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

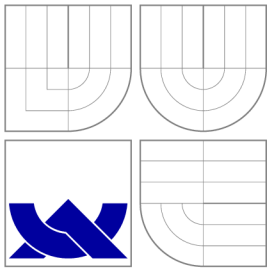
SROVNÁNÍ ÚSPĚŠNOSTI SIRI, CORTANY A GOOGLE

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

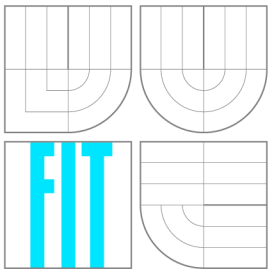
AUTOR PRÁCE
AUTHOR

LUCIE PROCINGEROVÁ

BRNO 2015



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

SROVNÁNÍ ÚSPĚŠNOSTI SIRI, CORTANY A GOOGLE

COMPARISON OF ACCURACY OF SIRI, CORTANA AND GOOGLE

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE
AUTHOR

LUCIE PROCINGEROVÁ

VEDOUCÍ PRÁCE
SUPERVISOR

Ing. IGOR SZŐKE, Ph.D.

BRNO 2015

Abstrakt

Cílem této práce je porovnat úspěšnost překladu mluveného slova do textu s využitím několika služeb. Primárně se jedná o aplikace od společností Apple Inc., Microsoft Corporation a Google Inc., avšak je zde zahrnuto také několik dalších aplikací, dostupných převážně online. Tento dokument obsahuje popis zadaného problému, rozbor postupu provádění přepisu u jednotlivých služeb. Následně jsou rozberány výsledky testu a porovnány s referenčními výstupy. Na závěr je uvedena diskuze těchto pokusů.

Abstract

The aim of this thesis is to compare the accuracy of translation of spoken word into text using several services. Primary it is about applications from Apple Inc., Microsoft Corporation and Google Inc., but there is also included several others, mostly available on-line. This document contains a descriptionn of the problem, analyzes the progress for each service. Subsequently, the test results are analyzed and compared with the reference outputs. In conclusion, there is a discussion of these experiments.

Klíčová slova

Siri, Cortana, Google, audio, řeč, srovnání, přepis, strojové rozpoznávání řeči

Keywords

Siri, Cortana, Google, audio, speech, comparsion, transcription, automatic speech recognition

Citace

Lucie Procingerová: Srovnání úspěšnosti Siri, Cortany
a Google, bakalářská práce, Brno, FIT VUT v Brně, 2015

Srovnání úspěšnosti Siri, Cortany a Google

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně pod vedením pana Ing. Igora Szőkeho, Ph.D. V práci jsem uvedla veškeré literární prameny a publikace, ze kterých jsem čerpala.

.....
Lucie Procingerová
20. května 2015

Poděkování

Ráda bych touto cestou poděkovala panu Ing. Igoru Szőkemu, Ph.D. za jeho pomoc a cenné rady při tvorbě této práce. Velké poděkování patří také moji rodině a přátelům za podporu v průběhu celého studia, zvláště během vypjatých situacích ve zkouškovém období a při zpracování bakalářské práce.

Obsah

Kapitola 1

Úvod

Řeč je v běžném pojetí výsadou člověka. Ostatní živočichové dokáží pouze porozumět slovům, ale sami mluvit neumí. Naučit se ovládat řeč je tedy jeden z nejdůležitějších momentů v životě člověka. Není to tak dávno, kdy se člověk naučil komunikovat také s počítačem. A to nejen prostřednictvím čísel a instrukcí, ale i pomocí mluveného slova. V současnosti je tedy běžné mluvit s počítačem, tabletem či mobilním telefonem a ušetřit spoustu času psaním, ať už ve škole, v práci či v běžném životě.

Tato práce si klade za cíl srovnat úspěšnost přepisu řeči do textové podoby zpracováním různými službami. Zejména se zaměřím na aplikace od známých společností jako je Apple Inc., Microsoft Corporation a Google Inc. Od těchto světoznámých firem jsou využity inteligentní hlasové asistentky Siri a Cortana. Google Inc. nabízí například Google Now, jenž je dostupný pro zařízení s operačním systémem Android. Nemalou část experimentování zaplní také další služby, dostupné on-line. V celé této práci se budu věnovat pouze přepisu nahrávek namluvených anglickými rodilými mluvčími. Čeština prozatím není příliš rozšířená a málokterá služba disponuje možností využít automatické přepisy česky namluvených nahrávek. Naopak, na jazyky ve světě hojně používané jako je čínština, španělština nebo ruština, můžeme narazit velmi často.

Mým záměrem je nastínit historii a vývoj komunikace člověka s výpočetní technikou, popis a rozbor tří hlavních aplikací a také popis některých dalších volně dostupných on-line služeb, které byly pro tuto práci zajímavé a využitelné. Samostatnou kapitolu představují nejen ukázky experimentů s jednotlivými službami, ale je tady předveden také způsob, jakým byly experimenty prováděny. Velkou měrou se věnuji přehlednému vyhodnocení vytvořených přepisů a porovnávání s referenčními výstupy. Představuji zde i klady a nedostatky jednotlivých, výše uvedených aplikací. Na závěr se věnuji vyhodnocení všech výsledků a diskuzi nad provedenými experimenty.

Kapitola 2

Rozpoznávání řeči

Tato kapitola popisuje historii vývoje rozpoznávání řeči a zaměřuje se i na obecný popis způsobu fungování služeb, které přepis poskytují. V dalších kapitolách jsou podrobněji rozebrány jednotlivé aplikace a v závěru se nachází tabulka s přehledným shrnutím.

2.1 Historie

Už statitísce let řeč umožňuje lidem, aby se navzájem domluvili a zároveň je jazyková různost jednou z největších překážek lidské komunikace. Lidé se však nepotřebují dorozumívat pouze mezi sebou, jejich pomocníci jsou i zvířata a v poslední době také stroje. Člověka již dlouhou dobu zajímá návrh a technologie systému pro rozpoznávání lidské řeči. Průkopníkem v analýze lidské řeči, a především v praktických konstrukčních provedeních tzv. „mluvících strojů“, byl Johann Wolfgang von Kempelen v druhé polovině 18. století a následně na něj nezávisle navázal Christian Gottlieb Kratzenstein [8]. Lidstvo dále zkoušelo nové metody a experimenty na zkonstruování systému pro analýzu, syntézu a rozpoznávání řečového signálu. Největší rozmach však nastal až s nástupem číslicových počítačů, kde mohly být uplatněny metody digitalizace a číslicové zpracování řečového signálu. Nicméně většina nových metod byla závislá na aktuálním pokroku dosaženém při rozvoji výpočetních systémů. Typický příklad je Fourierova analýza[1], která je dnes velmi významným prostředkem při zpracování signálů. V analogové verzi je známa již z první poloviny 18. století, ale diskrétní alternativa se stala aktuální teprve s příchodem číslicových počítačů v padesátých a šedesátých letech. Největší uplatnění nalezla v první polovině devadesátých let, kdy se poprvé objevily signálové procesory. Dnešní výkon počítačů je však dostačující na zpracování i vyhodnocení řečového signálu.

Podobný rozvoj lze sledovat v odvětví klasifikace řeči. V sedmdesátých a na počátku osmdesátých let byla nejvíce využívána klasifikace po jednotlivých izolovaných slovech a slovních spojeních. Nové způsoby klasifikace přinesl projekt DARPA-SUR[2], který ukázal nové technologie pro rozpoznávání souvislé řeči. V průběhu osmdesátých let se rozvinula metoda založená na statistickém přístupu ke zpracování řečového signálu, která je vhodná k rozpoznávání souvislé řeči. V současnosti se využívá způsob modelování řeči založený na tzv. Markovových modelech[3], které modelují kratší subslovní jednotky [11].

[1] Jean Baptiste Joseph Fourier byl francouzský matematik a fyzik, který se nejvíce proslavil zkoumáním Fourierových řad a jejich aplikací k problémům toků tepla.

[2] angl. Defence Advanced Project Agency, zkr. DARPA, angl. Speech Understanding Research, zkr. SUR

[3] Andrej Andrejevič Markov byl ruský matematik, jehož objevy byly později nazvány Markovovými řetězci.

Komunikace člověka s počítačem lze rozdělit na tři dílčí úlohy – syntéza řeči, rozpoznávání řeči a porozumění řeči. Rozdíl mezi pojmy je zobrazen na obrázku 2.1.

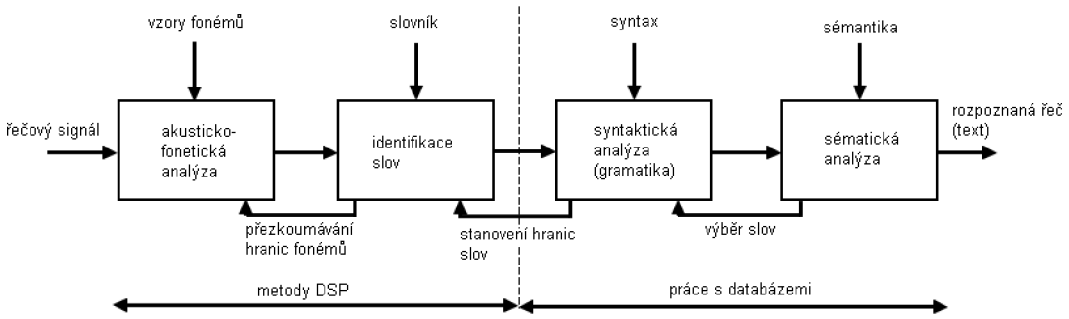
Obrázek 2.1: Rozdíl mezi syntézou řeči a rozpoznáváním řeči.

2.2 Současnost

V současnosti zažívají tyto služby velký rozmach. Je to zejména z toho důvodu, že výpočetní síla dnešních počítačů je dostatečná a veškeré přepisy zvládne běžný počítač či notebook. Na druhou stranu zaznamenaly obrovský pokrok také mobilní zařízení a hlasové ovládání telefonu či tabletu je naprosto běžné. Tuto funkci využívá velká část populace pro usnadnění komunikace se zařízením, např. hlasové vytáčení hovoru při řízení v autě. Také se hojně využívají přepisy nahrávek přednášek, jednání nebo konferencí místo psaní objemných textů ručně, jelikož se tímto způsobem zápisu dá ušetřit spousta času. Další možné využití lze najít v oblasti usnadnění přístupu pro tělesně postižené nebo zrakově postižené osoby. Příkladem může být automatické titulkování zpráv nebo hlasové povely pro ovládání telefonu. Další oblasti využití řečových technologií [14]:

- zdravonictví,
- ozbrojené síly,
- battle management.

Ovšem naučit stroje porozumnění lidské řeči, aby správně reagovaly na to, co jim na vstupu sdělíme ve svém vlastním jazyce, bez zvláštního kódování, které si jen namáhavě osvojujeme, není jednoduchý úkol. Lidský jazyk je totiž třeba chápat jako soustavu, která zakotvuje vzájemný vztah mezi obsahy vědomí a zvuky. Jazyk není jen soustava hlásek, tvarů a slov, ale také jednotek významových. K rozvoji experimentů s počítači s tzv. automatickým rozpoznáváním řeči a umělým intelektem je zapotřebí mnohem více odborníků. Zpracování řeči je totiž obor se širokým záběrem, kde se využívají poznatky věd přírodních, technických i humanitních. Často se tak můžeme setkat s tím, že se na vývoji nových technologií podílí specialisté z oblasti lingvistiky, psychologie, logiky, filozofie a dalších nově vzniklých oblastí vědeckého výzkumu [9]. Uplatnění jednotlivých lingvistických nauk v procesu rozpoznávání plynulé řeči zobrazuje blokové schéma na obrázku. 2.2.



