VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
FACULTY OF INFORMATION TECHNOLOGY

ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

SYNCHRONIZACE TEXTU A AUDIA

TEXT TO AUDIO ALIGNMENT

BAKALÁŘSKÁ PRÁCE
BACHELOR'S THESIS

AUTOR PRÁCE — VOJTĚCH ŠIKULA
AUTHOR

VEDOUCÍ PRÁCE — Ing. IGOR SZŐKE, Ph.D.
SUPERVISOR

BRNO 2017

Vysoké učení technické v Brně - Fakulta informačních technologií

Ústav počítačové grafiky a multimédií Akademický rok 2016/2017

Zadání bakalářské práce

Řešitel: **Šikula Vojtěch**
Obor: Informační technologie
Téma: **Synchronizace textu a audia**
Text to Audio Alignment
Kategorie: Zpracování řeči a přirozeného jazyka

Pokyny:

1. Nastudujte principy automatického zarovnání textu k audiu.
2. Navrhněte postupy, jak automaticky zarovnat text s odpovídajícím audiem. Uvažujte i postupy nezávislé na jazyku.
3. Implementujte navržené postupy. Zarovnejte vybraná data a porovnejte úspěšnost jednotlivých postupů.
4. Identifikujte místa (příčiny) selhání navržených postupů a pokuste se je vylepšit.
5. Znodnoťte dosažené výsledky a navrhněte směry dalšího vývoje.
6. Vyrobte A2 plakátek a cca 30 vteřinové video prezentující výsledky vaší práce.

Literatura:

- Dle pokynů vedoucího

Pro udělení zápočtu za první semestr je požadováno:

- Body 1 a 2, a část bodů 3 a 4 ze zadání.

Podrobné závazné pokyny pro vypracování bakalářské práce naleznete na adrese http://www.fit.vutbr.cz/info/szz/

Technická zpráva bakalářské práce musí obsahovat formulaci cíle, charakteristiku současného stavu, teoretická a odborná východiska řešených problémů a specifikaci etap (20 až 30% celkového rozsahu technické zprávy).

Student odevzdá v jednom výtisku technickou zprávu a v elektronické podobě zdrojový text technické zprávy, úplnou programovou dokumentaci a zdrojové texty programů. Informace v elektronické podobě budou uloženy na standardním nepřepisovatelném paměťovém médiu (CD-R, DVD-R, apod.), které bude vloženo do písemné zprávy tak, aby nemohlo dojít k jeho ztrátě při běžné manipulaci.

Vedoucí: **Szőke Igor, Ing., Ph.D.**, UPGM FIT VUT
Datum zadání: 1. listopadu 2016
Datum odevzdání: 17. května 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
Fakulta informačních technologií
Ústav počítačové grafiky a multimédií
612 66 Brno, Božetěchova 2

doc. Dr. Ing. Jan Černocký
vedoucí ústavu

Abstrakt

Tato bakalářská práce se zabývá synchronizací textu a audia. Jsou zde prezentovány jednotlivé práce, které se zabývaly daným problémem. K vyhodnocení jsou použita data z MGB Challenge 2015. Zkoumaná technika využívá fonémového přepisu a jeho následné zarovnání s přepisem. Zarovnání bylo provedeno podle různých modelů. Nejlepších výsledků bylo dosaženo při průniku dvou zarovnání podle modelů z kvalitních nahrávek.

Abstract

This bachelors thesis is dealing with text to audio alignment. I present here works which are dealing with same problem. For evaluation have been used data from MGB Challenge 2015. Technique used here is using phoneme transcription and its alignment with transcript. Alignment was done with different models. The best results have been achieved by intersection of two alignments from models from good records.

Klíčová slova

synchronizace textu a audia, fonémový rozpoznávač, MGB Challenge, G2P

Keywords

text-to-speech alignment, phoneme recognition, MGB Challenge, G2P

Citace

ŠIKULA, Vojtěch. *Synchronizace textu a audia.* Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Szőke Igor.

Synchronizace textu a audia

Prohlášení

Prohlašuji, že jsem tuto bakalářskou práci vypracoval samostatně pod vedením pana Igora Szőke. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

. . . . . .
Vojtěch Šikula
28. července 2017

Poděkování

Rád bych poděkoval Ing. Igorovi Szőkemu, Ph.D., za vedení této bakalářské práce, odbornou pomoc a rady.

