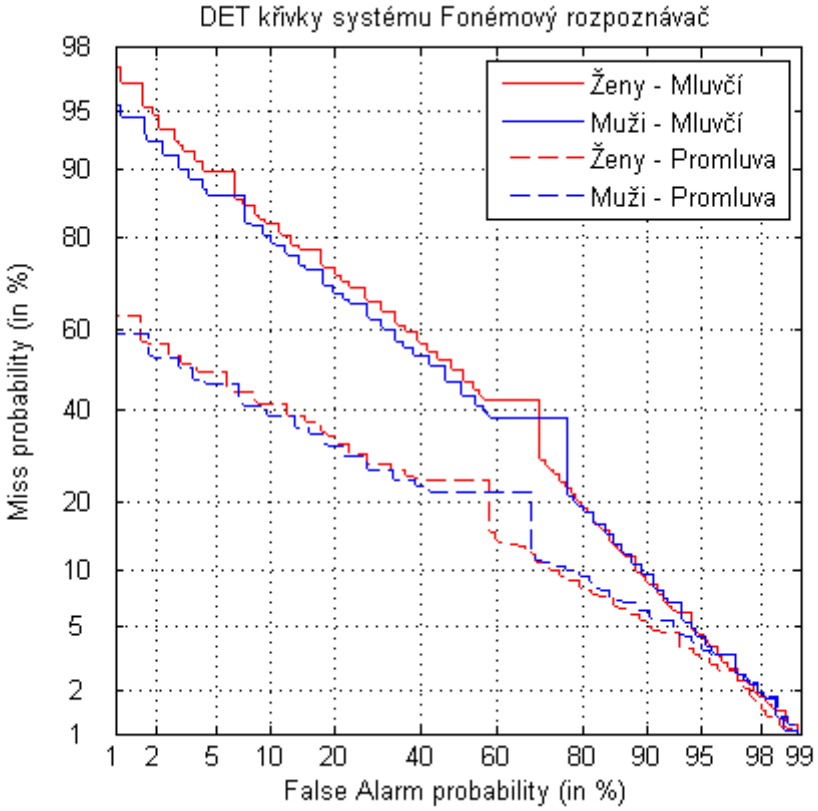
Fonémový rozpoznávač je výzkumný systém, který byl vyvinut právě pro takové testovací účely, kterými je i má bakalářská práce. Navíc jsem k němu jako student fakulty Informačních technologií VUT v Brně měl přístup, a proto bylo jeho využití jasnou volbou. Systém je založen na Neuronových sítích, které se nezřídka při rozpoznávání využívají, a proto bylo zajímavé vyzkoušet, jak si poradí s krátkými heslovitými nahrávkami, kterými se práce zabývá.

Nejprve bylo třeba zvolit slovník fonémů, použitý při rozpoznávání. Fonémový rozpoznávač nabízí český, anglický, maďarský a ruský. I když jsou trénovací a testovací nahrávky v angličtině,

zvolil jsem český slovník, díky kterému bude přepis řeči přesně kopírovat zvukový projev mluvčího, takže budou na první pohled vidět odlišnosti. U anglického slovníku by byl problém, že angličtina se jinak čte a jinak píše, takže systém by rozpoznané fonémy reprezentoval pro nás, Čechy, poněkud nesrozumitelným způsobem a nebylo by na první pohled poznat, že systém byl například chybně spuštěn. Pro lepší názornost uvedu příklad. Anglické slovo „aeroplane“ bude vysloveno jako „érouplejn“. Při použití českého slovníku budou fonémy vypadat podobně jako jednotlivá písmena vyslovené promluvy. Při použití anglického však systém uvede znaky, které pro Čecha nemusí být ihned rozpoznatelné a můžou vést v extrémním případě až k neodhalení chyby. Přepis do fonémů podle anglického slovníku by mohl vypadat takto: „ˈeərəˌpleɪn“. Jiná omezení by z této záměny slovníků neměla nastat, neboť pokud člověk dvakrát vysloví totéž slovo, bude vždy reprezentované stejnými fonémy v tomtéž jazyce. Jen budou vypadat jinak.

Dále systém na vstupu očekává textový soubor s cestami k nahrávkám a na výstup generuje MLF soubor, viz kapitola 2.1, který bude obsahovat výpis vstupních souborů spolu s výpisem fonémů, které byly nalezeny v každém souboru. Protože jsou však soubory databáze s nahrávkami umístěny v adresářové struktuře, bylo třeba ji dodržet i s těmito fonémovými přepisy. Proto byly MLF soubory rozsekány tak, aby v každém souboru byl jen přepis jedné nahrávky a takto byly uloženy do adresářové struktury jako REC soubory (record). Toto bylo provedeno pro trénovací i testovací množinu nahrávek.

V další fázi se porovnávaly přepisy trénovací a testovací nahrávky. K tomuto účelu posloužil *HResults* z HTK Toolkitu, viz kapitola 2.1. Tento nástroj umožňuje vyhodnocení výsledků rozpoznávače. Z míry podobnosti mezi fonémovými přepisy vygeneruje hodnoty přesnosti a správnosti, které jsou následně použity k sestavení grafu dole.



[Obr. 14] Výsledky Fonémového rozpoznávače

Systém	**Rozpoznávání mluvčího**		**Rozpoznávání promluvy**	
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**
DCF	0,4570	0,4406	0,2505	0,2400
EER	0,5009	0,4704	0,2739	0,2678

[Tab. 2] Hodnoty EER a DCF pro Fonémový rozpoznávač

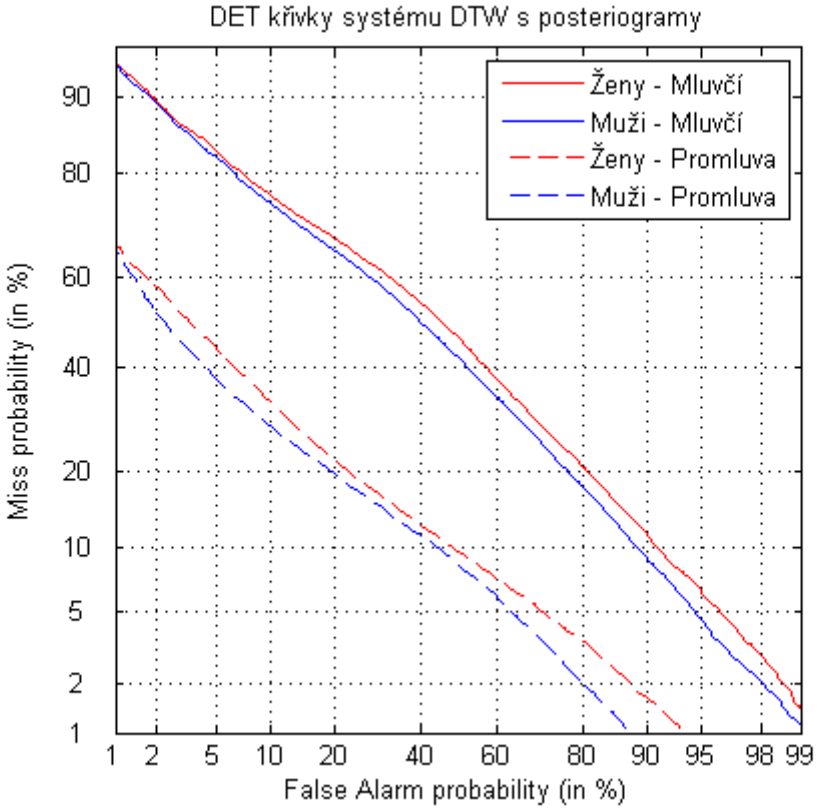
Jak lze vidět v grafu na obrázku [Obr. 14], fonémový rozpoznávač funguje mnohem lépe pro rozpoznávání slova než pro rozpoznávání mluvčího. To je očekávané chování, protože z fonémového přepisu lze jen těžko vyvodit charakteristiky, které by identifikovaly více mluvčích od sebe. Z grafů pro rozpoznávání slova je také patrné, že křivka začíná na ose y na mnohem menších hodnotách, než končí na ose x. To znamená, že systém bude fungovat lépe v levé části křivky při menší hodnotě nesprávného přijetí (Pfa), protože hodnota nesprávného odmítnutí (Pmiss) neporoste tak rychle. Hodnoty EER a DCF jsou uvedeny v tabulce [Tab. 2] a budou později porovnávány s ostatními systémy.

4.1.4 DTW s posteriogramy

Systém DTW využívající na vstupu přepis záznamu řeči do posteriogramů byl prvním testovaným DTW systémem. Pro tyto účely byl využit systém od pana Bc. Miroslava Skácela. Tento systém je navržen tak, že využívá distanční matici, kde se vzdálenosti mezi vektory z posteriogramů porovnávají mezi sebou. Pro výpočet je použita logaritmická kosinová vzdálenost (log-likelihood cosine distance). Dále využívá kumulační matici, ve které se sčítají vzdálenosti mezi vektory z distanční matice, a také matici váhovou, což je znormovaná matice kumulační podle délky cesty. Z váhové matice se najde nejlepší cesta pomocí back-trackingu. Čím nižší hodnotu systém vrátí, tím lépe dopadlo rozpoznávání.

Kvalitu rozpoznávání bude ovlivňovat především to, jak kvalitní byl přepis řeči do posteriogramu (pravděpodobnosti fonémů). Na tuto fázi zpracování řeči byl využit Fonémový rozpoznávač z části 4.1.3.

Systém je opět otestován na MIT databázi a výsledné skóre jsou zobrazeny pomocí DET křivek, které jsou níže. Tak jako u fonémového rozpoznávače je zde vidět, že i systém DTW při využití posteriogramů funguje mnohem lépe pro rozpoznávání slov než pro rozpoznávání mluvčího. Důvodem je právě posteriogram reprezentující obsah dané promluvy. Ten v sobě nenese příliš informací, ze kterých by se dal identifikovat mluvčí, který promluvu pronesl. Rozdíl mezi ženami a muži není příliš velký, nicméně lze pozorovat, že pro muže funguje systém o něco málo lépe.