Obrázek 2.2: Princip automatického rozpoznávání řeči. Zdroj: [9]

2.3 Metody rozpoznávání řeči

Z hlediska aplikovaných metod rozpoznávání lze klasifikátory řeči rozdělit na:

- fungující na principu **porovnání se vzory**,
- pracující s využitím **statistických metod**.

Skupina metod založená na principu porovnávání se vzory byla hojně využívána zejména v sedmdesátých a osmdesátých letech. Její použití bylo často spojeno s klasifikátory izolovaně vytvořených slov, tzn. že slovo je zde zpracováno jako celek, přičemž je zařazeno do třídy, k jehož vzorovému obrazu má nejmenší vzdálenost. U tohoto způsobu se ale objevil problém s tím, že každé slovo (dokonce několikrát namluvené stejným řečníkem) je různé dlouhé. Musela se tedy zjišťovat hodnota vzdálenosti mezi dvěma obrazy slov. První možností, jak se s tímto problémem vyrovnat, je lineární časová normalizace, tj. zkrácení nebo natažení promluv na stejnou konstantní délku. Když ale porovnáme dvě nahrávky stejného slova, zjistíme, že kromě odlišné délky se liší i poměrem délek vyslovování jednotlivých hlásek. Časová normalizace na jednotnou délku nemůže toto nelineární časové kolísání uvnitř slova postihnout. Tuto otázku následně vyřešila až aplikace metody dynamického programování[4], při kterém se hledá taková nelineární transformace časové osy jednoho obrazu, u níž dojde k porovnání obou obrazů s nejmenší výslednou vzdáleností.

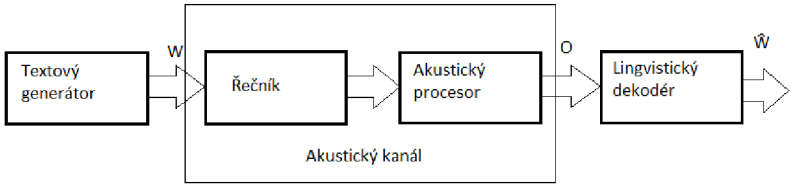
U druhé skupiny metod je klasifikace založena na statistických metodách, ve kterých jsou slova a celé promluvy modelovány pomocí skrytých Markovových modelů. Nejčastěji jsou konstruovány skryté Markovovy modely subslovních jednotek a promluva je pak modelována zřetězením těchto subslovních modelů [11]. Skryté Markovovy modely jsou více rozebrány v sekci 2.3.2.

2.3.1 Statistický přísup k rozpoznávání souvislé řeči

Jelikož technika rozpoznávání izolovaných slov založená na dynamickém programování i přes svoji výbornou účinnost ztratila svoji aktuálnost, jsou v současné době využívány statistické metody rozpoznávání řeči.

[4]Dynamické programování je odvětví optimalizace, kde je stěžejní myšlenkou rozklad problému na podproblémy. Tyto jsou dále řešeny a jejich řešení je ukládáno pro další potenciálně možné použití.

Základní schéma statistického přístupu k rozpoznávání mluveného slova je složeno z akustického procesoru a lingvistického dekodéru. Řečník a akustický procesor jsou spojeni pomocí akustického kanálu. Akustický procesor transformuje řečový signál na posloupnost značek nebo-li vektorů příznaků (většinou jeden vektor na každých 10 ms) a lingvistický dekodér převádí tuto posloupnost na řetězec slov. Rozpoznávání chápeme jako dekódování s maximální pravděpodobností. Toto schéma je naznačeno na obrázku 2.3.



Obrázek 2.3: Blokové schéma systému rozpoznávání řeči založené na statistickém přístupu. Zdroj: [11]

2.3.2 Skryté Markovovy modely

Skryté Markovovy modely (anglicky Hidden Markov Models, zkratka HMM) se začaly v širší míře využívat v 70. letech minulého století a v současné době stojí za rozvojem systémů na rozpoznávání řečových signálů. Hlavním cílem je stanovení podobnosti či rozdílnosti testovaného slova se vzory uloženými v paměti klasifikátoru. Proto se tyto metody založené na skrytých Markovových modelech řadí do kategorie rozpoznávání podle vzorů.

Metody založené na porovnávání obrazců se ve velké míře využívají z důvodu jejich snadné implementace, robustnosti a invariace z hlediska různých možností realizace slova, typů mluvčích, hluku pozadí a dalších náhodných rušivých jevů. I přes všechno vyjmenované se však řadí mezi metody s vysokou úspěšností rozpoznávání. Modelována mohou být nejen celá slova, ale i menší jednotky jako slabiky či fonémy apod. Při rozpoznávání slov hledáme model slova z tzv. kódové knihy, který by s největší pravděpodobností vygeneroval testované slovo.

Skryté Markovovy modely nemusí být využity pouze v oblasti zpracování řečových signálů, ale mohou být využity i pro jiné klasifikační aplikace, např. biometrické testy sítnic, otisky prstů nebo identifikace DNA.

Vzhledem k tomu, že Markovovův model je statistickým typem modelu, je nutná pro pochopení jeho matematického popisu orientace v základních pojmech z oborů matematické pravděpodobnosti a statistiky [9].

2.4 Rozpoznávače

Umístění rozpoznávačů v komunikačním prostředí může být lokální, tj. v našem počítači, v automobilu, v telefonu, avšak velmi časté je jejich umístění na konci telekomunikačního spoje (v telefonní ústředně, na serveru v bance atp.). Kromě rozpoznávání obsahu promluvy, může být požadováno rozpoznání (identifikace nebo verifikace) mluvčího, určení jazyka, kterým mluvčí hovoří, určení pohlaví mluvčího a další. Tomuto oboru se říká dolování informací ze spontánních řečových dat. Nejaktuálnějším, a pro tuto práci také hlavním předmětem

zkoumání, je však konstrukce systému pro rozpoznávání obsahu promluvy. S vývojem algoritmů a s možnostmi výpočetní techniky vznikly různé specifikace rozpoznávačů.

Dělení rozpoznávačů podle [10]:

- zavislé a nezávislé na mluvčím,
- telefonní nebo lokální,
- s malým nebo velkým slovníkem,
- schopnost rozpoznat izolovaná slova, izolovaná sousloví nebo souvislou řeč.

U všech výše uvedených typů velmi významně ovlivňuje náročnost konstrukce hledisko založené na počtu mluvčích. Každý mluvčí má svou charakteristickou výslovnost a rozpoznávač s ní tak může počítat. V případě rozpoznávače, který není závislý na mluvčím, musí být nalezeny rysy společné pro velký počet mluvčích. K takovému nalezení je ale potřeba obrovská spousta hlasových záznamů, ze kterých se pak společné charakteristiky vyjmou a specifika jednotlivých mluvčích se co nejvíce eliminují.

Díky tomu, že zvuky souvislé řeči obsahují neustále se měnící spektrum harmonických frekvencí a také hluk, mění se i hlasitost a rychlost řeči. Jeden a tentýž výraz vyslovený různými lidmi, anebo dokonce i stejným člověkem, který se nachází v různých psychických stavech, může mít odlišné spektrální a časové barvy. Tím se vyrázně komplikuje implementace universálního systému pro rozpoznávání řeči.

Komunikace se strojem může být rozdělena na dvě varianty. První možností je práce stroje s izolovanými slovy, pro která má v paměti uložené vzory. V případě druhé možnosti, kdy stroj bude zpracovávat souvislou spontánní řeč, musí rozpoznávač vzory hledat tak, že si nejprve musí uspořádat slova do delší sekvence a pak teprve identifikuje odpovídající promluvu. Algoritmus pro rozpoznávání souvislých čtených textů bude tedy o něco jednodušší.

V každém rozpoznávači můžeme nalézt tzv. slovník slov. V systémech, jenž rozpoznávají pouze izolovaná slova, dostačuje malý slovník (několik desítek slov). Vzory pak lze mít uloženy jako celá slova zaznamenaná ve fázi trénování. Avšak v případě rozsáhlejších slovníků již nelze systém jednoduše trénovat na všechna možná slova. Řešením je slovník vytvořený z foneticky přepsaných textů, které očekávaná slova obsahují a trénování musí probíhat na souborech promluv, které obsahují všechny fonetické elementy použité při přepisu reprezentativních textů do fonetické podoby [10].

2.5 Problémy při rozpoznávání řeči

Rozpoznávání řeči sice zaznamenalo obrovský pokrok, avšak stále se tyto systémy potýkají s problémy, které je ještě třeba řešit a zdokonalovat tak služby poskytující takové přepisy. V současnosti již funguje mnoho velmi kvalitních komerčních programů, ale i programů volně dostupných, ať už pro počítače či jakákoliv mobilní zařízení. Nyní zde nastíním několik možných problémů.

- Velká variabilita řečového signálu – zde se jedná o to, že každý řečník vysloví tutéž větu pokaždé jinak, dokonce stejný řečník často vystovuje jednu větu zcela odlišným způsobem.

- V řečovém signálu se výrazně projeví jakákoli změna prostředí (např. akustika místnosti, rušivé zvuky) nebo přenosového kanálu (např. jiný typ mikrofonu, řeč přenášená přes telefon).
- Při rozpoznávání plynulé řeči nelze jednoduše automaticky stanovit začátek a konec jednotlivých slov.
- Existují také překážky, jako např. porozumění sémantice řeči.
- Mohou se též objevit různé nonverbální projevy (kašlání, dýchání, pochybování), zvuky vnějšího okolí (klepání), neznámá slova či použití slangu apod.

Kapitola 3

Rešerše a popis aplikací pro rozpoznávání řeči

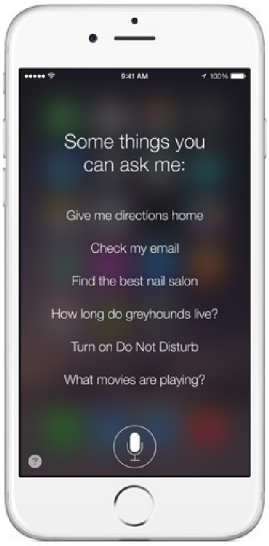
V této kapitole jsou jednotlivě popsány aplikace poskytující automatický přepis audio nahrávek do textu.

3.1 Siri

Siri je inteligentní osobní asistent(ka), jenž je součástí Apple iOS od verze 5. Aplikace využívá přirozený mluvený jazyk. Vznik Siri se datuje rokem 2007, kdy byla vymyšlena a posléze naprogramována třemi vývojáři ze společnosti DARPA. Oficiálně však byla představena až v první polovině roku 2010, kdy se ještě společnost Apple rozhodovala, zda tuto aplikaci poskytne také dalším operačním systémům. Nakonec Apple toto zamítl a nechal si Siri pouze pro sebe. Siri nelze využívat ani aplikacemi třetích stran. Do chytrého telefonu se dostala až na podzim roku 2011, kdy byla představena jako exkluzivní součást telefonu iPhone 4S. Je tedy integrována do telefonů iPhone 4s a novějších, tabletu iPad 3. generace a také do zařízení představených o rok později, tedy v roce 2012. Jedná se i o menší tablet iPad mini a přehrávač iPod touch 5. generace. Apple se mimo jiné rozhodl integrovat Siri také do chytrých hodinek Apple Watch, avšak zde jsou její funkce omezené a je zapotřebí vlastnit také telefon této značky.