Obsah

1 Úvod **2**

2 Zarovnání textu a audia **3**

3 G2P konverze **6**
3.1 Základní techniky 6

4 Fonémový rozpoznávač **8**

5 MGB Challenge **10**
5.1 Task 2 – zarovnání 10
5.1.1 Data 10
5.1.2 Skórování 11

6 Použité nástroje **12**

7 Experimenty **14**
7.1 Příprava dat 14
7.2 Porovnání ručního přepisu a automatického přepisu 15
7.3 Zarovnání 18
7.4 Zarovnání podle modelu `n-té` nahrávky 21
7.5 Spojení nahrávek 23
7.6 Zarovnání podle modelu `n-nejlepších` nahrávek 24
7.7 Kombinace dobrých a špatných nahrávek 25
7.8 Zarovnání podle Maďarského fonémového přepisu 27
7.9 Zarovnání podle modelu `n-té` nahrávky s maďarským fonémovým přepisem 29
7.10 Průnik zarovnaných nahrávek 29
7.11 Zarovnání grafému na grafém 32

8 Závěr **33**

Literatura **35**

Přílohy **37**

A Obsah přiloženého paměťového média **38**

Kapitola 1

Úvod

Tato práce se zabývá problémem zarovnání textu a audia. Jedná se o problém přiřazení k jednotlivým slovům přepisu čas jejich vyslovení v audiu. Nejedná se o tak žádaný úkol jako automatické rozpoznávání řeči, ale přesto má mnoho užitečných použití v mnoha aplikací. Zarovnání textu a audia může být využito v indexaci mluvených materiálů pro vyhledávání. Časové značky mohou být použity k automatické tvorbě titulků nebo pro automatické dolování a třídění trénovacích dat [11].

V kapitole 2 jsou uvedeny základní postupy a následné shrnutí několika prací, které se tímto problémem zabývaly. V kapitole 3 je vysvětlen základní popis převodu textu do jeho fonémové reprezentace, takzvaná G2P konverze (Grapheme-to-Phoneme conversion). Kapitola 4 je naopak věnována fonémovému rozpoznávači (Phoneme rocognition), který k audiu vygeneruje jeho fonémový přepis. Je-li tedy text i audio převedeno do stejné reprezentace, zbývá tedy jen zarovnat tyto dva přepisy na sebe. Využité metody, provedené experimenty a jejich výsledky jsou uvedeny v kapitole 7. Tato práce pracuje s daty z MGB Challenge 2015, kapitola 5 udává základní informace o ní. V kapitole 8 je popsáno zhodnocení a možnosti dalšího pokračování.

Kapitola 2

Zarovnání textu a audia

Úkolem zarovnání textu a audia se v poslední 20 letech zabývalo mnoho prací. Může se zdát, že se jedná o triviální problém, ale není tomu tak. Důkazem může být právě MGB Challenge z roku 2015, kde bylo zarovnání textu a audia jedním z úkolů. Existují různé důvody, kdy má tento úkol uplatnění. Jedním z možných uplatnění je indexace a vyhledávání v audiovizuálních nahrávkách (vyhledává se v textu a odtud vedou odkazy do audia), dále se hojně využívá v úlohách zpracování přirozeného jazyka.

Řešení tohoto problému můžeme rozdělit podle délky nahrávky. Je-li délka nahrávky v řádech několika vět, tak se používá takzvaného Nuceného zarovnání (Forced Alignment). Idea Nuceného zarovnávání spočívá v tom, že se provádí automatické rozpoznávání řeči, ale rozpoznávač je nucen rozpoznat jen tak, aby výsledek odpovídal přepisu. Tato technika dosahuje dobrých výsledků v kratších nahrávkách, ke kterým existuje kvalitní přepis. Nevýhodou je, že náročnost tohoto postupu se zvyšuje kvadraticky s délkou nahrávky a v případě nepřesného přepisu dosahuje neuspokojivých výsledků. Pro delší nahrávky se využívá podobný postup, ale je třeba řešit výše zmíněné nedostatky[11, 1].