[Obr. 15] Výsledky DTW systému s posteriogramy

Systém	**Rozpoznávání mluvčího**		**Rozpoznávání promluvy**	
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**
DCF	0,4308	0,4224	0,2040	0,1868
EER	0,4786	0,4542	0,2090	0,1963

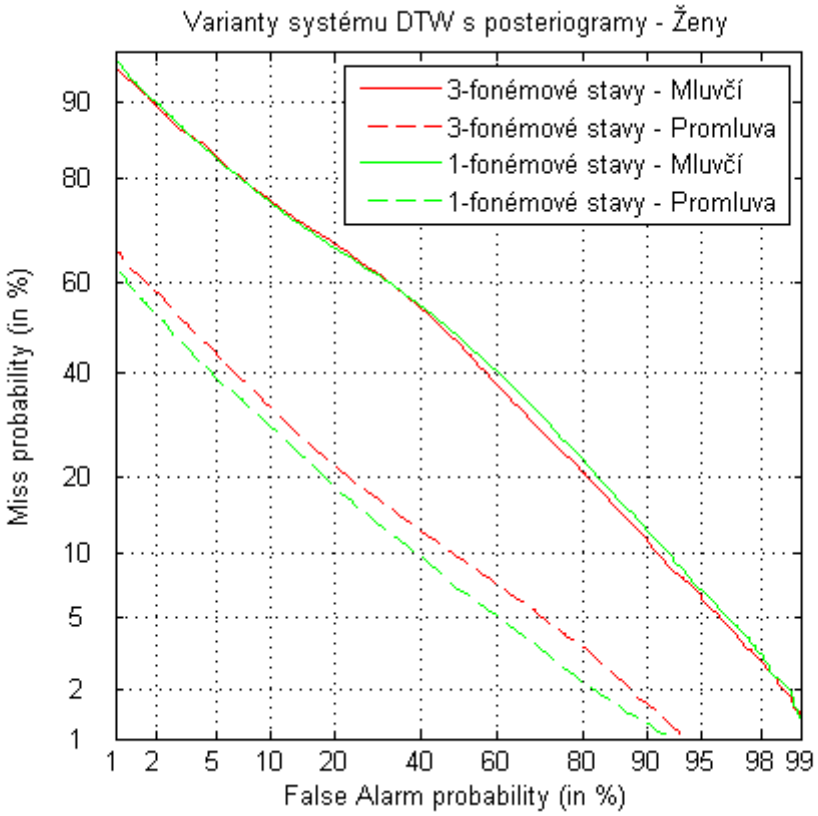
[Tab. 3] Hodnoty EER a DCF pro systém DTW s posteriogramy

Modifikace s jednoduchými fonémovými stavy

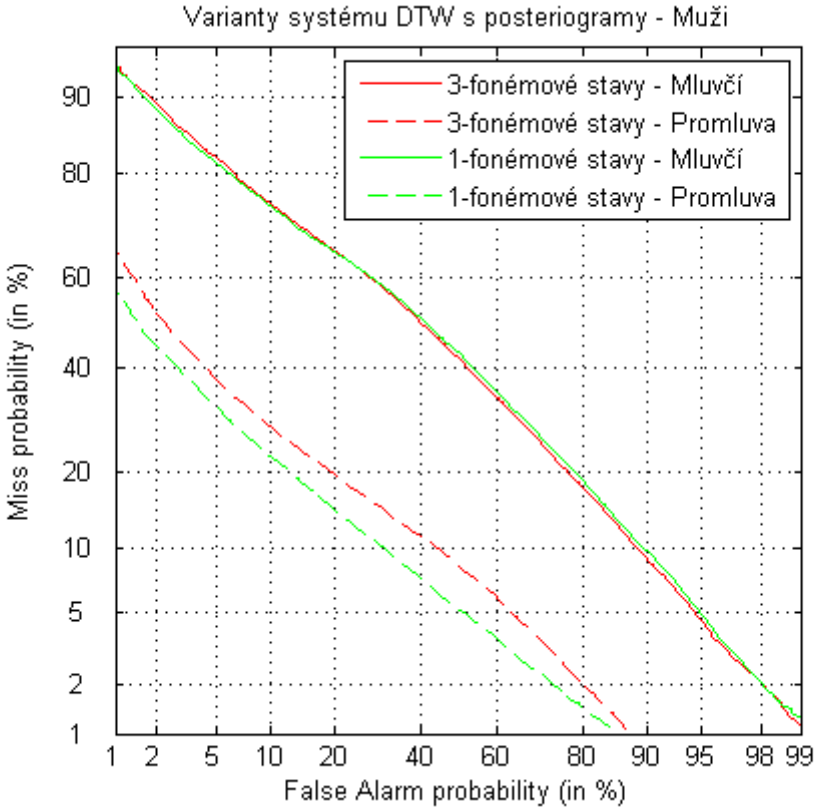
Většinou se pro reprezentaci promluvy pomocí fonémových pravděpodobností používá posteriogram, který má každý foném rozdělen na tři části. Jedná se o začátek, prostředek a konec fonému. Takový tvar mají i posteriogramy použité v systému DTW nahoře. Vzhledem k tomu, že čeština má 36 fonémů, bude fonémových stavů 108. Tento přístup zaručuje vyšší jemnost při klasickém rozpoznávání promluv, pro tuto aplikaci však nemusí být úplně vhodný. Hlavním problémem je to, že nahrávky, které má systém zpracovávat, jsou velmi krátké a navíc jsou použité pro jiný rozpoznávaný jazyk. Jemnější dělení může obecně přinést lepší úspěšnost rozpoznávání

pro češtinu, ale může zhoršit rozpoznávání pro jiné jazyky, jako angličtina. Je to proto, že v angličtině například neexistují některé specifické přechody mezi fonémy, které v češtině jsou. Například začátek písmene „R“ nemusí v angličtině existovat, takže dojde k rozpoznání jako prostředek tohoto písmene. Celé písmeno „R“ pak nebude rozpoznáno, protože systém musí nejprve projít začátek, potom prostředek a konec daného fonému. Pokud však začátek není rozpoznán, nebude rozpoznán ani celý foném. Když ale spojíme stavy do jednoho, tomuto problému se vyhneme. Matice posteriogramu potom na ose *y* reprezentuje přímo jednotlivé fonémy.

Použití jednoho stavu pro jeden foném se podle výsledků ukázalo jako dobré. Úspěšnost oproti původnímu algoritmu ještě trochu vzrostla. Na DET křivkách dole můžeme vidět rozdíly obou systémů. Původní křivka, kdy systém využíval tři stavy pro každý foném, je zobrazena červeně, systém po spojení fonémových stavů do jednoho je zobrazen zelenou křivkou. Pro přehlednost jsou grafy pro muže a ženy odděleny do samostatného obrázku. Pro identifikaci mluvčího se úspěšnost příliš nezměnila, ale pro identifikaci promluvy, o kterou u tohoto systému jde především, je lepší. Proto v dalších aplikacích bude již vstupní posteriogram algoritmu DTW vždy upraven na jednoduché stavy fonémů.



[Obr. 16] Zlepšení výsledků při použití 1-fonémových stavů – Ženy



[Obr. 17] Zlepšení výsledků při použití 1-fonémových stavů - Muži

Systém	**Rozpoznávání mluvčího**		**Rozpoznávání promluvy**	
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**
DCF	0,4284	0,4195	0,1874	0,1615
EER	0,4869	0,4598	0,1901	0,1650

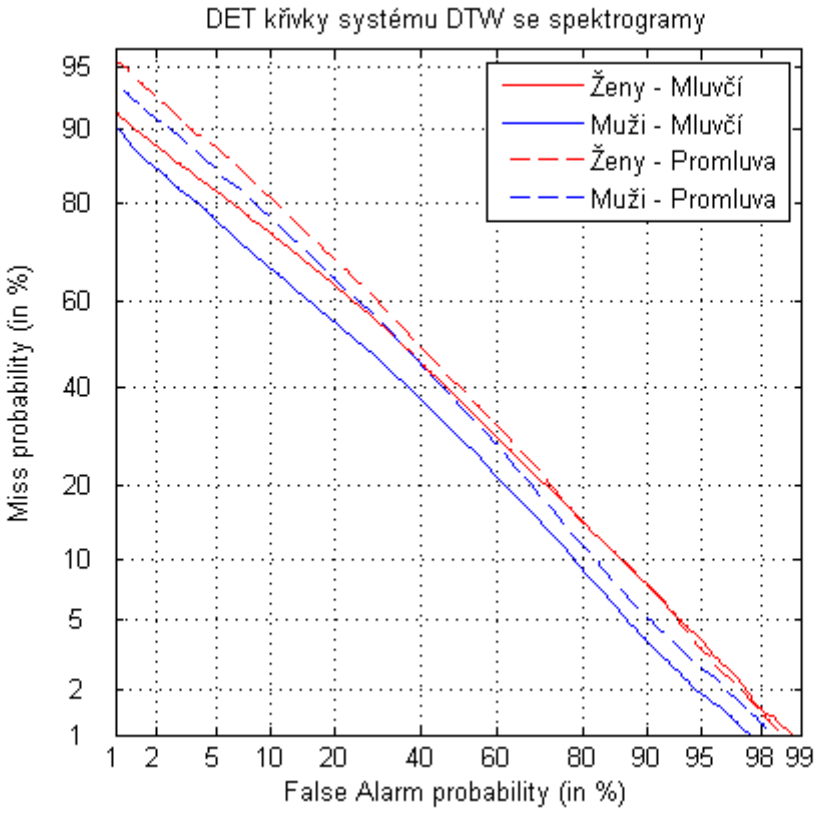
[Tab. 4] Hodnoty EER a DCF pro systém DTW s posteriogramy a s jednofonémovými stavy

Pro lepší důkaz o tom, že je skutečně výhodné využít tohoto spojení fonémových stavů, uvedu ještě relativní zlepšení při aplikaci systému pro rozpoznávání promluvy. Je to 9% pro ženy a pro muže dokonce 15,9%. Pro rozpoznávání mluvčího nemá příliš smysl se zlepšením zabývat, protože skóre je velmi špatné a nepředpokládá se tedy využití systému tímto způsobem.