Podporováno je několik světových jazyků (začátkem roku 2015 čítá kolem 25 jazyků), z nichž pro tuto práci je důležitá zejména angličtina. Při práci s asistentem je vyžadováno připojení k internetu. Převážná většina funkcí a rozšíření je navíc dostupná výhradně pro USA a Kanadu.

Modul rozpoznávání řeči je řešený za pomoci společnosti Nuance Communications. Siri začne fungovat ihned po zapnutí zařízení. Čím více se využívá, tím jsou odpovědi přesnější a relevantnější. Naučí se rozpoznat také tón řeči a dialekt. Siri přichází do kontaktu se širšími variacemi jazyku a tím se stává stále přesnější. Aplikace také využívá seznam kontaktů, kalendář nebo aktuální polohu [1].



Obrázek 3.1: Ukázka hlasového asistenta Siri[5]

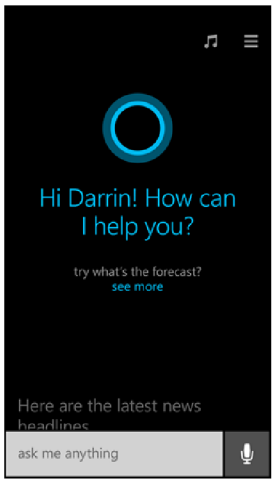
3.2 Cortana

Společnost Microsoft se svým operačním systémem pro chytré telefony rozhodně nezůstává pozadu. Již v době, kdy byla představena Siri od konkurenčního Apple, telefony s operačním systémem Windows Phone 7.5 obsahovaly jednoduchou digitální asistentku pro usnadnění používání některých funkcí např. při jízdě autem. S novějším operačném systémem Windows Phone 8.1 však přišla i nová asistentka nesoucí jméno Cortana. Toto jméno dostala podle známé hry Hallo 4. Je postavena na technologii Tellme, kterou Microsoft koupil již v roce 2007. Data jsou čerpána z vyhledávače Bing nebo z Wikipedie, používá stejně jako Siri informace uložené v telefonu. V učení o uživateli je Siri velmi podobná, avšak navíc obsahuje tzv. Cortanin notebook, kde se dají další informace o osobě, která aplikaci využívá, přidat manuálně. Jelikož Microsoft vyvíjel a také uvedl tuto asistentku o něco později než konkureční firmy Apple či Google, poučil se z některých chyb a také některé funkce vylepšil.

Narozdíl od společnosti Apple, která si trvá na uchování informací o zdrojových kódech apod., Cortana API je poskytována vývojářům aplikací třetích stran. Zatím jsou integrovány pouze dva jazyky - angličtina a čínština. Pro účel této studie je však hlavní jazyk anglický. Příkazy je možné zadávat dvěma způsoby, a to hlasem nebo vepsáním textového požadavku přímo do řádku aplikace. Nyní je Cortana dostupná také u Xbox One a po finální integraci do chytrých telefonů se chystá expandovat také na další platformy, např. do Windows 10.

Pro fungování Cortany musí být zřízen účet na webu společnosti Microsoft. Cortana pak odesílá osobní data přímo do Microsoftu a staví na tom pak své znalosti o uživateli [4].

[5]Zdroj: `https://www.apple.com/ca/ios/siri/`



Obrázek 3.2: Ukázka hlasového asistenta Cortana[6]

3.3 Google

Společnost Google nabízí více služeb poskytujících automatické rozpoznávání řeči, ovládání telefonu hlasovými příkazy nebo dokonce vyhledávání v prohlížeči pomocí diktování slov.

Budu se věnovat aplikaci s názvem Google Now (v české lokalizaci jako Chytré karty Google).

Google Now je dostupný na různých platformách. Využití nachází zejména na mobilních zařízeních se systémem Android (verze 4.1 a vyšší), také pro telefony od společnosti Apple a v neposlední řadě i jako doplněk do prohlížeče Google Chrome na počítačích. Začátkem roku 2015 Google oznámil, že zpřístupnil využití hlasového asistenta pro 40 aplikací třetích stran. Tento systém podporuje mnoho jazyků včetně češtiny.

Stejně jako konkurenční služby, Google Now je schopen si pamatovat některé informace o uživateli, čímž se stává přesnější. Pro vyhledávání informací používá tzv. Google knowledge graph. Knowledge graph je vlastní vylepšení Google vyhledávače o funkci vyhledávání na základě znalostní báze nebo informací získaných pomocí sémantického vyhledávání z jiných webů. Cílem je, aby uživatel získal odpověď na svoji otázku přímo po vložení dotazu a nemusel tak navštívit jinou stránku [3].

[6]Zdroj: `http://en.wikipedia.org/wiki/Microsoft_Cortana`

Obrázek 3.3: Ukázka aplikace Google Now[7]

3.4 Dictation

Přehledný nástroj s názvem Dictation je taktéž dostupný z webového rozhraní. Jeho funkcí můžeme využít na adrese **www.dictation.io**. Aplikace je dostupná i jako doplněk do prohlížeče Google Chrome pod názvem Voice Recognition. Nabízí převod řeči do textu v několika jazycích včetně angličtiny. Obsahuje také funkci vyexportování a uložení na lokální disk nebo na cloudové úložiště či odeslání na e-mail.

3.5 TalkTyper

Další vhodná služba pro testování je TalkTyper, dostupná na **www.talkytyper.com**. Jedná se o intuitivní a jednoduchou aplikaci přes webové rozhraní. Jeho použití není zpoplatněno. Kromě klasického diktování nabízí také možnost přeložit text do jiného jazyka. TalkTyper podporuje i češtinu a asi dvacet dalších jazyků. Disponuje několika dalšími funkcemi, jako zkopírovat text a vložit do e-mailu, sdílet text na sociální síť Twitter apod.

[7]Zdroj: `http://www.redmondpie.com/how-to-install-google-now-on-rooted-android-ics-devices-video/`

Obrázek 3.4: Ukázka služby TalkTyper

3.6 VoiceBase

Služba umístěná na adrese **www.voicebase.com** nabízí po přihlášení 200 hodin přepisu audio nahrávek a 20 hodin video nahrávek zdarma, další minuty jsou již zpoplatněny částkou $0,4. Kromě strojového přepisu je možné si objednat i lidský přepis. Práce s tímto nástrojem je velmi jednoduchá, stačí pouze nahrát audio záznam a nechat ho přepsat. Po zhotovení takového přepisu nabízí jeho editaci, sdílení a mnoho dalších rozšíření. Nástroj je dostupný také pro mobilní platformy, konkrétně pro zařízení běžící pod systémem Android a iOS.

3.7 Speechmatics

Aplikace dostupná na adrese **www.speechmatics.com** vyžaduje také registraci. Následně je nabídnuto 60 minut přepisu zdarma, další minuty jsou již placené. Služba je dostupná pouze pro britskou a americkou angličtinu, další jazyky zatím k dispozici nejsou. Po zhotovení transkripce je možnost nahlédnutí a následně také exportování do několik formátů.

3.8 SpokenData

Poslední nástroj, kterým se budu zabývat, nese název SpokenData (nebo také v české verzi Přepisovatel) a můžeme jej nalézt na adrese **www.spokendata.com**. Data se dají přidat pomocí klasického nahrávání audio nebo video záznamu, lze zadat i URL adresa nahrávky. Velká část služeb je zdarma, ale za doplňkové služby se připlácí. Po dokončení přepisu je odeslán informační e-mail a poté lze transkripci editovat a stáhnout v několik dostupných formátech. Dostupné jazyky jsou čeština, angličtina, ruština, čínština, španělština a slovenština.

3.9 Shrnutí nástrojů

V předchozím textu je představeno několik dostupných nástrojů, v tabulce 3.1 se nachází jejich přehledné shrnutí.

Existuje samozřejmě spousta dalších, ale občas bohužel ne moc dobře fungujících. Jsou dostupné také různé aplikace pro mobilní telefony či tablety. Z desktopových aplikací bych zmínila dvě nejlépe hodnocené služby. Jedná se o software od společnosti Newton Technologies, a.s. s názvem New Dictate 4 a dva programy od společnosti Fugasot, spol. s r.o. s názvem MyVoice a MyDictate. Další možností, kde najít nástroje pro převod řeči do textu, je např. v aplikacích jako Evernote, kde lze zaznamenávat poznámky pomocí hlasu podobně jako je tomu u aplikace Google Keep.

Služba	**Dostupnost**	**Cena**
Siri	Mobilní zařízení s iOS (iPhone 4S a vyšší)	Zdarma
Cortana	Mobilní zařízení s Windows Phone 7 a vyšší, OS Windows 10	Zdarma
Google Now	Mobilní zařízeni s Android 4.1 a vyšší, iOS, doplněk Chrome	Zdarma
Dictation	Webové rozhraní, doplněk Chrome	Zdarma
TalkTyper	Webové rozhraní	Zdarma
VoiceBase	Webové rozhraní, mobilní zařízení s Android a iOS	200 h zdarma
Speechmatics	Webové rozhraní	60 min zdarma
SpokenData	Webové rozhraní	Zdarma

Tabulka 3.1: Přehled

Kapitola 4

Metody testování

V této kapitole je podrobně rozepsáno, jak bylo testování prováděno.

4.1 Databáze audio nahrávek

Veškeré audio nahrávky byly poskytnuty vedoucím bakalářské práce, panem Ing. Igorem Szőkem, Ph.D. Jelikož byl k testování vybrán anglický jazyk, všechny dodané nahrávky jsou namluveny rodilými mluvčími ze Spojených států amerických, tudíž se jedná o americkou angličtinu.

Databázi bylo potřeba ručně projít a eliminovat špatně vytvořené nahrávky. Vyřazení se týkalo zejména nesrozumitelných, velmi tichých, často uchem těžce postřehnutelných, rozhovorů. Problémové je také překřikování více lidí, zvláště u jednoho společného mikrofonu, protože rozpoznávače nejsou schopny zaznamenat promluvy všech osob najednou.

Naopak pro důkladné otestování se záměrně nevyřazovaly různě zašuměné nahrávky. Také rozhovory, kde se objevuje bouchnutí dveří, pokašlávání, hlasité dýchání do mikrofonu, smích a podobné projevy, byly ponechány. V promluvách figurují ženy i muži. Objevují se vysoké i jemné a tišší ženské hlasy, také hrubý mužský hlas. Někteří lidé mluví rychle, jiní mají pomalejší promluvu a hojně využívají pauzy pro nadechnutí a přemýšlení. Jsou obsaženy také různé věkové kategorie – od dívky, které je 18 let až po staršího pána zřejmě v důchodovém věku[8]. Témata rozhovorů mají velmi široký záběr a týkají se např. sportu, cestování, domácích mazlíčků, knih, klonování, práce atd.

Databáze obsahuje čtyři typy promluv:

- telefonní hovor, záznam pouze jednoho kanálu,
- telefonní hovor, záznam obou kanálů,
- konference, každý má svůj mikrofon,
- konference, jeden společný mikrofon.

U telefonních hovorů jsou zahrnuty rozhovory žen i mužů. V případě záznamu obou kanálů se objevují kominace hovorů žena – žena, muž – muž a žena – muž. Na konferencích se většinou objevují kombinace žen i mužů.

[8] Věk osob byl odhadnut na základě obsahu promluvy nebo na základě zmínky o věku v rozhovoru.

4.2 Příprava záznamů

Celá databáze obsahuje nahrávky v různých formátech a různě dlouhé rozhovory. Konference měly dokonce více než hodinu trvání a bylo potřeba je zkrátit.