Zde jsou prezentovány některé práce, které se danou problematikou zabývají:

AUTOMATIC GENERATION OF HYPERLINKS BETWEEN AUDIO AND TRANSCRIPT[15]

Tato práce byla publikovaná již v roce 1997, a mnoho dalších se na ni odkazuje. Řeší problém zarovnání audia a textu v televizních pořadech. Jejich cílem není naprosto přesné zarovnání. Mezi jejich požadavky je, aby jejich systém neměl limit na délku zarovnávaného audia (až několik hodin), aby měl přesnost zarovnání na 1 nebo 2 sekundy a aby dokázal pracovat s nahrávkami s více řečníky a šumem okolí. Dále se jejich systém musí vyrovnat s nedokonalým přepisem. Jimi vyvinutý systém se nazývá ALTA (Automatic Linking Of Transcript And Audio) a skládá se z těchto částí:

- Analýza – nalezení vhodných míst pro segmentaci
- Segmentace – rozdělení audia a přepisu na menší části
- Jazykový model – z jednotlivých segmentů vytvořit jazykový model s využitím G2P
- Automatický rozpoznávač řeči – ze segmentu audia vytvořený fonémový přepis, pomocí skrytých markovských modelů

- Zarovnání na úrovni fonémů
- Rozhodnutí správnosti zarovnání, jestliže si systém není jistý, není tento úsek použit

Výsledky – na hodinové nahrávce nebylo zarovnáno 15 % slov a 95 % zarovnaných slov bylo v rozmezí 3 sekund.

A RECURSIVE ALGORITHM FOR THE FORCED ALIGNMENT OF VERY LONG AUDIO SEGMENTS[14]

Tato práce představuje rekurzivní algoritmus pro zarovnání nahrávek delších jak 1 hodina. Tento algoritmus pracuje s audiem reprezentovaným pomocí 13-ti Mel-Kepstrálními koeficienty dále zpracovaným pomocí rozpoznávače řeči SPHINX II.

Výsledný text se zarovná dynamickým programováním pomocí nalezení globálně nejlepšího zarovnání. Z tohoto zarovnání jsou vybrány místa s největší pravděpodobností správného zarovnání. Tímto se text a audio rozdělí na zarovnané a nezarovnané úseky. Na nezarovnaných úsecích se celý tento postup znovu aplikuje. Algoritmus končí pokud jsou zarovnány všechny úseky, nebo pokud nezarovnané úseky jsou menší než zadaná mez.

Důležitým aspektem tohoto postupu je, že zarovnání, která jsou s vysokou pravděpodobností chybná, jsou oddáleny do pozdějších iterací a nemají tedy vysokou důležitost.

Tento postup byl vyzkoušen na asi tříhodinové nahrávce obsahující kvalitní a méně kvalitní úseky. 98 % procent slov bylo zarovnáno s chybou do 0,5 sekundy.

TEXT-TO-SPEECH ALIGNMENT OF LONG RECORDING USING UNIVERSAL PHONE MODEL[9]

Cílem této práce je vytvořit co nejmenší fonémový model, který je schopen zarovnat s vysokou přesností text k audiu, prezentované řešení je do značné míry jazykově nezávislé.

Jednotlivé fonémy jsou klasifikovány pomocí HMM do skupin (plosivy, frikativy, ...), text se převede pomoci G2P. Audio se rozdělí na vhodné úseky podle míst ticha a na těchto úsecích probíhá nucené zarovnání.

Výsledný systém byl vyzkoušen na audio-knihách různých jazyků (angličtina, němčina, francouzština, finština a bulharština), kde při klasifikaci samohlásek do osmi skupin a souhlásek do čtyř skupin dosáhly zarovnání na věty se standardní odchylkou menší než 0,5 s pro všechny testované jazyky kromě bulharštiny (1,3 s).

AUTOMATIC ALIGNMENT AND ERROR CORRECTION OF HUMAN GENERATE TRANSCRIPTS FOR LONG SPEECH RECORDINGS[12]

Tato práce zkoumá chyby, které mohou vzniknout při ručním přepisování. A dále přestavuje způsob jeho zarovnání s audiem s korekcí chyb. Představený způsob se skládá ze tří kroků: Pomocí automatického rozpoznávače řečí se získá přepis nahrávky. Rozpoznávač je ovlivněn ručním přepisem. Automatický přepis je zarovnán s ručním přepisem a rozdělen po úsecích, kde došlo ke shodě. V dalším kroku se úseky zarovnají pomocí upraveného nuceného zarovnání, který dovoluje vložení, změnu a smazání slov. V poslední kroku jsou jednotlivé segmenty z prvního kroku znovu vyhodnoceny automatickým rozpoznávačem, který může upravit slova podle informací získaných z předchozího kroku.

Systém byl vyzkoušen na audiu ze tří přednášek, kde se pomocí vyvinutého systému snížila jejich chybovost o 14 %.