4.1.5 DTW se spektrogramy

Další možností, jak algoritmus DTW využít, je analyzovat spektrogram řeči namísto posteriogramu. Díky posteriogramu sice získáme přesnější informace o obsahu promluvy, ale příliš nepoznáme, zda se jedná o správného mluvčího. Spektrogram je podobná matice jako posteriogram, rozdíl je však v tom, že na vertikální ose nejsou fonémy, ale zvukové frekvence. Podle frekvencí, které člověk při mluvení využívá, je možné jej snáz identifikovat, protože nejsme omezeni sledováním řeči a hledáním odlišností jen v tom, jak byla promluva pronesena (zda koktal, mluvil pomalu/rychle, natahoval písmeno „á“ apod.), ale můžeme využít i anatomické struktury hlasového aparátu (jak vysoko posazený má člověk hlas, zda mluví dýchavičně apod.).

Tento komplexní přístup však může být i nevýhodou. Pokud by například mluvčí byl nemocný tak, že by se jeho hlas dočasně změnil (rýma) nebo by byl v jiném psychickém rozpoložení než obvykle (unavený, energický), mohl by vzniknout rozdíl ve frekvenční oblasti i když by promluva byla stále stejná. Pokud by však byl systém využíván samostatně a bez dalšího zpracování, je zřejmě lepší použít metodu, která bude respektovat jak mluvčího (frekvenční oblast), tak promluvu (fonémovou oblast). Pro tyto účely byl využit systém ze zdroje [18].



[Obr. 18] Výsledky DTW systému se spektrogramy

Systém	Rozpoznávání mluvčího		Rozpoznávání promluvy	
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**
DCF	0,4162	0,3737	0,4432	0,4218
EER	0,4274	0,3813	0,4472	0,4256

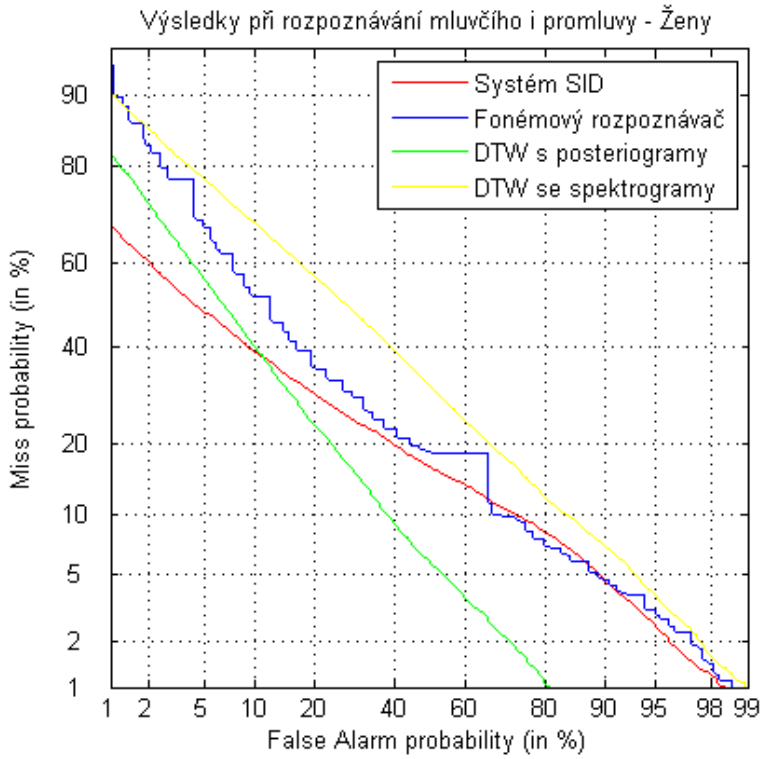
[Tab. 5] Hodnoty EER a DCF pro systém DTW se spektrogramy

Výsledné DET křivky jsou zobrazeny nahoře. Podle očekávání lze vidět, že systém je docela dobře úspěšný pro obě úlohy, jak ověřování mluvčího, tak i promluvy. Cenou za komplexnost je ovšem přesnost. Pokud porovnáme hodnoty EER z DTW systému využívajícího posteriogramy s tímto spektrogramovým DTW systémem, zjistíme, že pro identifikaci mluvčího je nynější systém lepší, o mnoho se však zhoršil pro identifikaci promluvy. Bude proto nejspíš vhodný pro použití tam, kde je zapotřebí využít jeden komplexní systém, který nebude mít příliš vysoké výpočetní požadavky. Pokud bychom však chtěli vyšší přesnost, bude nutné kombinovat více metod dohromady a využít výhod, které mají. Opět lze také pozorovat, že systém je o něco úspěšnější pro muže než pro ženy.

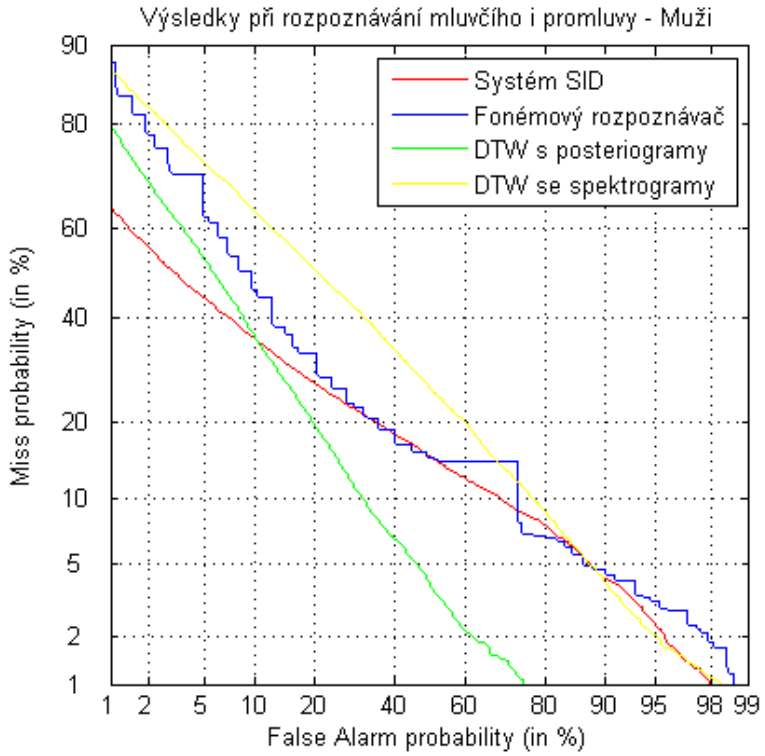
4.2 Fúze systémů

Každý systém má některé výhody a nevýhody. Například systém DTW s posteriogramy se velmi osvědčil pro identifikaci promluvy. Pro identifikaci mluvčího však selhává. Naopak systém SID má o něco větší úspěšnost pro identifikaci mluvčího. Systém DTW se spektrogramy má úspěšnost docela špatnou a z toho důvodu je v této práci nepoužitelný. Fonémový rozpoznávač je konkurentem posteriogramového systému DTW, ale úspěšností jej nepřekoná.

Protože mým cílem bylo dosáhnutí co nejlepšího výsledku a také prozkoumání všech možností, rozhodl jsem se využít výhod, které představují opačně orientované systémy (na mluvčího, na promluvu), a udělat spojení neboli „fúzi“ takových dvou systémů. Pro tyto účely jsem vybral ty systémy, které pro své zaměření dávaly nejlepší výsledky. Je to systém DTW využívající posteriogramy pro identifikaci promluvy a systém SID pro identifikaci mluvčího. Na obrázcích [Obr. 19] a [Obr. 20] jsou zobrazené všechny testované systémy jako křivky v jednom grafu pro lepší přehlednost. Lze se přesvědčit, že dva výše zmíněné systémy opravdu ve své kategorii vyhrály (DTW a SID), neboť oba leží nejblíže k bodu [0;0], což znamená, že dávají nejlepší výsledky.



[Obr. 19] Porovnání testovaných systémů při rozpoznávání mluvčího i promluvy – Ženy

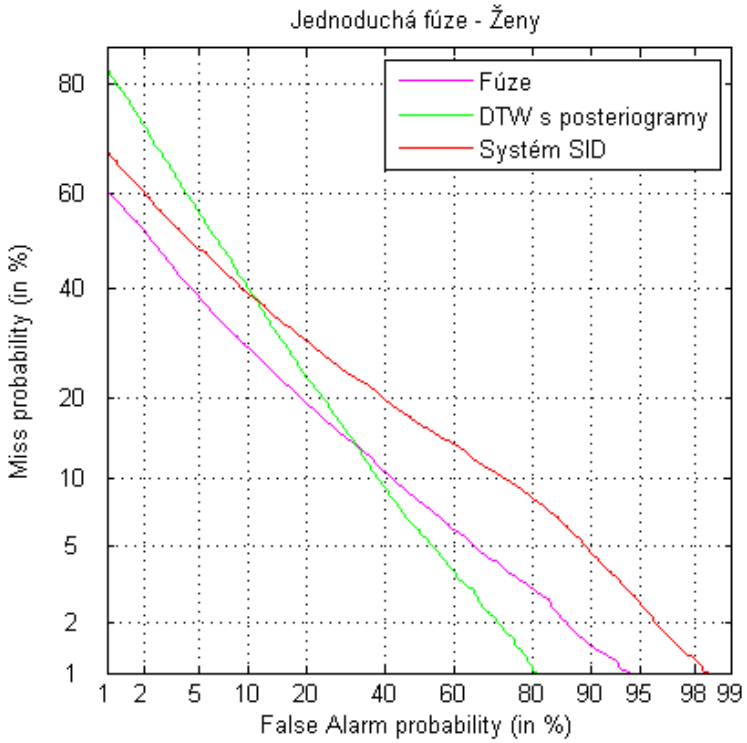


[Obr. 20] Porovnání testovaných systémů při rozpoznávání mluvčího i promluvy – Muži