4.2.1 Formát WAV

Část dodaných dat je ve formátu WAV (nebo také WAVE, angl. Waveform Audio Format). Tento formát patří do rodiny RIFF formátů vytvořených pro výměnu dat mezi programy a je nekomprimovaný (nezmenšený). Nabízí vysokou kvalitu, avšak na úkor kapacitní náročnosti. Formát má výhodu i v tom, že do něj lze ukládat zvuk v jakékoliv kvalitě (vzorkovací frekvence, bitová hloubka) a také vícekanálový zvuk [5].

Protože je tento formát standardem ve zpracování řeči a všechny služby ho bez problému akceptují, nebylo třeba data nijak konvertovat.

4.2.2 Formát SPH

Větší část nahrávek však byla ve formátu SPH (SPHERE, z anglického Speech Header Resources). Tento formát se často využívá ve spojení se zpracováním řeči a je definován Národním institutem standardů a technologií (NIST, angl. National Institute of Standards and Technology). Zde už bylo nutné konvertování do WAV formátu, jelikož vkládání SPH souborů podporuje pouze služba Speechmatics.com.

Ke konvertování je k dispozici několik možností. Přímo NIST poskytuje pro převod software zvaný **The NIST SPeech HEader REsources (SPHERE) Package Version 2.6** dostupný pro UNIXové systémy[9].

Další možností je využít program **SoX**[10], který nabízí převod, čtení a spouštění různých formátů a navíc je multiplatformní. Spouští se přes příkazový řádek a jeho použití je jednoduché. Po instalaci je konvertování možné spustit takto:

```
sox file.sph file.wav
```

Jednoduchý program s názvem **sph_convert**[11] poskytuje také The Linguistic Data Consortium (zkr. LDC) sídlící na Univerzitě v Pensylvánii. I tento software je multiplatformní a spouští se přes příkazový řádek.

Konverzi lze také provést některými komerčními programy. Nabízí se např. **Adobe Audition**[12], který zvládne většinu formátů přečíst i konvertovat do běžnějších a standardních formátů.

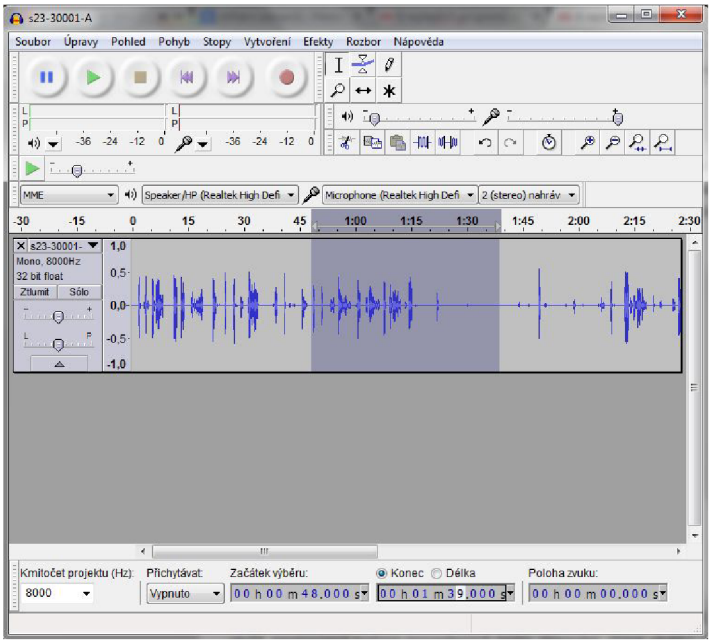
4.2.3 Střih zvukových záznamů

Zkrácení audio nahrávek bylo potřebné zejména kvůli testování Siri, Cortany a Google, protože diktování zabere delší čas a každá sekundová pauza v rozhovoru znamená ukončení poslouchání a čekání na další pokyn od uživatele. Více o tomto testování je rozepsáno v kapitole 5. Úprava délky také znamenala výběr částí rozhovoru, kde se plynule hovoří a vynechaly se tak části se zdlouhavým váháním a přemýšlením (typicky se jedná o slova uh, um, oh, yeah, která se do skórování nezahrnují).

[9] Dostupný na `http://www.itl.nist.gov/iad/mig//tools/`
[10] Dostupný na `http://sox.sourceforge.net/`
[11] Dostupný na `https://www.ldc.upenn.edu/language-resources/tools/sphere-conversion-tools`
[12] Dostupný na `https://creative.adobe.com/cs/products/audition`

Původní nahrávky telefonních rozhovorů měly délku od čtyř do deseti minut. U konferencí se délka pohybovala kolem 30 až 70 minut.

Pro stříhání a editaci zvukových záznamů je možné využít několik programů. Nejoblíbenější a také volně dostupný audio editor nese název **Audacity**. Je uživatelsky přívětivý, multiplatformní a umožňuje zvuky nahrávat, přehrávat, editovat, importovat a exportovat do několika formátů. Náhled softwaru můžeme vidět na obrázku 4.1.



Obrázek 4.1: Ukázka audio editoru Audacity

Celkem bylo nutné upravit a zkrátit 64 nahrávek. Více informací je v tabulce 4.1

	Počet záznamů	**Průměrná délka**	**Délka záznamů celkem**
Telefonní hovory 1 kanál	20	1 min 1 s	21 min 18 s
Telefonní hovory 2 kanály	31	56 s	28 min 54 s
Konference více mikrofonů	6	53 s	5 min 15 s
Konference 1 mikrofon	7	52 s	6 min 7 s
Celkem	64	3 min 45 s	61 min 36 s

Tabulka 4.1: Přehled zvukových záznamů

4.2.4 Zpracování nahrávek

Jelikož se netestovaly pouze aplikace Siri, Cortana a Google, ale je zahrnuto více služeb poskytující automatický přepis do textu, probíhalo testování mírně odlišně.

Služby dostupné přes webové rozhraní umožňující jednoduché vkládání připravených WAV zvukových souborů (patří sem služby Speechmatics, Spokendata a Voicebase) byly pro zpracování nahrávek nejjednodušší. U Speechmatics a Voicebase se navíc můžeme setkat s drag and drop nahráváním a je tak možné vložit několik záznamů najednou. Přepis do

textu u těchto služeb trvá při minutovém záznamu průměrně deset minut. Poté je možné přepis editovat a stáhnout na lokální disk v počítači. Každá služba ovšem nabízí stažení transkripce v různých formátech, v následující tabulce 4.2 je uvedeno, které formáty jsou k dispozici.

Speechmatics	JSON, TXT, XML
Spokendata	HTML, SRT, TRS, TXT, WebVTT, XML
Voicebase	PDF, RTF, SRT

Tabulka 4.2: Dostupné formáty přepisů

Pro získání přepisu z webových služeb, které neumožňují vložit jednoduše WAV soubor (týká se Talktyperu a Dictation) bylo zapotřebí připojit k počítači lepší mikrofon, protože se výkon integrovaného mikrofonu ukázal jako nedostatečně výkonný. K testování byl využit zapůjčený mikrofon značky Yenkee YMC1010BK.
Text z těchto služeb byl přímo ukládán do počítače ve formátu STM. Více informací o tomto formátu obsahuje sekce 4.3.

Aplikace Siri, Cortana a Google Now se testovaly mírně odlišným způsobem. Jedná se o osobní asistenty, jak již bylo zmíněno v kapitole 3, konkrétně v sekcích 3.1, 3.2 a 3.3.

Siri

Testování přepisu audia do textu pomocí osobní asistentky Siri prováděné na tabletu iPad Mini s verzí opračního systému 8.3, která byla v době vypracování této práce nejaktuálnější. Získání přepisu je možné několika způsoby, např. jako psaní e-mailu nebo poznámky.

Cortana

Pro získání automatického přepisu do textu byl původní záměr využít testovací verzi operačního systému Windows 10, avšak zde nastal problém s tím, že Cortana ještě nemá plně zpřístupněné některé funkce. Neumí zatím zaznamenávat poznámky či psát e-maily. Proto bylo přepisování prováděno na mobilním telefonu Nokia Lumia 920 s verzí operačního systému Windows Phone 8.1. Možnost získání transkripce je pouze jediný, a to psaním poznámky, která se následně uloží do integrované aplikace OneNote.

Google Now

Aplikace od Google je dostupná pro zařízení s iOS a Android, ale více možností, jak získat přepis audia do textu, nabízí telefony a tablety s OS Android. Pro testování byl využit tablet značky Lenovo A3500-FL s verzí Android 4.4.2. Podobně, jako je tomu u konkurenčních aplikací Siri a Cortana, Google Now umožňuje přepis nahrávky do textu pomocí psaní e-mailu nebo poznámky. Zapsaná poznámka se uloží do zařízení (např. do OneNote nebo Evernote).

4.3 Úprava přepisů

V dodané databázi byly obsaženy také referenční přepisy všech audio záznamů ve formátu STM.

4.3.1 Formát STM

Soubory uložené v tomto formátu (z angl. Segment Time Marked) jsou referenční, tedy obsahují textový přepis promluvy, který byl opravdu řečený. Používá se pro skórování a obsahuje několik informací o zvukové stopě (např. název zvukové stopy, kanál, id mluvčího atd.) [2].

```
single_0edc5a27 1 single_0edc5a27 150.165 150.995 with your parents
single_0edc5a27 1 single_0edc5a27 152.315 152.755 oh
single_0edc5a27 1 single_0edc5a27 158.805 159.285 nice
single_0edc5a27 1 single_0edc5a27 160.805 161.485 perfect
single_0edc5a27 1 single_0edc5a27 168.855 170.195 right wow that's awesome
```

Obrázek 4.2: Ukázka struktury soubor ve formátu STM

4.3.2 Odstranění slov váhání a sjednocení přepisů

Při skórování se často vypouští slova jako oh, uh, yeah, mm apod., protože není příliš důležité hodnotit, zda se správně zaznamenala. Druhým faktorem, proč se slova odstraňují je to, že ne všechny služby tato slova zahrnují do přepisů. Bylo tedy nutné přepisy sjednotit a slova eliminovat.

Každá aplikace poskytuje různé formáty přepisů a bylo potřeba je všechny sjednotit. Kromě Speechmatics, Spokendata a VoiceBase žádná ze služeb nezapisuje časy jednotlivých promluv. Bylo tedy nutné odstranit i časy a ostatní informace z referenčních přepisů. Ukázka rozdílnosti je znázorněna na obrázku 4.3.2.

Speechmatics: `SPEAKER: F1 Thank you thank you you are young person.`
Spokendata: `00:00.7 - 00:07.7 you sound like you are young person.`
VoiceBase: `1 00:00:00,11 --> 00:00:04,84 Thank you thank you you are young.`
Siri: `Thank you like you are you are you are you.`

Obrázek 4.3: Ukázka rozdílů mezi přepisy

4.3.3 Formát MLF

Po získání a úpravě všech přepisů bylo ještě nutné překonvertovat všechny transkripce do formátu MLF (z angl. Master Label File), který je určený pro skórování v programu `HResults`. Tento program je z balíku HTK[13] (Hidden Markov Model Toolkit), který byl vyvinut na půdě Univerzity v Cambridge [7]. O použití při testování pojednává kapitola 5. Na obrázku 4.4 je ukázka, jak soubor MLF vypadá.
Pro konverzi mi byl poskytnut jednoduchý skript v Bashi.