A SYSTEM FOR AUTOMATIC ALIGNMENT OF BROADCAST MEDIA CAPTIONS USING WEIGHTED FINITE-STATE TRANSDUCERS[13]

Jedná se o systém vyvinutý na universitě v Edinburgu, který se zúčastnil MGB Challenge 2015. Jako mnoho automatických řečových rozpoznávačů využívají dekodér založený na vážených konečných převodníků (WFST – Weighted Finite State Transducer). Jeden převodník se skládá z gramatiky, ze slovníku, z kontextové závislosti a ze skrytých markovských modelů. Dekódování je provedeno nalezením nejkratší cesty daného převodníku.

Systém funguje na dvojitém procházení přes data. V prvním průchodu se z celé nahrávky vytvoří konečný převodník, kde každé slovo je potenciální vstup a konečný stav, vyhovuje všem podřetězcům původního textu. Ve druhém průchodu se používá deterministický převodník, který dovoluje vynechání slov, přidáním přechodu `skip` s pevně danou pravděpodobností. Konečné převodníky byly implementovány pomocí Kaldi tools. Použitý akustický model byl trénován na 640 hodinách z MGB Challenge training set [8]. Systém dosáhl 0,8965 f-hodnoty na MGB vývojové sadě.

Kapitola 3

G2P konverze

Tato kapitola se zabývá konverzí grafému na foném, vychází převážně z článku Joint-sequence models for grapheme-to-phoneme conversion[10]. G2P je zkratka od Grapheme To Phoneme což by se do češtiny mohlo přeložit jako „grafém na foném“, jelikož v češtině neexistuje zavedený pojem, budeme v dalším textu používat výhradně zkratku G2P. Grafém je základní, dále již nedělitelná jednotka psaného jazyka[4]. Mezi grafémy se řadí například písmena, číslice, interpunkční znaménka, atd. Foném je naopak nejmenší jednotka mluveného jazyka[5], která má rozlišovací funkci v daném jazyce. Úkolem G2P konverze je tedy nalezení správné výslovnosti k danému textu. Má důležitou roli ve zpracování přirozeného jazyka především v syntéze nebo rozpoznávání řeči.

Kdyby měl jazyk čistě fonetický pravopis, mezi grafémem a fonémem by existoval vztah 1 ku 1. K takovým jazykům se například řadí umělý jazyk esperanto. Nicméně většina jazyků nemá fonetický pravopis a vztah mezi grafémem a fonémem je nejasný a závisí na kontextu.

Vlastní G2P konverzi předchází normalizace – expanze zkratek a čísel.

3.1 Základní techniky

- *Přímé mapování*

  Má-li jazyk fonetický pravopis, tak se grafém shoduje s fonémem a G2P konverze není potřeba (esperanto, z části čeština), většina jazyků však používá nefonetický pravopis.

- *Vyhledávání ve slovníku*

  Mezi nejjednodušší techniku patří vyhledávání ve slovníku. Ač dosahuje dobrých výsledků má mnohá úskalí. Tvorba takových slovníků je značně náročná a zabírá velké množství paměti, takže je její využití nepraktické, zvláště pak v mobilních a vestavěných systémech.

- *Pravidlový převod (rule based)*

  Pravidlový převod naopak převádí text na fonémy pomocí předem definovaných pravidel, například pomocí konečných převodníků. Jako doplněk využívá slovník výjimek, který obsahuje slova, která se neřídí podle definovaných pravidel (např. přejatá slova). Nevýhoda tohoto přístupu je složitost tvorby daných pravidel (je zapotřebí odborník), je potřeba brát v úvahu různé nepravidelnosti.

- *Převod řízený daty (data-driven aproach)*

  Hlavní myšlenka toho přístupu je, jestliže poskytneme dostatek příkladů, mělo by být možné předpovědět převod na základě podobnosti se slovy jejichž převod zná. Hlavní výhodou je, že místo náročného vytváření pravidel stačí pouze poskytnout dostatek příkladů. Hlavní techniky:

  - Lokální klasifikace

    předpokládá zarovnání trénovacích dat po jednotlivých písmenech. Jedno písmeno může být zarovnáno k 0 (epsilon), 1 nebo více fonémům, takzvané zarovnání 1-k-n, toto zarovnání se vytváří například pomocí dynamického programování. Vlastní převod se prování sekvenčně zleva doprava pro každé písmeno. Ke každému písmenu se přiřadí jedna z množiny povolených sekvencí fonémů. Výběr se provádí na základě kontextu daného písmene (odtud název lokální klasifikace), nejčastěji pomocí neuronových sítí nebo rozhodovacích stromů.
  - Největší podobnosti (nearest-neighbor). Tato technika je založena na hledání v trénovacích datech slova nebo části slov, které se nejvíce podobají slovu, které má být přepsáno. Výsledný přepis je odvozen právě z těchto nalezených slov.
  - Další možností je použití pravděpodobnostních metod.