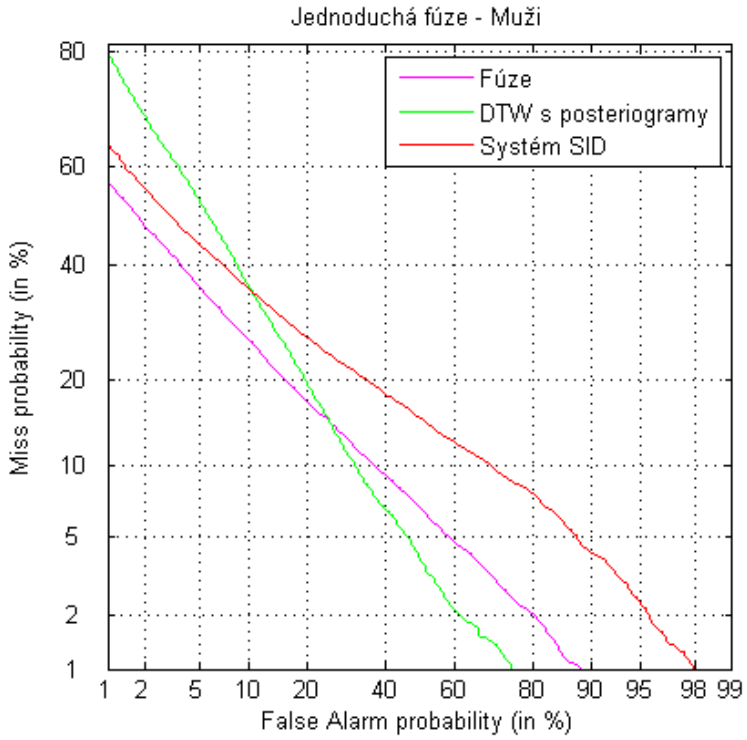
Fúze dvou různě zaměřených systémů je nejlepší způsob jakým docílit, aby systém správně reagoval na různé situace. Takováto fúze již vyžadovala upravit trénovací listy. Dosud bylo pro rozpoznávání mluvčího využíváno rozdělení trénovacích a testovacích nahrávek podle obrázku [Obr. 12a] v kapitole 3.3 a pro rozpoznávání promluvy podle obrázku [Obr. 12b]. Oba druhy systémů však akceptují nebo zamítají jinou množinu nahrávek a tyto množiny se překrývají. Pokud vyjdeme z obrázku [Obr. 12c], bude nyní třeba za oprávněné žadatele o přístup považovat pouze ty, kteří padnou do levého horního kvadrantu. Zbylé tři kvadranty budou nyní znamenat odmítnutí. Grafy na obrázcích [Obr. 19] a [Obr. 20] mají právě takovéto rozložení nahrávek. Je to proto, abychom mohli snadno porovnávat systémy s následujícími fúzemi. Zajímavostí v grafu je bod, ve kterém se grafy kříží. Systém, který je lepší v jedné části grafu, je v další části horší. Fúze by měla využít výhod obou systémů a udělat dobrý systém ve všech částech grafu.

4.2.1 Jednoduchá fúze

První problém, který bylo třeba vyřešit, byl rozsah hodnot systémů. Každý z testovaných systémů funguje na jiném principu a jeho výstupem jsou proto jiné hodnoty. Někdy je nejlepším skórem číslo nejblíže k nule, jindy je to číslo co nejvyšší. Proto bylo třeba takovéto různé rozsahy sjednotit. Tento problém řeší tzv. Globální *Mean and Variance Normalization*. Všechny hodnoty obou systémů jsou pomocí této metody převedeny do stejného rozsahu. Skóre obou systémů takto normalizovaná se poté sečtou jedna ku jedné. Pokud tedy pro každou promluvu existovalo jedno skóre ze systému DTW a jedno ze systému Speaker ID, nyní je pro každou promluvu jedno společné skóre (fúze). Obrázky [Obr. 21] a [Obr. 22] ukazují DET křivky, které porovnávají úspěšnosti této fúze s původními systémy.



[Obr. 21] Jednoduchá fúze s Mean and Variance normalizací s původními systémy – Ženy



[Obr. 22] Jednoduchá fúze s Mean and Variance normalizací s původními systémy – Muži

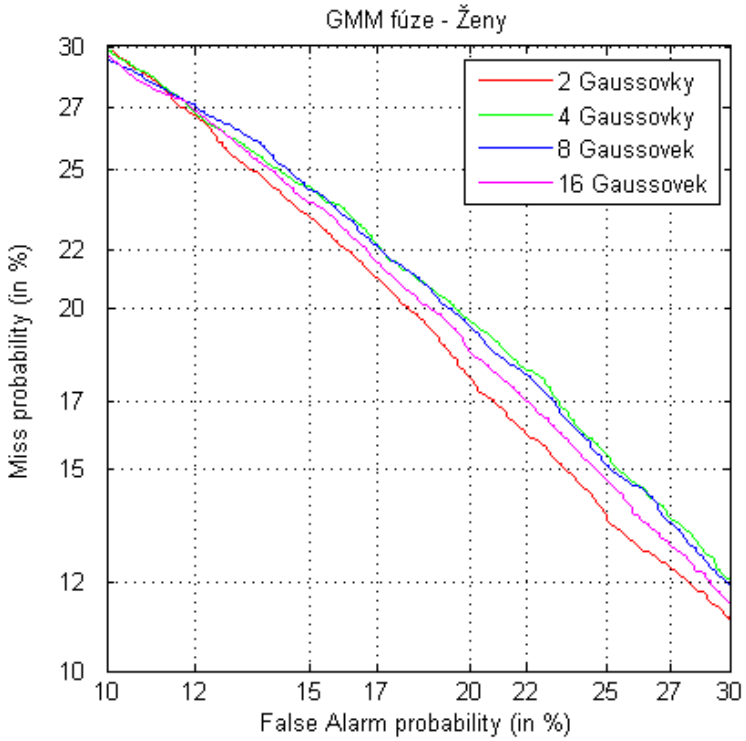
Pohlaví mluvčího	Žena	Muž
DCF	0,1888	0,1735
EER	0,1938	0,1816

[Tab. 6] Hodnoty EER a DCF pro jednoduchou fúzi DTW a SID

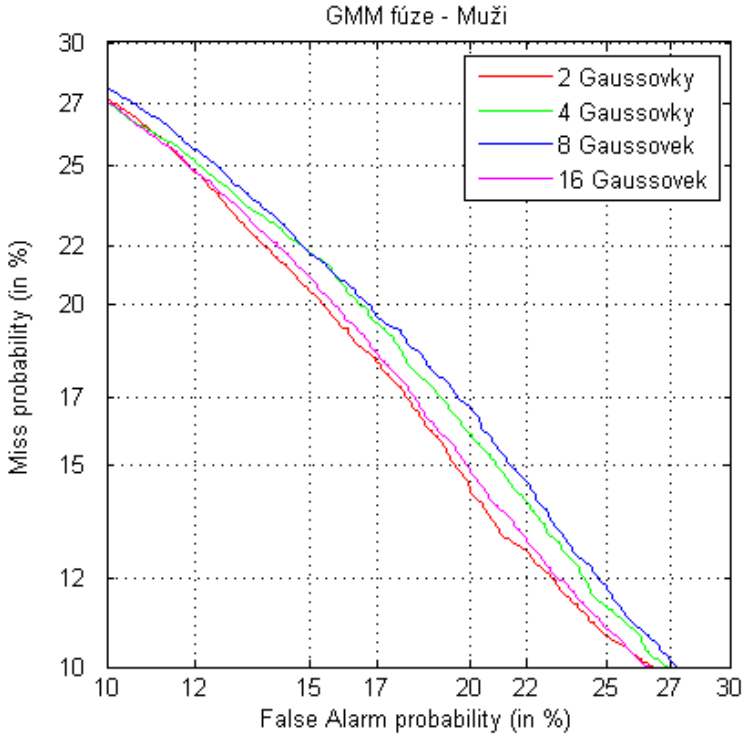
Jak vidíme v tabulce [Tab. 6] a v grafech na obrázcích [Obr. 21] a [Obr. 22], fúze byla poměrně úspěšná. Sloučením dvou různých metod rozpoznávání bylo dosaženo zlepšení skóre (menší hodnota EER a DCF) a přiblížení DET křivky k počátku. Při pozornějším prohlédnutí DET křivky fúze zjistíme, že se tato v jednom bodě kříží s křivkou DTW a dále už pokračuje s horším skóre než samotný systém DTW. Proto je třeba zdůraznit, že použití této fúze má smysl, pokud chceme, aby pravděpodobnost nesprávného odmítnutí byla větší nebo rovna pravděpodobnosti nesprávného přijetí (Pmiss >= Pfa). Při dodržení této podmínky bude systém pracovat jak má tj. lépe než systémy ze kterých je složen. Toto omezení nebude vadit při použití fúze tam, kde se vyžaduje větší zabezpečení za cenu možného odmítnutí oprávněného žadatele. Tyto požadavky můžou být například v bankách. Zde je nesprávné schválení přístupu situací, která se nesmí stát, protože by mohlo hrozit zcizení velkého množství peněz. Opačné požadavky může mít například policie, když se bude snažit pomocí podobného systému zjistit, zda na průkazné nahrávce je hlas některého s obviněných. Pokud systém všechny ihned odmítne jen proto, že je nahrávka zašuměná a nevezme v potaz, že jeden člověk měl skóre shody mnohem vyšší než ostatní, i když to nestačilo na nastavený práh, nebude policii tento systém příliš potřebný.

4.2.2 Fúze založená na Gaussovských modelech

Fúze GMM funguje vlastně jako učící se systém. Teorie této metody je podrobněji popsána v kapitole 2.3.1. Tato fúze by měla být schopná pomocí Gaussovských modelů spojit oba systémy tak, aby výsledná DET křivka věrněji kopírovala oba systémy v celé její délce. U této fúze bude také záležet na tom, kolik Gaussových křivek bude zvoleno. Více křivek znamená jemnější modelování systému, ale jak bylo patrné u fonémových stavů, větší jemnost nemusí vždy znamenat lepší výsledek. Počet Gaussových křivek také určuje složitost modelu. Proto jsem zkoušel systém se dvěma, čtyřmi, osmi a šestnácti Gaussovými křivkami. Cílem bylo zjistit, jestli se budou výsledky výrazně lišit.