[13]Dostupný na `http://htk.eng.cam.ac.uk/`

```
#!MLF!#
"*/I_000000_000000.rec"
States
I
mean
"*/but_000000_000000.rec"
like
you’re
"*/and_000000_000000.rec"
to
Vermont
```

Obrázek 4.4: Ukázka souboru ve formátu MLF

Kapitola 5

Průběh testování

V této kapitole jsou uvedeny možnosti, jakým způsobem lze testování provádět a hodnotit. Představen je také využitý software `HResults` pro hodnocení úspěšnosti.

Vyhodnocování kvality systémů rozpoznávání řeči spočívá v porovnávání množiny rozpoznaných posloupností slov s množinou referenčnímích textů (tzn. tím, co bylo ve skutečnosti řečeno). Především nás zajímá úspěšnost rozpoznávání na úrovni promluv, slov nebo hlásek. Vyhodnocování úspěšnosti se provádí algoritmem a výsledkem je číslo, tabulka nebo statistický výstup.

5.1 HResults

Pro vyhodnocení úspěšnosti přepisu použitých aplikací byl zvolen program `HResults` ve verzi 3.4.1, který je dostupný na školním serveru Merlin, na němž běží operační systém Centos 6.6. Je to analytický program fungující na principu čtení značkových souborů a porovnávání s korespondujícími referenčními přepisy. Konkrétní použití softwaru `HResults` na dodané databázi je demonstrováno v kapitole 6.

Program je schopen vypočítat počet nahrazení, vymazání nebo vložení hlásek do slov. Pro analýzu výstupu rozpoznávače se využívá dynamické programování. Pokud tedy použijeme výpočet pro úspěšnost přepisu promluvy, získáme jednoduchý stastický výstup pro celý porovnávaný soubor.
Příklad výstupu:

```
--------------------------- Overall Results --------------------
SENT: %Correct=20.00 [H=2, S=8, N=10]
WORD: %Corr=66.67, Acc=58.97 [H=52,D=19,S=7,I=6,N=78]
================================================================
```

V prvním řádku výsledků je zobrazena úspěšnost celých vět, která funguje na principu porovnání všech značkovacích souborů, které jsou stejné jako značkovací soubory v referenčních přepisech. Úspěšnost rozpoznání slov demonstruje druhý řádek. Ve výsledcích můžeme sledovat několik hodnot.

Vysvětlení významu písmen:

- H – počet správně rozpoznaných slov,
- D – smazání jednoho znaku,
- S – náhrada jednoho znaku za jiný znak,
- I – vložení jednoho znaku,
- N – počet slov v referenčním souboru.

Určení procentuální úspěšnosti správně rozpoznaných slov je dáno následujícím vzorcem[14]:

$$\%Corr = \frac{H}{N} \times 100\,\%$$

Určení úspěšnosti celkové je dáno vztahem:

$$Acc = \frac{H - I}{N} \times 100\,\%$$

Program `HResults` se spouští přes příkazový řádek.
Příklad použití:

```
HResults [options] hmmList recFiles
```

Položka `hmmList` může obsahovat seznam modelů. `RecFiles` jsou přepisy ve formátu LAB (značkové soubory) nebo MLF. Ovšem pro skórování MLF souborů je zapotřebí přidat ještě parametr `-I` a použití poté vypadá následovně:

```
HResults -I hmmList recFiles
```

V balíku HTK je mnoho dalších programů pro zpracování výsledků z rozpoznávačů řeči. Obsahuje software pro trénování modelů, skript pro manipulaci se značkovacími soubory, konvertor do formátů potřebných pro hodnocení rozpoznávačů řeči a mnohé další. Seznam a také popis, k čemu jednotlivé programy slouží a jak se používají, můžeme nalézt v knize **The HTK Book** [7].

Pro testování kvality přepisů se využívají i jiné metody. Příkladem může být např. využití Levenshteinovy metody nebo výpočet intervalů spolehlivosti. Řeč lze také hodnotit pomocí clusterů. Pojem cluster značí třídu, do které daná promluva spadá. Pokud tyto informace známe, můžeme např. hodnotit, zda nahrávka namluvená ženou měla vyšší úspěšnost než nahrávka namluvená mužem [15].

[14] Vzorce vychází z publikace [7]

Kapitola 6

Vyhodnocení

Všechny výstupy z aplikací pro rozpoznávání řeči byly porovnány s referenčními soubory programem `HResults`. Zpracování probíhalo na školním serveru Merlin a výstupy byly zaznamenány do textových souborů.
Spouštění probíhalo tímto způsobem:

```
HResults -I reference.mlf /dev/null vystup.mlf > vysledek.txt
```

Obsah souboru `vysledek.txt` potom vypadá následovně:

```
====================== HTK Results Analysis =======================
Date: Sun May 14 13:46:18 2015
Ref : reference.mlf
Rec : vystup.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=0.00 [H=0, S=8, N=8]
WORD: %Corr=21.43, Acc=1.02 [H=21, D=36, S=41, I=20, N=98]
===================================================================
```

6.1 Telefonní rozhovory (1 kanál)

Z dodané databáze telefonních rozhovorů bylo zpracováno celkem 20 audio záznamů. V tabulce 6.1 jsou vypsány naměřené hodnoty. Uvedená čísla jsou v procentech a udávají procentuální úspěšnost správně rozpoznaných slov.
Seznam zkratek použitých v tabulkách 6.1, 6.2, 6.3 a 6.4:

- SM – Speechmatics,
- SD – SpokenData,
- TT – TalkTyper,
- VB – VoiceBase.

Průměrnou procentuální úspěšnost správně rozpoznaných slov demonstruje poslední řádek tabulky 6.1 níže.

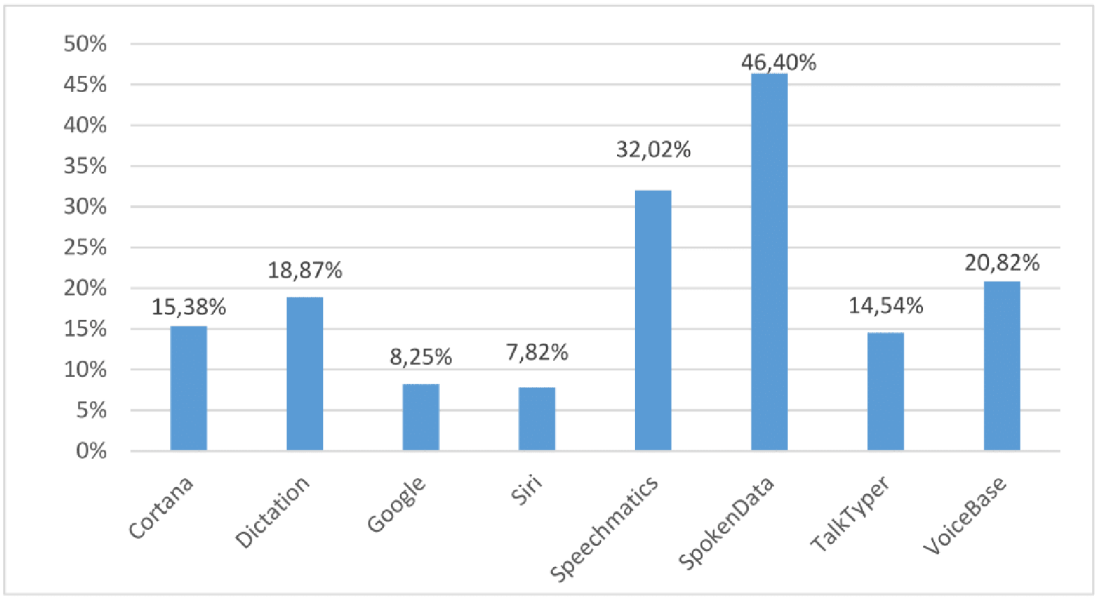
Pořadí	Cortana	Dictation	Google	Siri	SM	SD	TT	VB
1	27,5	22,5	4,8	4,1	47,5	54,56	23,11	35,77
2	0	6,76	2,7	0	3,28	53,54	8,11	8,11
3	6,17	13,58	1,23	2,49	12,35	26,41	6,64	6,38
4	27,03	35,14	4,05	4,05	37,84	66,67	26,16	14,78
5	4,7	8,72	0	0	9,4	51,36	0	0
6	9,5	9,91	0	0,9	15,32	41,19	4,5	5,45
7	16,43	18,1	11,2	7,92	28,42	56,16	14,33	11,64
8	11,2	12,3	7,89	8,3	41,2	36,54	19,2	26
9	24,61	22,5	13,8	14,51	51,8	56,1	19,6	21,3
10	3,1	7,98	0	1,18	27,1	26,14	7,97	17,84
11	23,85	28,9	14,49	15,34	46,19	58,17	22,19	31,94
12	16,54	17,33	8,95	7,35	32,11	49,17	19,36	21,49
13	23,11	24,38	11,14	10,99	28,7	46,13	19,6	17,32
14	7,71	8,98	5,95	6,03	12,06	32,11	5,55	23,65
15	19,97	21,17	11,59	12,34	44,87	53,18	16,14	27,49
16	17,5	26,16	13,2	9,8	40,62	50,9	21,07	24,6
17	7,27	14,52	6,56	7,41	30,11	23,09	11,76	15,43
18	12,8	18,11	10,79	8,27	38,6	49,13	7,28	30,7
19	20,27	30,9	20,18	20,8	44,47	43,72	19,13	41,3
20	28,27	29,5	16,4	14,58	48,42	53,66	19,01	35,16
Průměr	15,38	18,87	8,25	7,82	32,02	46,4	14,54	20,82

Tabulka 6.1: Výsledky telefonních rozhovorů (1 kanál)

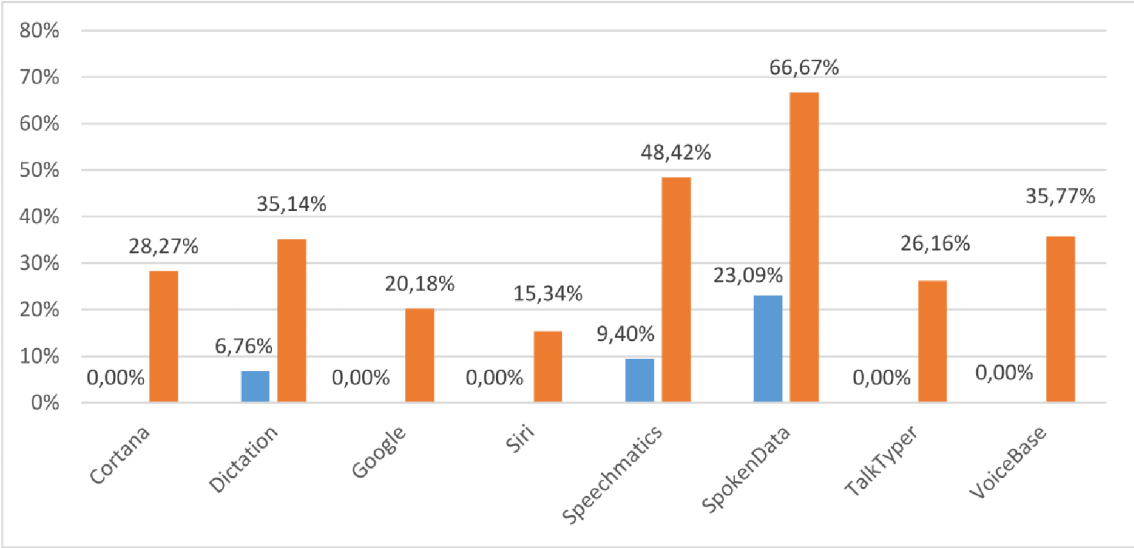
Naměřené průměrné hodnoty jsou pro lepší přehlednost zobrazeny ve sloupcovém grafu 6.1, ve kterém je zřetelné, že služba SpokenData průměrně nabývá téměř padesátiprocentní úspěšnosti správně rozpoznaných slov a dosahuje tak nejlepších výsledků ze všech.