Kapitola 4

Fonémový rozpoznávač

Tato kapitola je založena podle[16]. Fonémový rozpoznávač hraje důležitou roli v automatickém zpracování přirozeného jazyka. Využívá se například v detekci řeči, hledání klíčových slov nebo identifikaci jazyka. V této práci se využívá k zarovnání textu a audia.

Fonémový rozpoznávač převádí audio signál na posloupnost fonémů. Je třeba uvést, že fonémový rozpoznávač je závislý na jazyku, různé jazyky mohou mít různé fonémy, například španělština má dva druhy hlásky r (ere a erre), které dokáží odlišit od sebe různá slova (např. pero - ale, perro - pes), ale v češtině má slovo vždy stejný význam, bez ohledu na to jestli se vysloví s ere nebo erre. Fonémové rozpoznávače kromě fonémů rozeznávají ticho a neřečový hluk.

Příklad výstupu fonémového rozpoznávače:

```
000000 5200000 pau -43.996017
5200000 6200000 int -21.674377
6200000 7200000 u -22.883202
7200000 8500000 n -21.741669
8500000 9200000 f -21.345512
```

První číslo udává čas začátku fonému, druhé číslo čas konce, dále je rozeznaný foném a poslední číslo udává s jakou věrohodností se doopravdy jedná o daný foném.

Existují fonémové rozpoznávače vyvinuté nad různými principy. Nejčastější jsou fonémové rozpoznávače používající neuronové sítě nebo skryté markovské modely (HMM).

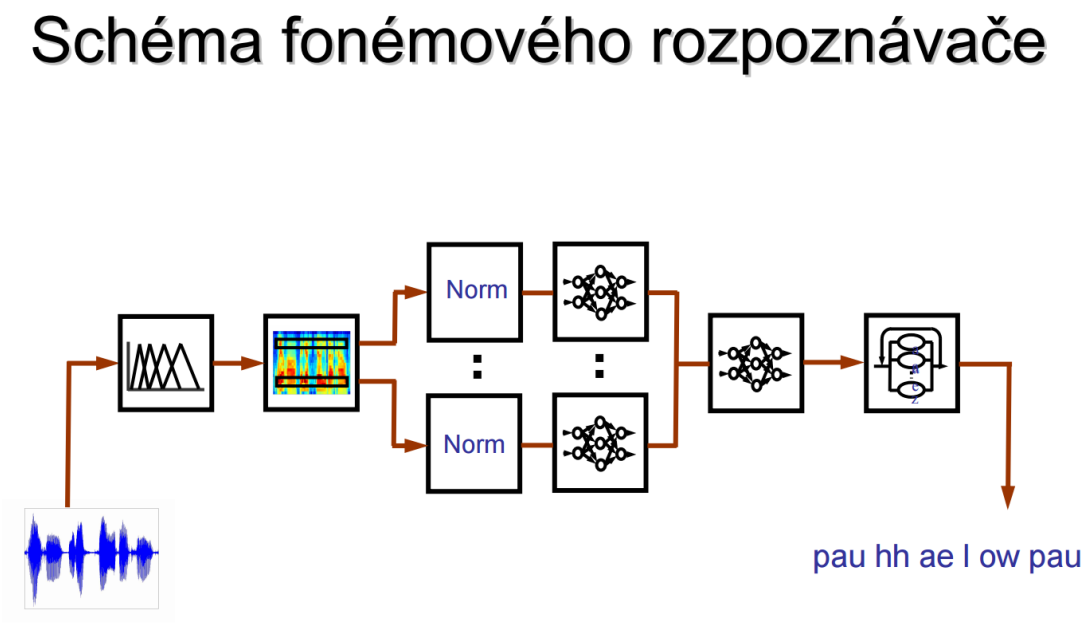
Než se použije některá z výše uvedených metod je třeba vhodně upravit data. Jedná se hlavně o zredukování množství dat. Toho je dosáhne rozdělení na segmenty dlouhé například 20 ms a překrývající se o 10 ms. Tyto segmenty jsou modulovány Hammingovým oknem:

$$w(n) = a - b * cos(\frac{2\pi n}{N-1}),$$

kde

- `a` a `b` jsou vhodné konstanty
- `N` počet vzorků v segmentu

Na tato data je dále aplikována Mel-Frekvenční kepstrální analýza, která zredukuje počet dat typicky na 13 koeficientů. Na těchto datech pak probíhá učení neuronové sítě nebo skrytých markovských modelů.