[Obr. 23] Výsledky GMM fúze podle počtu Gaussových křivek - Ženy

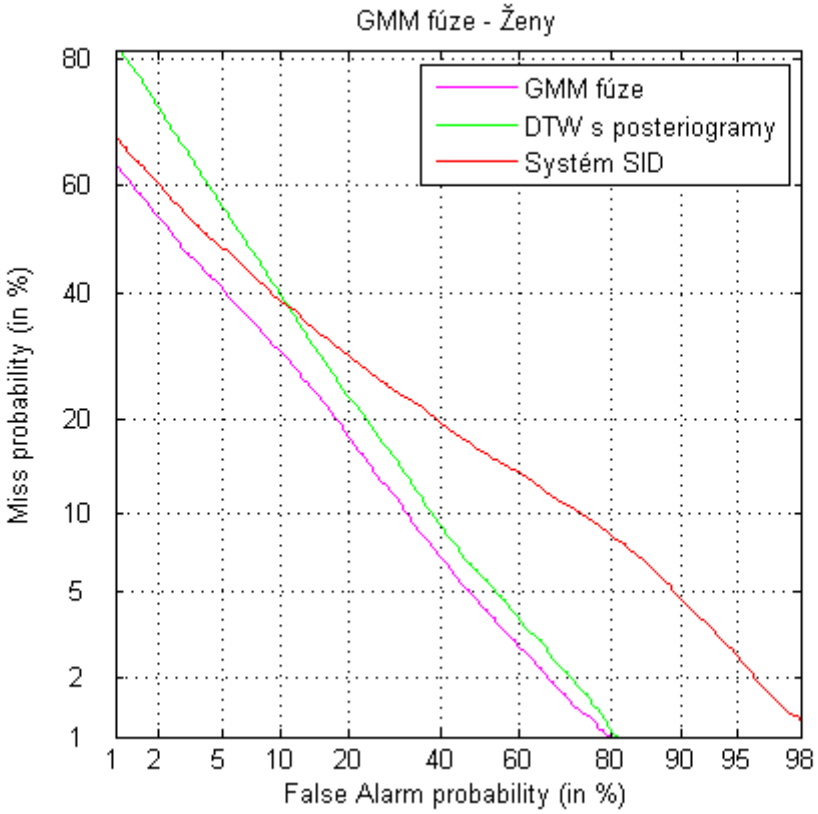


[Obr. 24] Výsledky GMM fúze podle počtu Gaussových křivek - Muži

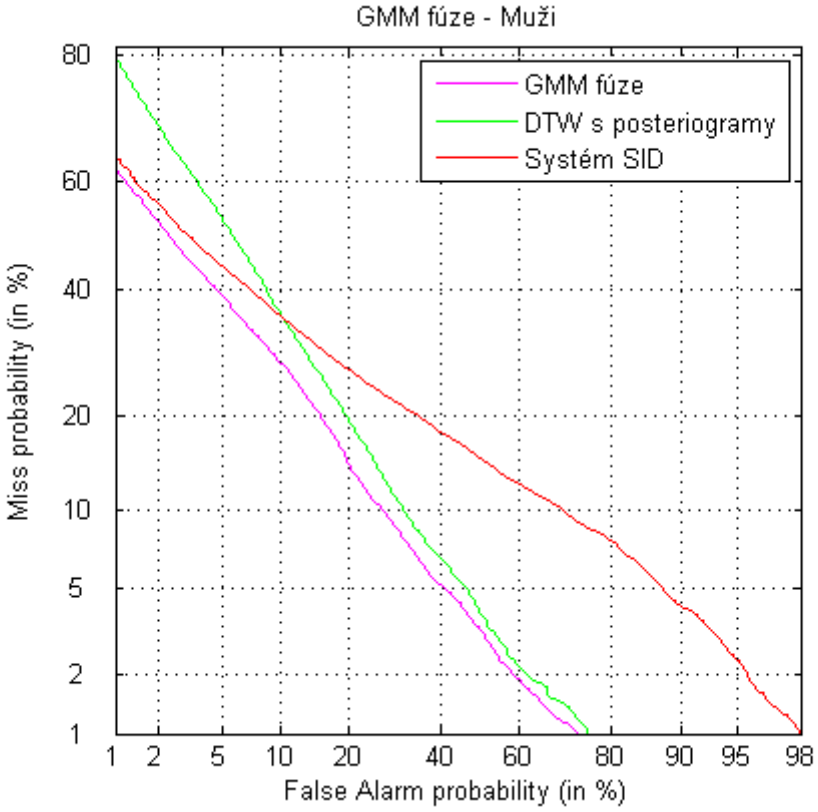
Počet křivek	**2**		**4**		**8**		**16**	
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**	**Žena**	**Muž**	**Žena**	**Muž**
DCF	0,1878	0,1707	0,1950	0,1791	0,1952	0,1816	0,1924	0,1736
EER	0,1896	0,1748	0,1973	0,1798	0,1968	0,1835	0,1938	0,1764

[Tab. 7] Hodnoty EER a DCF pro fúzi pomocí GMM

Z tabulky [Tab. 7] můžeme vyčíst, že nejlepší výsledky dával systém se dvěma Gaussovými křivkami, neboť jsou zde hodnoty Equal Error Rate nejmenší. Velmi blízko je také systém se šestnácti křivkami. V grafech na obrázcích [Obr. 23] a [Obr. 24] vidíme rozdíl mezi jednotlivými křivkami v bližším detailu kolem bodu EER. Protože se dvě Gaussovy křivky osvědčily jako nejlepší parametry GMM fúze pro tuto úlohu, bude fúze v následujícím textu a obrázcích prezentována s tímto nastavením dvou křivek. Na obrázcích [Obr. 25] a [Obr. 26] níže tedy můžeme vidět tuto variantu v porovnání s původními systémy DTW a SID.



[Obr. 25] Porovnání GMM fúze s původními systémy – Ženy



[Obr. 26] Porovnání GMM fúze s původními systémy – Muži

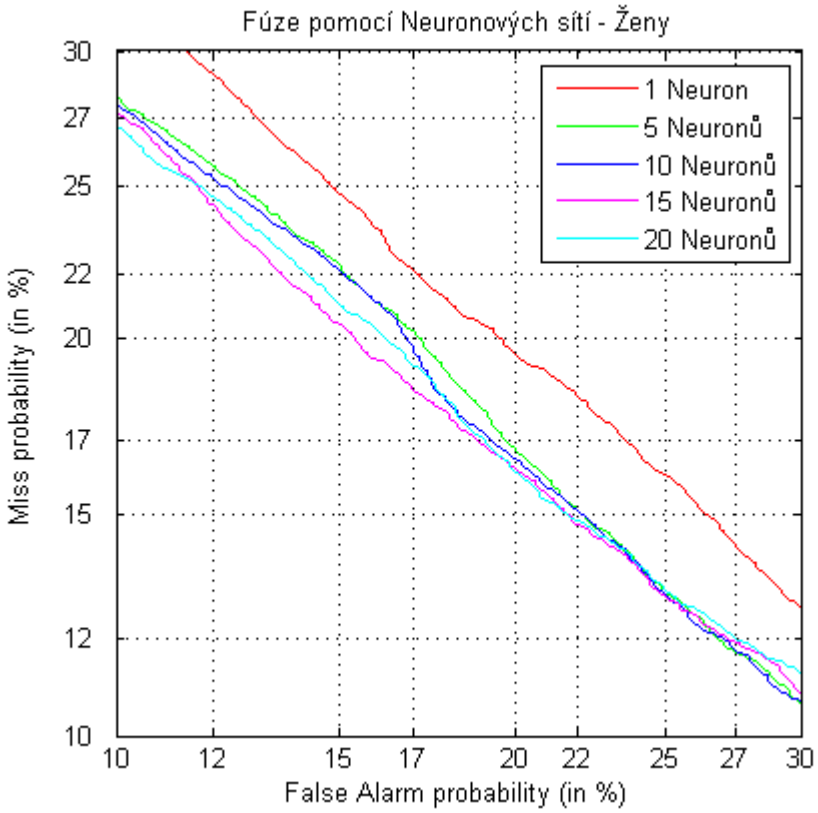
Výsledky pokusu s GMM fúzí jsou docela dobré a podle očekávání tento způsob odstraňuje problém předchozí jednoduché fúze, kde docházelo k překřížení křivky s původním systémem, takže při určitém extrémním nastavení by systém dával horší výsledky než původní. Když ovšem srovnáme tuto GMM fúzi s předchozí, zjistíme, že při rovnosti pravděpodobnosti nesprávného přijetí a pravděpodobnosti nesprávného odmítnutí (nastavení Pfa=Pmiss) má tato fúze o několik tisícin horší skóre. Je to ovšem zanedbatelná hodnota, která je daní za to, že systém v jiných poměrech pravděpodobností Pfa a Pmiss funguje mnohem lépe. Jde hlavně o případ Pmiss < Pfa. V této fúzi skutečně ke křížení již nedochází, a proto vždy dostaneme lepší výsledek, než bychom dostali ze samotného systému DTW nebo Speaker ID. Proto lze fúzi doporučit pro velkou škálu využití. Jako u samotných systémů i zde lze pozorovat o něco lepší výsledky systému pro muže než pro ženy.