V grafu 6.2 můžeme sledovat minimální a maximální naměřené hodnoty. V případě hodnoty rovné nule aplikace zachytily pouze pár slov, avšak v porovnání s referenčními výsledky se ukázalo, že zachycená slova nepatří do referenční množiny. Hodnoty kolísají právě z důvodu odlišných nahrávek. Při zašuměných a méně srozumitelných nahrávkách jsou hodnoty nižší. Naopak u nahrávek bez okolního ruchu rozpoznávače dosahují vyšších hodnot.

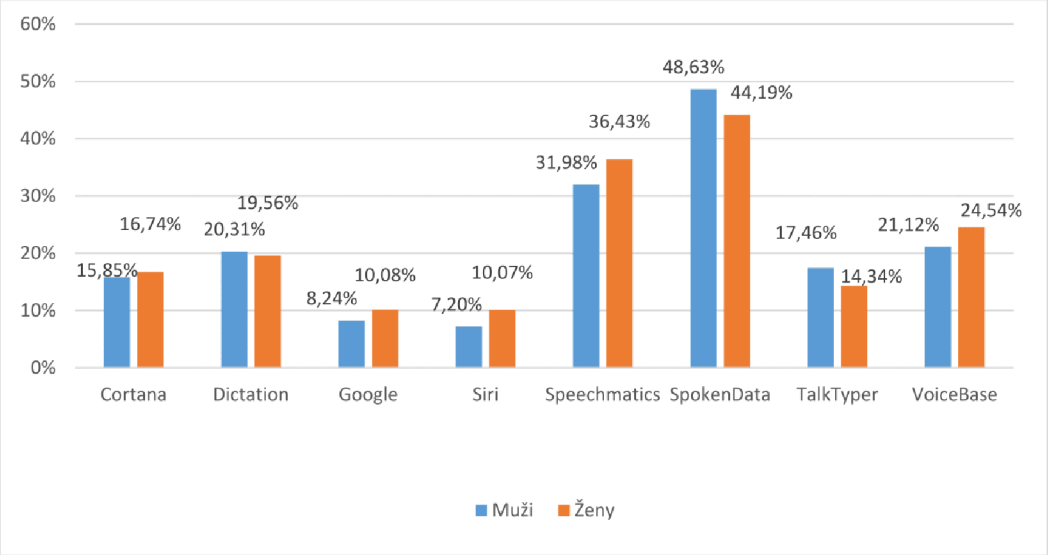
Pro srovnání je zde uveden i graf 6.3, kde nalezneme průměrné hodnoty úspěšnosti správně rozpoznaných slov u žen a mužů. Po důkladnějším prohlédnutí lze zjistit, že úspěšnost správně rozpoznaných slov je mírně vyšší u žen.



Obrázek 6.1: Graf reprezentující průměrné výsledné hodnoty z tabulky 6.1



Obrázek 6.2: Graf zobrazující minimální a maximální dosažené hodnoty



Obrázek 6.3: Rozdíl mezi úspěšností rozpoznání u žen a mužů

6.2 Telefonní hovory (2 kanály)

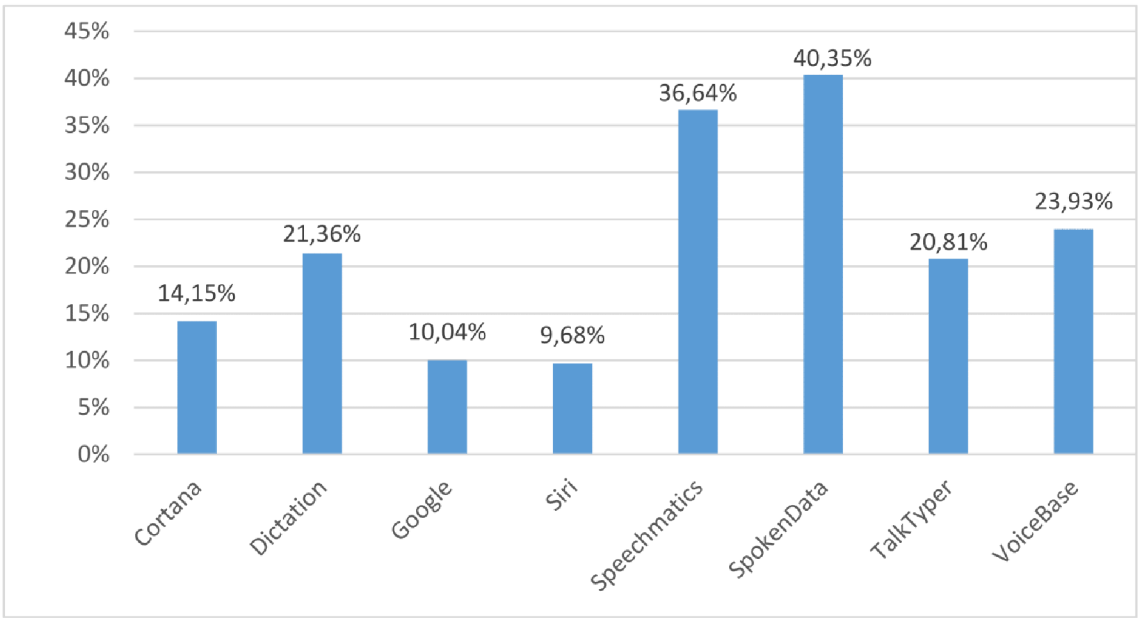
Pro dvoukanálové záznamy hovorů bylo zpracováno celkem 31 nahrávek. V tabulce 6.2 jsou zobrazeny všechny naměřené hodnoty v procentech. Poslední řádek tabulky reprezentuje průměrnou procentuální úspěšnost správně rozpoznaných slov. Pod tabulkou následuje graf 6.4, ve kterém jsou naneseny naměřené průměrné hodnoty úspěšnosti rozpoznání u jednotlivých služeb. I zde aplikace SpokenData dosahuje nejvyšších hodnot.

V grafu 6.5 stojí za povšimnutí minimální rozdíl u služeb SpokenData a Speechmatics mezi jejich nejvyššími naměřenými hodnotami.

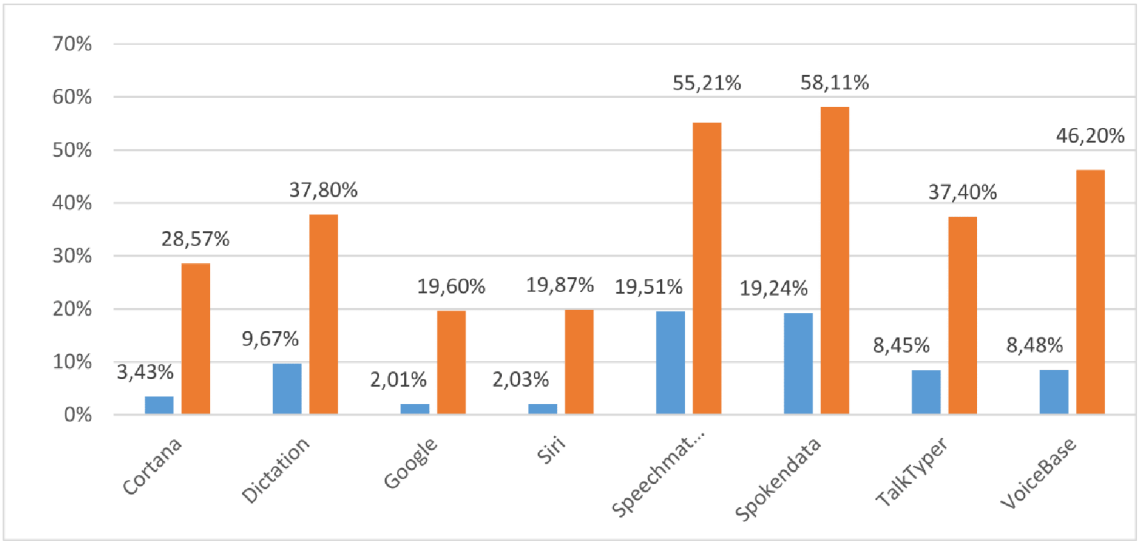
U těchto telefonních rozhovorů komunikovaly vždy dvě osoby. Bylo tedy možné mít kombinace muž – muž, žena – žena a muž – žena. V grafu 6.6 jsou zobrazeny průměrné hodnoty úspěšnosti u všech tří kombinací. Je zřejmé, že kombinace žena – žena má nejvyšší úspěšnost. Naopak kombinace muž – žena jsou vždy nejméně úspěšné.

Pořadí	Cortana	Dictation	Google	Siri	SM	SD	TT	VB
1	28,57	22,73	19,6	9,52	31,75	29,16	24,91	8,48
2	14,19	24,59	15,36	12,31	52,54	54,89	36,61	32,49
3	3,43	13,3	6,06	7,67	29,84	36,71	11,09	19,11
4	14,2	29,7	10,1	12,2	34,03	39,93	25,89	33,73
5	17,19	37,8	11,49	13,24	55,21	58,11	37,4	46,2
6	7,52	17,84	3,21	6,7	20,33	19,24	8,45	19,53
7	18,45	18,9	16,74	15,6	37,34	48,01	19,17	32,19
8	16,4	19,91	12,41	10,97	44,15	47,33	24,13	29,5
9	11,27	24,64	8,32	9,11	41,4	47,65	21,88	34,59
10	22,07	35,6	16,47	14,17	52,13	55,73	30,71	43,51
11	13,58	24,65	10,9	10,88	34,12	33,65	26,17	30,18
12	8,82	16,54	5,13	4,32	29,9	27,46	12,3	16,4
13	26,9	31,14	19,11	18,6	51,73	54,29	34,1	34,23
14	7,09	16,01	2,27	5,14	26,54	23,75	16,23	17,24
15	11,87	20,15	9,94	9,98	29,18	38,55	22,17	20,74
16	15,81	17,24	6,18	4,26	39,35	39,46	19,53	23,51
17	19,22	27,1	15,45	16,34	47,34	51,29	32,19	31,46
18	26,33	27,51	17,81	17,47	43,23	49,73	25,17	24,11
19	15,43	17,54	9,08	10,25	31,73	39,76	24,18	27,46
20	9,61	19,97	12,91	10,46	26,88	24,34	12,92	14,67
21	7,76	23,11	4,23	5,07	23,16	30,49	16,2	17,41
22	7,22	9,67	2,11	3,01	19,51	29,13	13,64	10,89
23	10,8	15,41	11,24	13,4	30,15	36,19	13,97	19,4
24	18,46	21,33	16,45	13,67	45,23	49	24,59	29,46
25	7,48	12,37	6,92	4,18	28,45	30,53	13,81	18,38
26	9,15	16,44	5,41	5,66	24,8	22,15	17,3	19,55
27	27,64	35,14	16,4	19,87	46,78	48,1	24,61	31,59
28	11,89	16,49	6,6	6,61	36,79	41,55	14,3	18,1
29	8,54	11,94	4,37	3,14	47,44	53,02	14,64	16,4
30	12,4	16,41	2,01	2,03	36,4	35,91	15,12	8,98
31	9,41	21,1	6,82	4,37	38,44	55,64	11,69	12,22
Průměr	14,15	21,36	10,04	9,68	36,64	40,35	20,81	23,93