Obrázek 4.1: Schéma fonémového rozpoznávače. Převzato z [16]

Naučené systémy mohou být již použity k rozpoznávání nad daty, která také byla upravena Mel-Frekvenční Kepstrální analýzou. Výsledkem je takzvaná matice aposteriorních pravděpodobností, kde každý foném představuje jeden řádek, sloupce pak jednotlivé časové okamžiky. Výsledný rozeznaný foném v daném okamžiku je ten s největší pravděpodobností.

Kapitola 5

MGB Challenge

Tato kapitola je shrnutí [8] MGB Challenge. Dále v části o skórování jsou vysvětleny použité statistické pojmy[6].

MGB Challenge (Multi-Genre Broadcast Challenge) je projekt zabývající se problémy z oblasti zpracování přirozeného jazyka, jedná se o oficiální výzvu (challenge) `2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)`. V roce 2015 se jednalo o 4 úkoly:

- Task 1: Rozpoznávání řeči (Speech-to-text transcription).
- Task 2: Zarovnání (Alignment) audia k titulkům.
- Task 3: Dlouhodobé rozpoznávání řeči (Longitudinal speech-to-text transcription) na úsecích nahrávek ze stejných pořadů.
- Task 4: Dlouhodobá diarizace mluvčího (Longitudinal speaker diarization) – identifikace mluvčího na více nahrávkách.

Data poskytnutá k jednotlivým úkolům byla poskytnuta BBC (British Broadcasting Corporation) z jejích kanálů BBC1, BBC2, BBC3 a BBC4 z pořadů vysílaných v rozmezí 1. dubna 2008 až 19. května 2008. Audio data jsou více žánrová, snažící se pokrýt všechen rozsah televizního vysílání.

V roce 2016 výzva obsahovala arabské nahrávky z televizní stanice Aljazeera. Tento rok (2017) výzva zahrnuje více žánrové nahrávky z youtube a z televizní stanice Aljazeera.

Tato práce se zabývá výhradně problémem zarovnávání textu a audia (task 2).

5.1 Task 2 – zarovnání

Cílem tohoto úkolu je zarovnat jednotlivá slova z přepisu k jejich vyslovení v audiu.

5.1.1 Data

K tomuto úkolu byly k dispozici titulky, s kterými byl pořad původně vysílán, zarovnání vygenerovaná pomocí automatického rozpoznávače řeči a ručně vyrobené zarovnání na jednotlivá slova. Televizní titulky se často neshodují s reálně vyslovenými slovy. Důvodem může například být zvýšení kvality porozumění, kdy dojde k odstranění některých slov, kde je promluva příliš rychlá. V titulkách se tedy mohou vyskytovat slova, která nebyla řečena a naopak nemusí obsahovat všechna slova, která byla řečena.

5.1.2 Skórování

Všechny zkoumané prvky spadají do jedné ze čtyř kategorií:

- skutečně negativní (true-negative) – správně určený negativní prvek
- skutečně pozitivní (true-positive) – správně určený pozitivní prvek
- falešně negativní (false-negative) – špatně určený pozitivní prvek
- falešně pozitivní (false-positive) – špatně určený negativní prvek

Přesnost (precision)

– je definována jako poměr prvků určených správě ku všem určeným prvkům

$$precision = \frac{true\ negatives}{true\ negatives + false\ positives}.$$

Výtěžnost (recall)

– na rozdíl od přesnosti je výtěžnost definována jako poměr správně určených výsledků ke všem pozitivních prvků

$$recall = \frac{true\ negatives}{true\ negatives + false\ positives}.$$

F-hodnota (f-measure)

– pak kombinuje přesnost a výtěžnost do jednoho čísla

$$f - measure = 2 * \frac{precision * recall}{precision + recall}.$$

K vyhodnocení byl dodán skript, který počítá přesnost a výtěžnost a z nich počítá f-hodnotu. Slovo je počítáno jako správně zarovnané, jestliže se jeho začátek a konec liší maximálně o 100 ms od referenční hodnoty. V tomto úkolu bylo povoleno použít pouze slova z titulků, ale nemusela být použita všechna. Správná zarovnání se tedy počítala pouze ze slov použitých před zpracováním. Z vyhodnocení byla odstraněna slova, která spadají do intervalů, kde se jednotlivé promluvy překrývají.