4.2.3 Fúze založená na Neuronových sítích

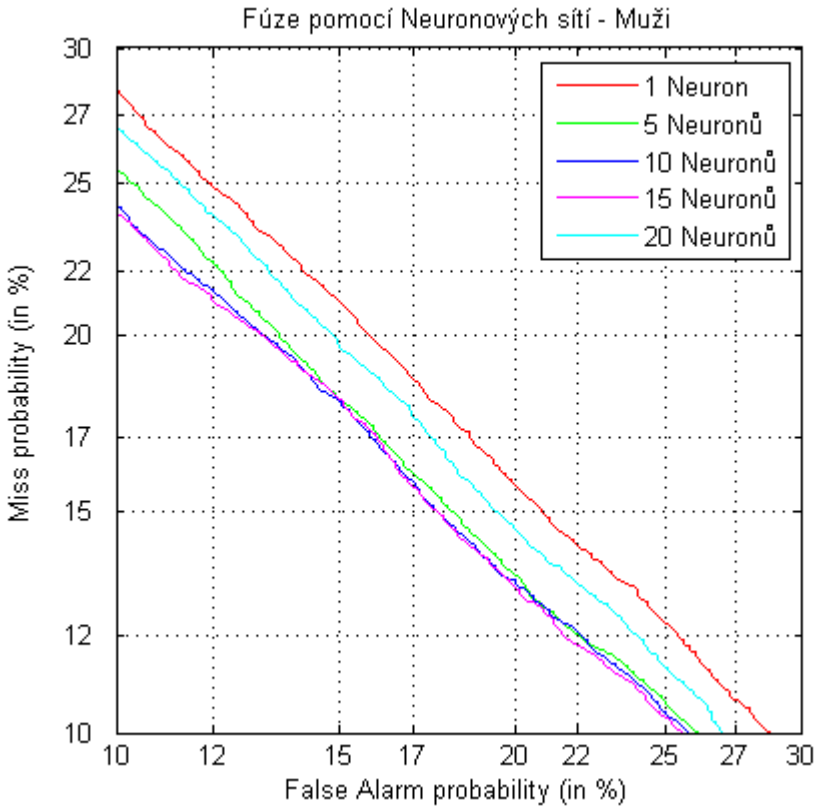
Dalším typem fúze, která by mohla přinášet dobré výsledky je fúze pomocí neuronové sítě. Neuronová síť je vlastně učící se systém navzájem propojených uzlů s váhami. Při trénování

systému dochází k úpravám těchto vah až do stavu, kdy jsou nastaveny na takovou úroveň, že další trénování by už nemělo smysl, neboť by se hodnota vah měnila jen zanedbatelně. Takovýto naučený systém pak může dávat dobré výsledky.

Využití jako fúze dvou systémů není úplně běžné, nicméně je zajímavé vyzkoušet všechny možnosti. U neuronové sítě bylo především důležité zjistit závislost výsledku na nastavení parametrů neuronové sítě. Výsledek může ovlivnit hlavně počet neuronů ve vnitřní „skryté“ vrstvě sítě. Experimenty byly nejprve prováděny se sítí, která obsahovala 10 neuronů. Poté byl tento počet měněn, aby bylo možné vypozorovat, jestli bude mít větší jemnost sítě (větší počet neuronů) zásadní vliv na výsledné skóre. Počet neuronů ve skryté vrstvě jsem postupně měnil na 1, 5, 10, 15 a 20. Na obrázcích [Obr. 27] a [Obr. 28] jsou zobrazené DET křivky pro výsledky s takto nastavenými sítěmi.



[Obr. 27] Výsledky NN fúze podle počtu neuronů ve skryté vrstvě – Ženy



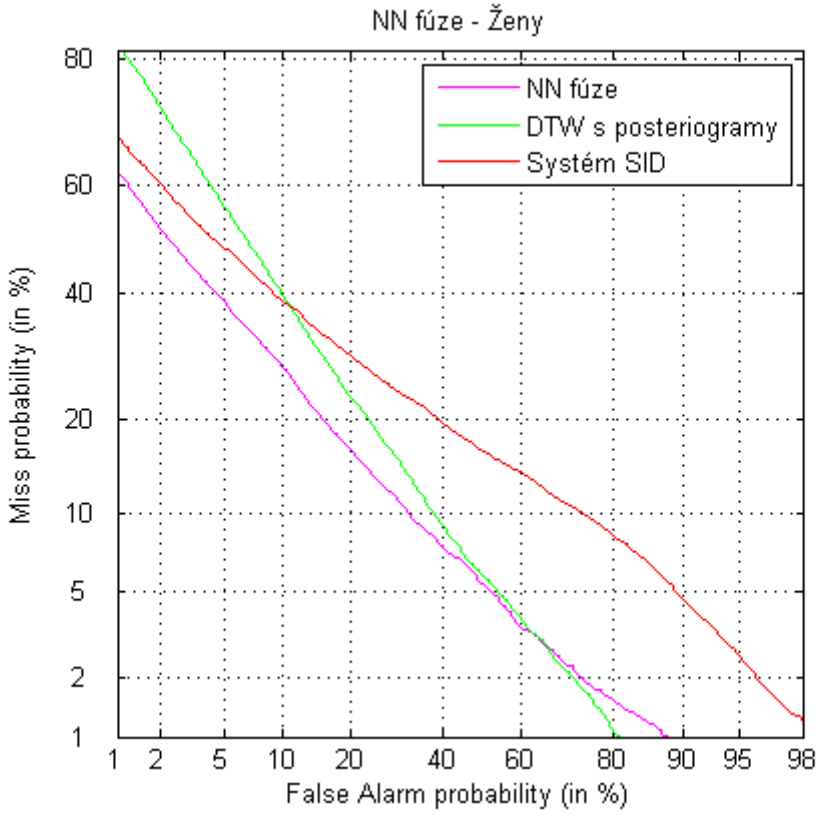
[Obr. 28] Výsledky NN fúze podle počtu neuronů ve skryté vrstvě - Muži

Počet neuronů ve skryté vrstvě	**1**		**5**		**10**	
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**	**Žena**	**Muž**
DCF	0,1952	0,1775	0,1831	0,1647	0,1797	0,1630
EER	0,1971	0,1782	0,1841	0,1650	0,1799	0,1638
Počet neuronů ve skryté vrstvě	**15**		**20**			
Pohlaví mluvčího	**Žena**	**Muž**	**Žena**	**Muž**		
DCF	0,1760	0,1627	0,1793	0,1712		
EER	0,1784	0,1639	0,1805	0,1727		

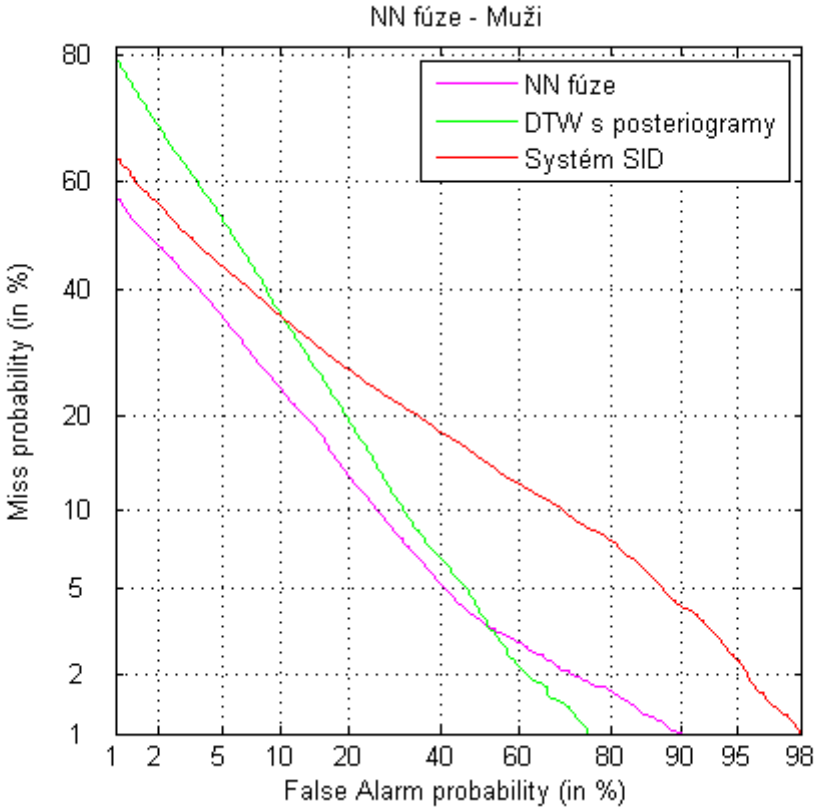
[Tab. 8] Hodnoty EER a DCF pro fúzi pomocí Neuronové sítě

Z předchozích grafů můžeme vidět, že pro tuto aplikaci není potřeba dělat síť zbytečně velikou. Výsledky se výrazněji liší hlavně při použití 1 nebo 5 neuronů. Další přidávání počtu neuronů už má menší efekt. Je tedy potřeba najít hranici mezi velikostí sítě a zlepšením skóre. Zde se zdá, že nejoptimálnější počet je něco mezi 10 a 15 neurony. Větší přidávání už nepřinese přílišné zlepšování, ale spíše nastanou problémy s nutností delšího trénování sítě a delšího vyhodnocování. Přesto však zjištěné rozdíly ve výsledcích nejsou až tak velké. Křivky jsou od sebe vzdálené poměrně málo, přesto je ale na místě se zabývat optimálním nastavením, protože pokud bude systém použit na rozpoznávání hesel či v jiné bezpečnostní aplikaci, i drobné zlepšení je vítané. V tabulce [Tab. 8] jsou zvýrazněné nejlepší hodnoty EER a DCF. Nejlepší jsou opravdu u 15neuronové sítě, jenom EER pro muže je o tři desetitisíciny horší než u 10neuronové sítě, což je zanedbatelné.

Tímto pokusem bylo tedy určeno, že optimální počet neuronů sítě pro tento případ krátkých nahrávek je něco kolem patnácti a síť s tímto počtem neuronů tedy bude v sekci 4.3 využita ke srovnání s ostatními fúzemi. Nelze ovšem říci, že pro každou úlohu rozpoznávání bude 15 neuronů optimální počet. Vždy záleží na vlastnostech, které má systém mít. Například pro rozpoznávání mluvčího z dlouhé promluvy bude přesnější než z krátkých nahrávek, spojované systémy dají skóre, která budou hodnotami blíže u sebe (menší odchylky) a ideální počet neuronů pro fúzi bude jiný. Bude opět potřeba experimenty vyzkoumat, jak velká síť bude správně modelovat tento systém. Záleží také na množství nahrávek dostupných pro trénování.