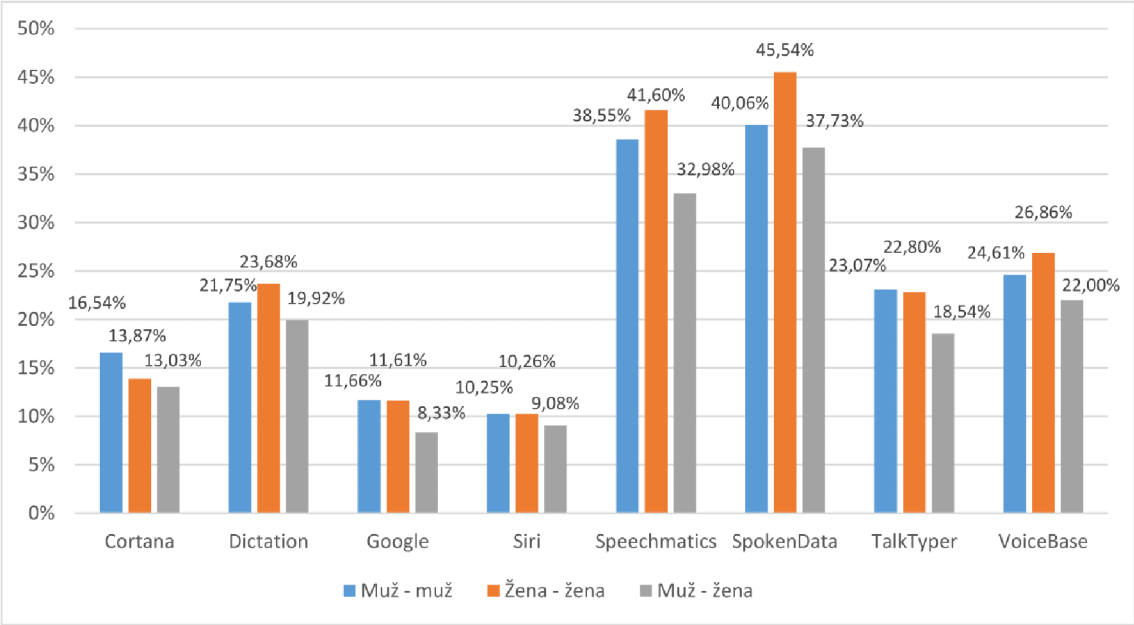
Tabulka 6.2: Výsledky telefonních rozhovorů (2 kanály)



Obrázek 6.4: Graf reprezentující průměrné výsledné hodnoty z tabulky 6.2



Obrázek 6.5: Graf zobrazující minimální a maximální dosažené hodnoty



Obrázek 6.6: Rozdíl mezi úspěšností rozpoznání u žen a mužů

6.3 Konference

U konferencí bylo zapotřebí zpracovat celkem 13 nahrávek z databáze. Konference jsou rozlišeny podle počtu mikrofonů.

6.3.1 Konference (více mikrofonů)

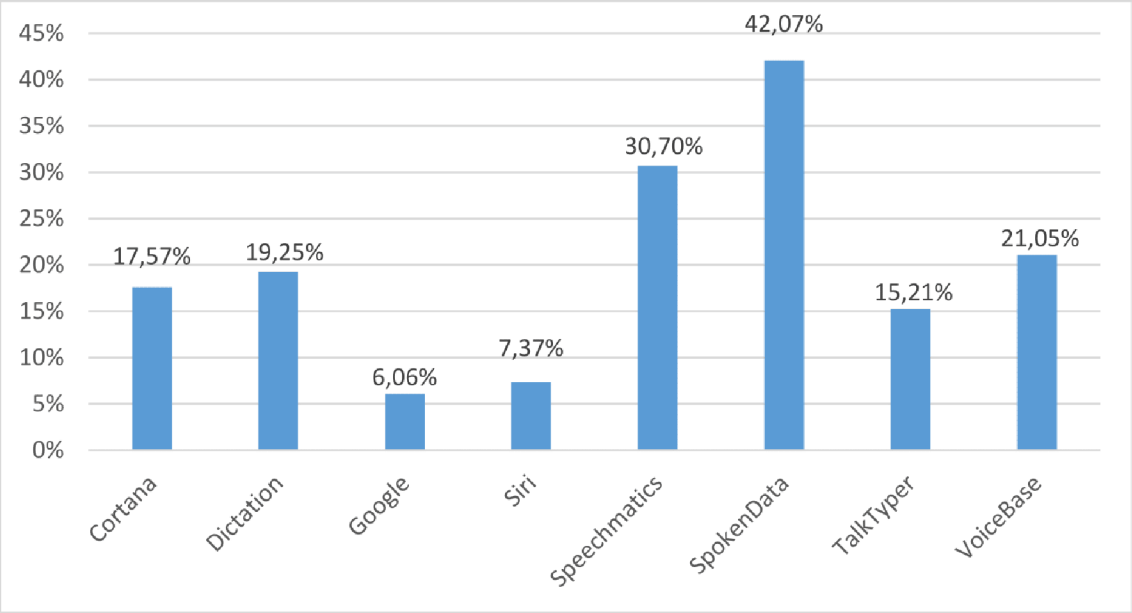
V tabulce 6.3 jsou zobrazeny výsledky šesti nahrávek a v posledním řádku se nachází průměrná procentuální úspěšnost správně rozpoznaných slov.

Na tomto testování lze snadno poznat, jak si rozpoznávače poradí s více hlasy (v tomto případě čtyři až pět hlasů) a také s překřikováním mluvčích. Promluvy totiž mohou snadno splývat a není již tak jednoznačně slyšitelné, kdy končí promluvy jedné osoby a začíná hovořit jiná osoba. V grafu 6.7, který reprezentuje výsledky z posledního řádku tabulky, si můžeme povšimnout, že průměrná procentuální úspěšnost správného rozpoznání není nijak výrazně nižší oproti výsledkům z telefonních hovorů.

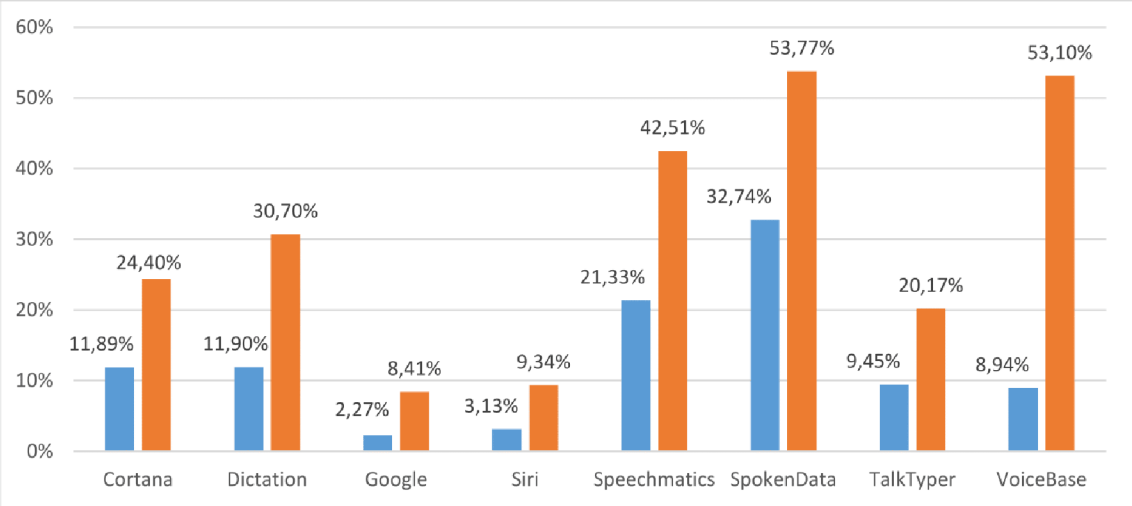
Z grafu 6.8 zobrazujícího minimální a maximální dosažené hodnoty úspěšnosti plyne, že nejvyšší úspěšnost klesla průměrně o sedm procent. Služba VoiceBase vykazuje naopak vyšší dosažené maximum oproti předešlým hodnotám v grafech 6.5 a 6.2.

Pořadí	Cortana	Dictation	Google	Siri	SM	SD	TT	VB
1	13,04	17,46	2,27	9,34	31,82	37,99	14,48	53,1
2	22,17	22,51	8,24	8,55	29,18	38,94	18,4	11,37
3	19,78	20,61	5,58	6,1	42,51	53,77	17,38	17,55
4	24,4	30,7	8,41	8,98	36,21	41,24	20,17	21,6
5	11,89	12,32	4,15	3,13	23,17	32,74	11,38	13,74
6	14,15	11,9	7,7	8,1	21,33	47,76	9,45	8,94
Průměr	17,57	19,25	6,06	7,37	30,7	42,07	15,21	21,05

Tabulka 6.3: Naměřené hodnoty u konferencí (více mikrofonů)



Obrázek 6.7: Graf reprezentující průměrné výsledné hodnoty z tabulky 6.3



Obrázek 6.8: Graf zobrazující minimální a maximální dosažené hodnoty

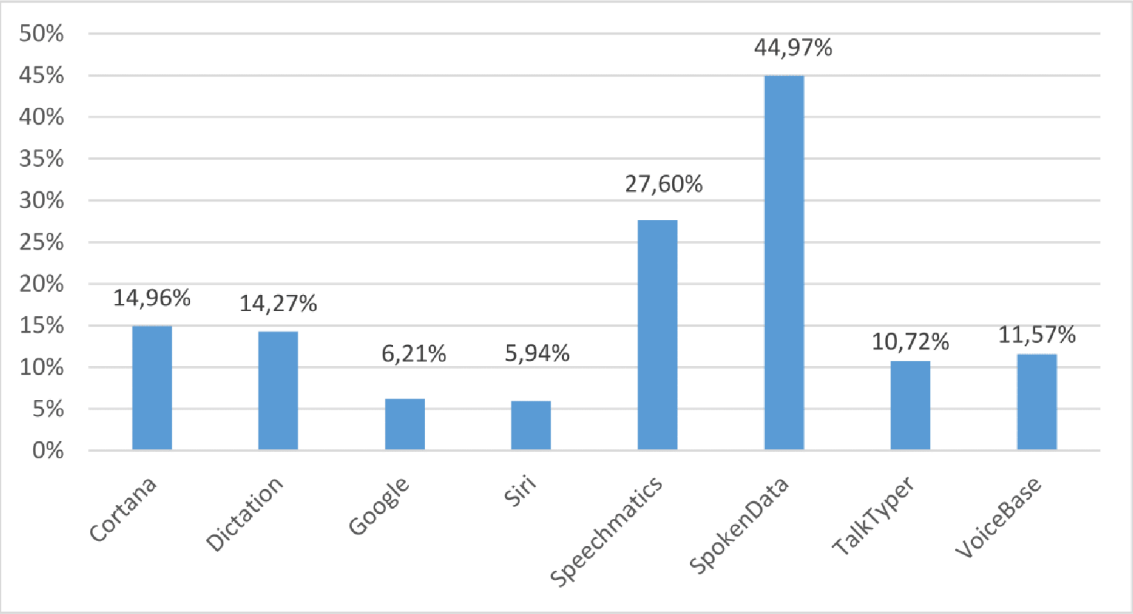
6.3.2 Konference (1 mikrofon)

U konferencí s jedním společným mikrofonem může být problémem různá vzdálenost mikrofonu od mluvčích. Nicméně, aplikace dosahovaly průměrných hodnot od šesti do čtyřiceti procent. Graf 6.9 zaznamenává mírný pokles průměrných hodnot v porovnání s grafem 6.7 u konferencí s mikrofony pro každého mluvčího.