Kapitola 6

Použité nástroje

SoX

SoX (Sound eXchange) je víceplatformní program, který dokáže převést různé druhy audio formátů na jiné formáty. Dále dokáže aplikovat různé efekty nad audio soubory a navíc dokáže nahrávat a přehrávat některé audio soubory[1].

Zde byl využit pro převod audio souborů ve formátu `wav` do formátu `raw`, se kterým dokáže pracovat fonémový rozpoznávač phnrec.

```
sox input.wav -r 8000 -b 8 output.raw
```

- input.wav – zdrojový soubor
- -r 8000 – vzorkovací frekvence výstupního souboru
- -b 8 – velikost jednoho vzorku
- output.raw – výstupní soubor

Phnrec

Phnrec je fonémový rozpoznávač vyvinutý na Fakultě Informačních Technologií na Vysokém Učení technickém v Brně (FIT VUT). Využívá klasifikátor na bázi neuronových sítí a Viterbiho algoritmus na dekódování fonémů. Obsahuje systém na anglický, český, maďarský a ruský fonémový přepis[2].

```
phnrec -c PHN_HU_SPDAT_LCRC_N1500 -i input.raw -o output.rec
```

- -c PHN_HU_SPDAT_LCRC_N1500 - použitý systém pro rozpoznávání
- -i input.raw – vstupní soubor
- -o output.rec – výstupní soubor

G2p-alignment.py

Skript g2p-alignment.py byl použit k trénování modelu zarovnání a k zarovnání grafému na foném. Vyvinul ho Mirko Hanneman na FIT VUT.

Trénování:

[1] http://sox.sourceforge.net/

[2] http://speech.fit.vutbr.cz/cs/software/phoneme-recognizer-based-long-temporal-context

```
g2p-alignment.py train-align input.g2p.labels
      output.graphones output.costs output.alignment beam
```

- train_align – trénování a zarovnání
- input.g2p.labels – data nad kterými se trénuje model a která budou zarovnána
- output.graphones output.costs – soubory do kterých bude uložen výsledný model
- output.alignment – vypočítané zarovnání
- beam – maximální relativní vzdálenost mezi grafémem a fonémem

Zarovnání:

```
g2p-alignment.py align input.g2p.labels input.graphones input.costs
      output.alignment beam
```

- align – zarovnání
- input.g2p.labels – data, nad kterými se trénuje model a která budou zarovnána
- input.graphones output.costs – soubory obsahující model, podle kterého se budou zarovnávat vstupní data
- output.alignment – vypočítané zarovnání
- beam – maximální relativní vzdálenost mezi grafémem a fonémem

Alignment2mlf.py

Stejně jako skript g2p-alignment.py i alignment2mlf.py vyvinul Mirko Hanneman. Jeho úkolem je k vypočtenému zarovnání přidat časové značky z fonémového přepisu a výsledek uložit ve formátu `mlf`.

```
alignment2mlf.py false input.rec input.alignment > output.mlf
```

- false – pouze pro testovací účely
- input.rec – fonémový přepis s časovými údaji
- input.alignment – výsledné zarovnání
- output.mlf – výsledný soubor ve formátu mlf

Score-alignment.py

Jedná se o skript který vyvinul Thomas Hain speciálně pro MGB Challenge. Používá se pro skórování zarovnaných dat.

```
score-alignment.py --ignoreintervals input.uem --match_bound bound
   --script reference.stm reference.ctm input.ctm > result.txt
```

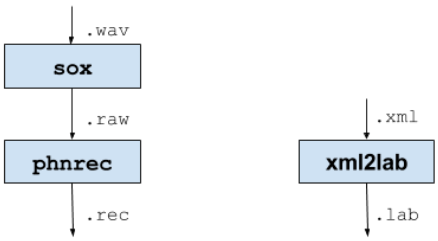
- --ignoreintervals input.uem – časové úseky, které se nepočítají do výsledného skóre
- --match_bound bound – přesnost shody
- --script reference.stm – referenční přepis
- input.ctm – skórované zarovnání
- result.txt – výsledky zarovnání