[Obr. 29] Porovnání fúze s Neuronovými sítěmi s původními systémy – Ženy



[Obr. 30] Porovnání fúze s Neuronovými sítěmi s původními systémy – Muži

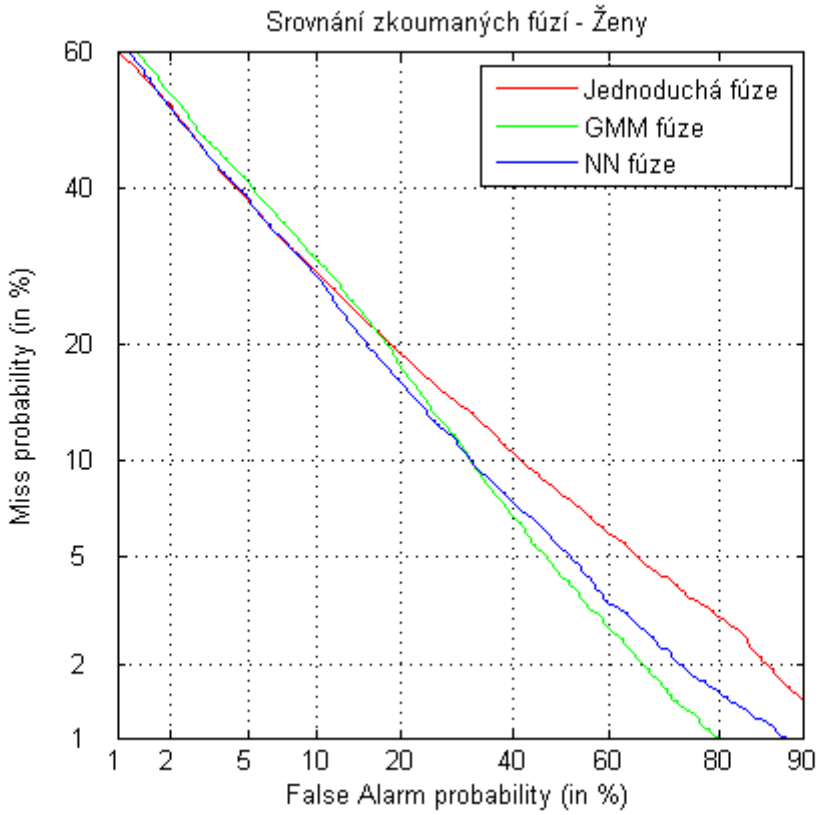
V grafech na obrázcích [Obr. 29] a [Obr. 30] vidíme, že tato fúze funguje nejlépe ve stejných případech jako jednoduchá fúze, a to při Pmiss >= Pfa. Při splnění této podmínky bude fungovat lépe než fúze jednoduchá i GMM. Horší skóre však dává při větší pravděpodobnosti nesprávného přijetí (Pfa). Opět zde dochází ke křížení křivek, a to pravděpodobně z toho důvodu, že systém Speaker ID v této části křivky dává výsledky spíše horší.

4.3 Vyhodnocení fúzí

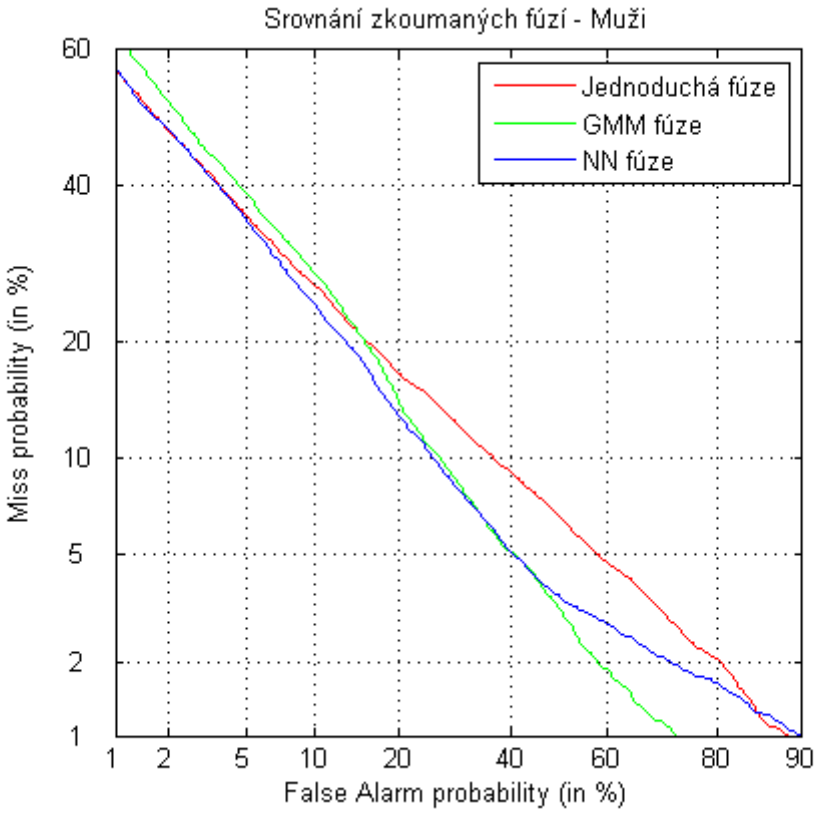
V částech 4.2.1, 4.2.2 a 4.2.3 byly prezentovány výsledky různých fúzí vybraného systému pro rozpoznávání promluvy a systému pro rozpoznávání mluvčího. Projděme si tedy, který typ fúze měl největší úspěšnost a který je naopak nejhorší. Pro porovnání výsledků jsou DET křivky všech tří fúzí zobrazeny do jednoho grafu pro ženy [Obr. 31] a pro muže [Obr. 32].

Jednoduchá fúze využívající Mean and Variance normalizaci byla použita v základním tvaru, tedy jako jednoduchá normalizace hodnot obou systémů a následné sečtení hodnot skóre. V grafech na obrázcích [Obr. 31] a [Obr. 32] je zobrazena červenou křivkou. Ze čtyř testovaných variant GMM fúze byl pro vyhodnocení použit systém pracující se dvěma Gaussovými křivkami, neboť

větší počet křivek neměl až tak zásadní vliv na konečné výsledky systému. Na obrázcích [Obr. 31] a [Obr. 32] je křivka GMM fúze označena zelenou barvou. Fúze pomocí neuronových sítí je v těchto grafech vyznačena modře. Je pro ni využita varianta s patnácti neurony ve skryté vrstvě, protože také dávala nejlepší výsledky.



[Obr. 31] Porovnání výsledků fúzí – Ženy



[Obr. 32] Porovnání výsledků fúzí - Muži

Z grafů je vidět, že nejhůře dopadla jednoduchá fúze. To je očekávaný výsledek, protože tato fúze je nejjednodušší a neumí vidět chování systémů v širších souvislostech. Není možné ji natrénovat ani pomocí ní vytvořit nějaký model, který bude reprezentovat daný systém. Obě složitější fúze dopadly lépe. Není úplně jednoduché říci, která je lepší a která horší, protože to bude záležet na zamýšleném využití. Pokud bychom se dívali na hodnoty EER, které najdeme v grafech tam, kde se hodnota Pfa rovná Pmiss (Pfa = Pmiss), bude vítězem fúze založená na neuronových sítích. Pro většinu využití je vhodné právě toto nastavení systému, takže využití fúze pomocí neuronových sítí bude nejuniverzálnější. Pokud půjdeme v grafu k nižším hodnotám Pfa, neboť bude na systém požadavek, že pravděpodobnost nesprávného přijetí musí být malá i za cenu větší pravděpodobnosti nesprávného odmítnutí (Pfa < Pmiss), dostaneme lepší výsledky opět z fúze s neuronovými sítěmi. Tyto požadavky budou při aplikaci systému tam, kde je třeba vysoké bezpečnosti. Například trezor banky, kam nebude vpuštěn nikdo, kdo skutečně nemá oprávnění vstoupit. Pokud bude náhodou odmítnut i oprávněný mluvčí, například ředitel banky, raději promluvu zopakuje, pokud si systém nebude jist. Druhá část grafu s nízkými hodnotami Pmiss, kde Pmiss < Pfa, ovšem ukazuje změnu. Čím menší bude požadovaná pravděpodobnost nesprávného odmítnutí, tím lépe bude fungovat GMM fúze oproti neuronovým sítím.

Z výsledků tedy vyplývá, že fúzi pomocí neuronových sítí lze doporučit pro většinu standardních a bezpečnostních aplikací. Pro požadavky menší pravděpodobnosti nesprávného odmítnutí, je ale lépe využít fúzi GMM, díky níž bude systém fungovat o něco lépe. Jednoduchou

fúzi s normalizací Mean and Variance příliš doporučit nelze, neboť dopadla nejhůř, pokud však potřebujeme nějakou jednoduchou metodou spojit dohromady výsledky dvou rozpoznávacích systémů, pak je dobré vyzkoušet i tuto fúzi, protože výsledky budou lepší, než když fúzi nepoužijeme vůbec.

5 Závěr

V této bakalářské práci jsem se věnoval textově závislému rozpoznávání mluvčího s využitím krátkých nahrávek. Vysvětlil jsem, proč je dobré zabývat se krátkými nahrávkami v úlohách rozpoznávání, a také jsem popsal problémy, které z toho plynou. Také jsem vysvětlil, že pro dosažení dobrých výsledků bude nutné zabývat se nejen rozpoznáváním mluvčího, ale i rozpoznáváním textu. Popsal jsem i důležité znaky každého mluvčího, které jsou použitelné pro jeho rozpoznání, a zařadil je do kategorií.

V další části jsem se zabýval myšlenkou dobrého skóre rozpoznávání, kterého je při takto krátkých nahrávkách možno dosáhnout kombinací rozpoznávání textu a hlasu, tedy fúzí dvou různě orientovaných systémů. Následovalo bádání po různých metodách a systémech, které by byly schopny s dobrými výsledky rozpoznávat mluvčího, nebo promluvu. Všechny vybrané systémy jsem následně popsal a zhodnotil jejich možnosti a výhody pro tuto úlohu. Pro rozpoznávání promluvy jsem vybral Fonémový rozpoznávač dostupný v rámci fakulty, který sliboval lokalizování fonémů v promluvě a výstup ve formě posteriogramu, nebo přepisu do fonémového řetězce. Pro zpracování fonémového řetězce jsem zvolil program HResults z HTK Toolkitu. Dalším systémem pro rozpoznávání textu, který jsem vybral, byl Key Word Spotting od společnosti Phonexia. Jednalo se o systém, který je již nasazený v terénu, takže sliboval, že bude dobře otestován. Nakonec jsem zvolil také algoritmus DTW, který je základním algoritmem pro rozpoznávání promluvy podle reference. Jeho možnost srovnání nahrávek, které jsou posunuté v čase, nebo mají různou délku, sliboval dobré skóre i při různých variabilitách v projevech mluvčího. Algoritmus DTW jsem se rozhodl vyzkoušet pro spektrogramy i posteriogramy. U spektrogramů byl předpoklad, že budou fungovat pro rozpoznávání mluvčího i pro rozpoznávání promluvy s podobným skóre, kdežto při použití posteriogramů bylo očekáváno, že bude lépe fungovat pro rozpoznávání promluvy. Jako systém rozpoznávání mluvčího jsem vybral systém Speaker ID od společnosti Phonexia, který je v terénu také používán.