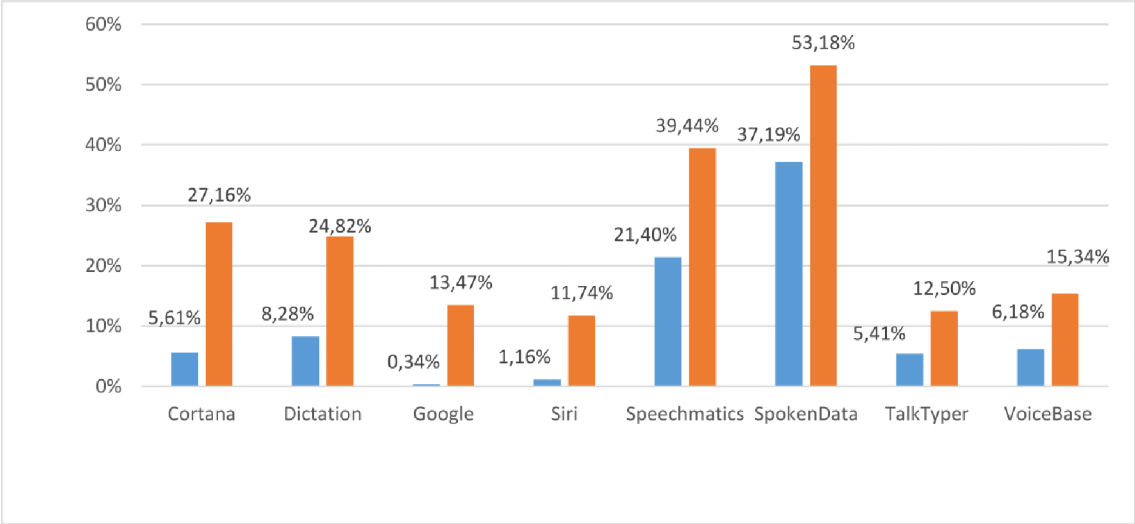
I v grafu 6.10 lze sledovat, že maximální možné hodnoty úspěšnosti poklesly ve srovnání s nejvyššími hodnotami u telefonních rozhovorů.

Pořadí	Cortana	Dictation	Google	Siri	SM	SD	TT	VB
1	12,53	13,08	5,39	4,79	25,01	42,79	5,41	8,91
2	27,16	24,82	13,47	11,74	39,44	53,18	12,5	13,81
3	8,99	10,11	6,1	7,28	26,14	44,94	11,34	13,12
4	5,61	10,86	5,29	3,11	21,4	44,91	12,23	15,34
5	18,1	21,34	7,98	8,39	30,69	49,91	10,6	13,82
6	7,2	8,28	0,34	1,16	22,39	37,19	12,09	9,84
7	25,1	11,4	4,89	5,11	28,11	41,9	10,88	6,18
Průměr	14,96	14,27	6,21	5,94	27,6	44,97	10,72	11,57

Tabulka 6.4: Naměřené hodnoty u konferencí (1 mikrofon)



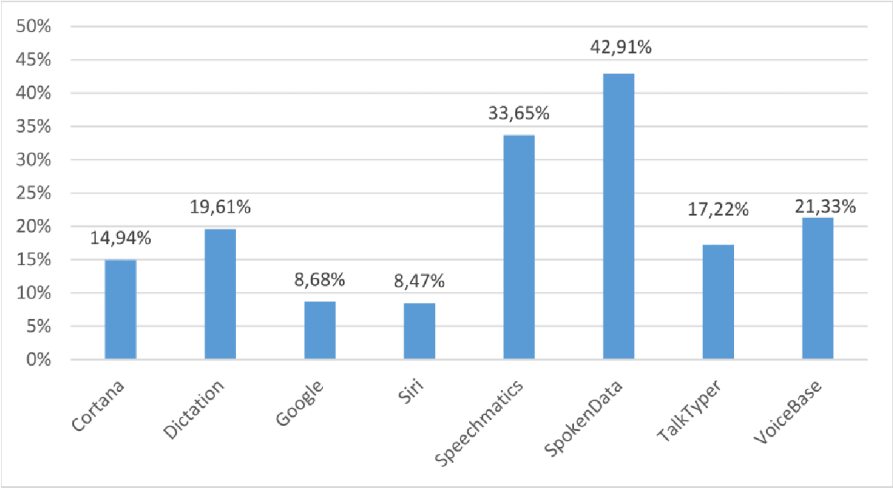
Obrázek 6.9: Graf reprezentující průměrné výsledné hodnoty z tabulky 6.4



Obrázek 6.10: Graf zobrazující minima a maxima dosažených hodnot

6.4 Celkové srovnání

Z předchozích grafů je zřejmé, že se při různých nahrávkách za rozličných podmínek úspěšnost správného rozpoznání mění. Se stoupající obtížností záznamů klesá procento úspěšnosti.
Celkové vyhodnocení je vidět v grafu 6.11. Nejvyšší úspěšnosti dosahovaly služby, jimž lze nahrávka poskytnout ve formě WAV souborů (nebo případně zvukové záznamy v jiných formátech). U diktovacích aplikací jako je Dictation a TalkTyper také velmi záleželo na zašumění nahrávek a nemálo ovlivnily výsledky vnější podmínky. Při použití výkonnějšího mikrofonu a kvalitních reproduktorů by úspěšnost byla jistě vyšší. Stejně tak tomu je i u osobních asistentů od společností Apple, Microsoft a Google.



Obrázek 6.11: Graf reprezentující průměrné výsledné hodnoty z tabulek 6.1, 6.2, 6.3 a 6.4

Kapitola 7

Závěr

Cílem této práce bylo srovnat úspěšnost přepisu řeči do textu u aplikací poskytující automatický přepis audio nahrávek do textu. Při realizaci bylo zjištěno, že tyto služby mají dnes vysokou úroveň a zpracované výsledky jsou toho důkazem.

Zpracování řeči je v současnosti velmi aktuální téma, ať už se jedná o syntézu lidského hlasu nebo přepis mluvené řeči do textu. Řečové systémy se totiž využívají v různých odvětvích – od ovládání strojů hlasem, přes čtení audio knih až k mluvení s virtuálním asistentem v mobilním telefonu. Vědy o řeči se vyvíjí již několik století a v podstatě jsou ve vývoji neustále až do současnosti. Nejvíce zvládnutá oblast je právě vytváření umělé řeči. Postup v těchto oblastech velkou měrou závisí na rozvoji nových technologií a výpočetní technice. Tento vývoj jde však velmi rychle kupředu a lze sledovat čím dál častější představování nových produktů na různých technických konferencích a výstavách. Výrobci a firmy vedou konkureční boj, přebírají mezi sebou odborníky, pořizují nové technologie a patenty a snaží se přijít na trh v co nejkratším čase s novinkou, která zaujme zákazníky. Takto se často projevují právě společnosti, jejichž aplikace byly zkoumaným předmětem této práce.

Nicméně osobní asistenti příliš neuspěli v takových testech. Mnohem více se vyznamenaly služby, které jsou dostupné přes webové rozhraní. Nejvyšší úspěšnost při rozpoznávání dosahovala služba SpokenData, následují služby Speechmatics a VoiceBase. Při komunikaci s vývojáři Speechmatics bylo zjištěno, že mezi firmami Speechmatics a VoiceBase je navázána spolupráce a některé technologie mají tyto služby společné či postavené na podobném základu. Využití osobních asistentů je vhodnější zejména pro kratší příkazy a ovládání některých funkcí mobilních zařízení hlasem. V případě diktování je nutné komunikovat s přístrojem přímo, bez přílišného šumu a také bez delších pauz. Dalším omezením mohou být jazyky. Je však jen otázkou času, kdy budou služby a aplikace využitelné i pro češtinu a další jazyky.

Uplatnění řečových technologií je možné i na dalších zařízeních. Příkladem mohou být chytré hodinky, brýle nebo samořídicí auto společnosti Google, kterému lze nadiktovat např. cílovou adresu. Velký ohlas vzbuzují také technické novinky usnadňující život handicapovaným lidem. Jde o automatické titulkování televizních pořadů, automatický překlad znakové řeči do češtiny, převod mluvené řeči do znakové řeči a jiné oblasti. Dalším okruhem uplatnění znalostí o řečových systémech může být real-time překlad mluvené řeči do cizího jazyka. Tuto novinku nedávno představila společnost Microsoft v souvislosti s uvedením nového programu Skype Translator Preview. Zde by jistě stálo za vyzkoušení takový software otestovat na podobných datech, jako tomu bylo v této práci.

Seznam obrázků

Seznam tabulek

Literatura

[1] Apple - Informace o Siri. https://support.apple.com/cs-cz/HT204389.

[2] NIST - Questions and Answers concerning the Speech Recognition Task.T. http://www.itl.nist.gov/iad/mig/tests/sdr/2000/SRT_FAQ.html.

[3] Wikipedie - Google Now. http://en.wikipedia.org/wiki/Google_Now.

[4] WindowsPhone - Cortana. http://www.windowsphone.com/cs-cz/how-to/wp8/cortana/meet-cortana.

[5] Ing. Schimmel, J.: Formáty zvukových souborů na PC. *Elektrorevue*, 2007, ISSN 1213-1539, [Online], [cit. 2015-05-16].
URL `http://www.elektrorevue.cz/clanky/01007/index.html#riff`

[6] Kolektiv autorů: *Pravidla českého pravopisu.* Academia, 2005, ISBN 80-200-1327-X.

[7] Kolektiv autorů: *The HTK Book.* Cambridge University Engineering Department, 2006.
URL `http://htk.eng.cam.ac.uk/prot-docs/htkbook.pdf`

[8] Mišun, V.: *Tajemství lidského hlasu.* Vysoké učení technické v Brně - Nakladatelství VUTIUM, 2010, ISBN 978-80-214-3499-8.

[9] Prof. Ing. Sigmund, M., CSc.: *Rozpoznávání řečových signálů.* Vysoké učení technické v Brně, 2007, ISBN 978-80-214-3526-1.

[10] Prof. Ing. Uhlíř J., CSc. a kolektiv: *Technologie hlasových komunikací.* Praha: Nakladatelství ČVUT, 2007, ISBN 978-80-01-03888-8.

[11] Psutka J.; Müller L.; Matoušek J.; Radová V.: *Mluvíme s počítačem česky.* Praha: Academia, 2006, ISBN 80-200-1309-1.

[12] Sgall, P.; Daneš, F. a kolektiv: *Cesty moderní jazykovědy: jazykověda a automatizace.* Orbis, 1964.

[13] Sgall, P.; Hajičová, E.; Piťha, P.: *Učíme stroje česky.* Praha: Panorama, 1982.

[14] Yemchenko, V.: *Moderní technologie pro ovládání PC hlasem.* Bakalářská práce, Vysoká škola ekonomická v Praze, 2011.

[15] Řezáček, P.: *Automatické vyhodnocení výstupů systému rozpoznávání řečit.* Bakalářská práce, Západočeská univerzita v Plzni, 2012.

Dodatek A

Obsah CD

- Technická zpráva ve formátu PDF v adresáři `/thesis/`
- Zdrojové kódy technické zprávy ve formátu LaTeXv adresáři `/thesis-latex/`
- Plakát v adresáři `/poster/`
- Video v adresáři `/video/`

Dodatek B

Plakát