Dále jsem zkoumal možnosti, jak lze provést spojení systémů dohromady. Popsal jsem jednoduchou fúzi využívající Mean and Variance normalizaci, fúzi využívající GMM modely a fúzi pomocí Neuronových sítí. Jako zdroj dat (nahrávky), na kterých byly systémy trénovány a testovány, jsem zvolil databázi MIT Mobile Device Speaker Verification Corpus. Jako základní nástroj na vyhodnocování a porovnávání výsledků jednotlivých systémů jsem využil program DETware, který umožňuje zobrazování výsledků ve formě DET křivek.

Další část byla věnována samotným experimentům s jednotlivými systémy a zvolenými metodami. Průběh každého experimentu jsem pečlivě popsal, zaznamenal výsledky a sestrojil grafy s DET křivkami. Experimentováním jsem zjistil, že z použitých metod a systémů je pro rozpoznávání promluvy nejlépe aplikovat algoritmus DTW na předem vytvořené posteriogramy. Jedna podkapitola byla věnována také experimentům s počtem fonémových stavů v posteriogramu. Úspěšnost systému (EER) byla 19,01% pro ženy a 16,50% pro muže. Pro rozpoznávání mluvčího měl největší úspěšnost (EER) systém SID, a to 35,64% pro ženy a 35,78% pro muže, což bylo očekávané, protože jako jediný z použitých systémů je přímo pro tuto úlohu vyvinut.

Následovalo samotné vytváření jednotlivých fúzí mezi těmito dvěma nejlepšími systémy. Všechny fúze byly prováděny v prostředí MATLAB a výsledky každé z fúzí byly opět popsány a

vyhodnoceny. Fúze byly také testovány pro různé nastavení základních parametrů. U GMM fúze to byl počet Gaussových křivek. Jako ideální se ukázalo využití 2 křivek. Fúze s Neuronovými sítěmi nabízela variabilitu v počtu neuronů. Z pokusů vychází, že optimální počet je okolo 15. Z výsledků jsem také zjistil, že pro tak krátké nahrávky, jaké byly použity při této práci, je skutečně na místě se fúzí rozpoznávání mluvčího a rozpoznávání promluvy zabývat, protože nám umožní využít výhod obou systémů a skóre rozpoznávání oproti samostatným systémům zvýšit. Nejlepší výsledky byly získány z fúze pomocí Neuronových sítí. Úspěšnost byla 17,84% (EER) pro ženy a 16,38% pro muže, což je relativní zlepšení 49,9% u žen a 54,2% u mužů oproti samotnému akustickému rozpoznávání. Tato fúze se velmi hodí pro všechny standardní a bezpečnostní aplikace, u kterých preferujeme bezpečnost před komfortem (systém nastavený tak, že při sebemenší nejistotě o identitě mluvčího jej raději nechá promluvu zopakovat, než by vpustil podvodníka), nebo tam, kde mají obě vlastnosti systému stejnou váhu. V určitých případech by však byla lepší fúze GMM. To platí při preferování většího komfortu před bezpečností (systém přijme i nahrávky s horším skórem). Podrobnější vyhodnocení výsledků testovaných fúzí lze nalézt v kapitole 4.3.

Systém SID pravděpodobně nefunguje zcela optimálně kvůli příliš krátkým nahrávkám. Možnost pro zlepšení by do budoucna mohla být v adaptaci akustického SID systému. Nejprve by byl systémem natrénován adaptační soubor a ten by byl následně použit při samotném rozpoznávání. Výsledkem této adaptace systému by mělo být zlepšení skóre. Další možností je zrevidování jednotlivé kroků systému a snaha o optimalizaci. Systém v sobě například obsahuje normalizaci, která se aktuálně počítá přes okno dlouhé 3 sekundy. Využité nahrávky však nejsou tak dlouhé, což může být důvod k horšímu skóre rozpoznávání. Další možností by mohlo být natrénování menšího systému. Redukováním počtu parametrů by se systém také mohl stát úspěšnější pro krátké nahrávky. Zajímavé by mohlo být i vyzkoušení jednoduchého GMM-UBM systému (*Universal Background Model*).

Obecně lze z většiny grafů vypozorovat, že všechny systémy fungovaly lépe pro muže než pro ženy. Tento poznatek byl získán z experimentů na datech z databáze *MIT Mobile Device Speaker Verification Corpus*. Zajímavým tématem by tedy mohlo být i otestování dostupných systémů s různými testovacími databázemi a vypozorování, zda je to záležitost pouze této databáze, nebo některých využitých systémů.

6 Literatura

[1] PHONEXIA S.R.O. Phonexia Keyword Spotting. [online]. 2006 [cit. 2013-05-08]. Dostupné z: http://phonexia.com/docs/white/AcousticKWS_v3_cze.pdf

[2] MARTIN, A., G. DODDINGTON, T. KAMM, M. ORDOWSKI a M. PRZYBOCKI. The DET curve in assessment of detection task performance. [online]. [cit. 2013-05-08]. Dostupné z: http://www.itl.nist.gov/iad/mig/publications/storage_paper/det.pdf

[3] MATĚJKA, P., P. SCHWARZ, J. ČERNOCKÝ a P. CHYTIL. Phonotactic Language Identification using High Quality Phoneme Recognition. [online]. 2005 [cit. 2013-05-08]. Dostupné z: http://www.fit.vutbr.cz/~matejkap/publi/2005/eurospeech2005.pdf

[4] GOLD, B. a N. MORGAN. *Speech and audio signal processing: Processing and perception of speech and music*. New York: John Wiley and Sons, 2000. ISBN 0-471-35154-7.

[5] PHONEXIA S.R.O. Phonexia Speaker Identification. [online]. 2006 [cit. 2013-05-08]. Dostupné z: http://www.phonexia.com/download/demo-sid

[6] MCCULLOCH, W. S. a W. PITTS. A logical calculus of the ideas immanent in nervous aktivity. *Bulletin of Mathematical Biophysics*, 5:115-133, 1943.

[7] ŠÍMA, J. a R. NERUDA. *Teoretické otázky neuronových sítí*. Vyd. 1. Praha: MATFYZ press, 1996. ISBN 80-85863-18-9.

[8] MARTIN, A. F. Speaker Databases and Evaluation. *Encyclopedia of Biometrics* [online]. 2010 [cit. 2013-05-08]. Dostupné z: http://www.itl.nist.gov/iad/mig//publications/storage_paper/Biometrics_Encyclopedia_Entry_Voice.pdf

[9] UNIVERSITY OF CAMBRIDGE. HTK Speech Recognition Toolkit. [online]. 2003 [cit. 2013-05-08]. Dostupné z: http://htk.eng.cam.ac.uk/

[10] YOUNG, S., D. KERSHAW, J. ODELL, D. OLLASON, V. VALTCHEV a P. WOODLAND. The HTK Book: for HTK Version 3.1. [online]. 1995 [cit. 2013-05-08]. Dostupné z: http://htk.eng.cam.ac.uk/docs/docs.shtml

[11] MEHROTRA, K., K. MOHAN a S. RANKA. *Artificial neural networks*. Vyd. 1. Cambridge: MIT Press, 1997. ISBN 0-262-13328-8.

[12] HUANG, X., A. ACERO a H. HON. *Spoken language processing: a guide to theory, algorithm, and system development*. Vyd. 1. New Jersey: Prentice-Hall, 2001. ISBN 0-13-022616-5.

[13] REYNOLDS, D. Gaussian Mixture Models. [online]. [cit. 2013-05-08]. Dostupné z: http://www.ll.mit.edu/mission/communications/ist/publications/0802_Reynolds_Biometrics-GMM.pdf

[14] ČERNOCKÝ, J. Definice směsí Gaussových rozdělení. In: UPGM FIT VUT BRNO. *Rozpoznávání řeči - HMM: Přednáškové materiály*. Brno, 2010, s. 38.

[15] SILOVSKÝ, J. *Generativní a diskriminativní klasifikátory v úlohách textově nezávislého rozpoznávání a diarizace mluvčích*. Liberec, 2011. DISERTACNÍ PRÁCE. Technická univerzita v Liberci. Vedoucí práce Prof. Ing. Jan Nouza, CSc.

[16] MATĚJKA, P. SPEECH@FIT, Brno University of Technology, Czech Republic. *Speaker recognition*. Brno, 2009. Přednáška BEST summer school.

[17] PHONEXIA S.R.O. Phonexia Speaker Identification White paper: Speaker identification technology V2. [online]. 2011 [cit. 2013-05-08]. Dostupné z: www.phonexia.com

[18] ELLIS, D. Dynamic Time Warp (DTW) in Matlab. [online]. 2003 [cit. 2013-05-11]. Dostupné z: http://www.ee.columbia.edu/~dpwe/resources/matlab/dtw/

7 Seznam použitých zkratek

SID – Speaker Identification systém

MLF – Master Label File

REC – Record

HTK - Hidden markov models Toolkit

KWS – Key Word Spotting

GUI – Graphical User Interface

ML – Maximum Likelihood

DTW – Dynamic Time Warping

MIT – Massachusetts Institute of Technology

NIST – National Institute of Standards and Technology

DET – Detection Error Tradeoff

EER – Equal Error Rate

DCF – Decision Cost Function

Pfa – Probability of False Alarm detection

Pmiss – Probability of Miss detection

GMM – Gaussian Mixture Model

NN – Neural Network

8 Seznam příloh

Příloha 1. CD s textem práce