Kapitola 7

Experimenty

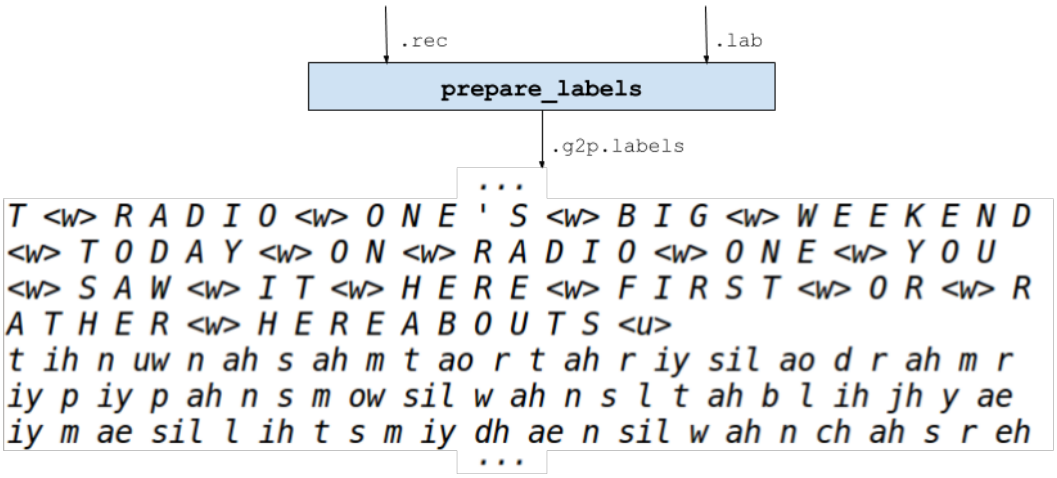
7.1 Příprava dat

Jako první je třeba z dodaných XML souborů získat přepis, který budeme zarovnávat. Nás zajímá ruční přepis (`transcript_human`). Skript `xml2lab.sh` vytvoří `lab` soubor kde je každé slovo na vlastním řádku. Dále je třeba získat fonémový přepis z `wav` nahrávky. Prvně je třeba převést audio soubor z formátu `wav` na formát `raw`, k tomu byl využit linuxový nástroj `sox`. Se souborem ve formátu `raw` už dokáže pracovat fonémový rozpoznávač `phnrec`, který vytvoří fonémový přepis audia. Tento přepis má formát
`<začátek fonému> <konec fonému> <foném> <věrohodnost fonému>`
a je uložen v souboru s koncovkou `.rec`.



Obrázek 7.1: Schéma přípravy dat

Z těchto dvou souborů se vytvoří soubor s koncovkou `.g2p.labels`, který se skládá ze dvou řádků. První obsahuje přepis, kde každý grafém je oddělen mezerou, každé slovo značkou `<w>` a celý přepis značkou `<u>`. Druhý řádek pak obsahuje jednotlivé fonémy oddělené mezerou.



Obrázek 7.2: Schéma přípravy dat

7.2 Porovnání ručního přepisu a automatického přepisu

Jako první experiment bylo zvoleno porovnání ručního přepisu s automatickým přepisem, oba přepisy jsou obsaženy v dodaných xml souborech, jedná se o přepis `transcript_human` a `transcript_lsdecode`. Porovnání bylo provedeno spočítáním nejmenší vzdálenosti (Minimum Edit Distance – MED) mezi těmito přepisy, a z této vzdálenosti vypočítán Word Error Rate (WER) a Word Accuracy (WAcc).

Minimum Edit Distance

Minimum Edit Distance nebo-li MED je číslo udávající nejmenší možný počet operací, které je nutné provést, aby se z porovnaného řetězce stal zdrojový řetězec. Zarovnání dvou řetězců pomocí Minimum Edit Distance se využívá v různých odvětvích např. v strojovém překladu, v rozpoznávání řeči nebo také v biologii, kde se využívá k porovnání dna[3].
operace:

- Vložení (Insertion) – vložení znaku
- Smazání (Deletion) – smazání znaku
- Substituce (Substitution) – náhrada jednoho znaku za jiný

```
K - Á V A     K R Á V A     K Á V A
| X | | |     | | | | |     | | | |
K R Á V A     K R Á V A     L Á V A

smazání       vložení       náhrada
```

Obrázek 7.3: Příklad možných operací.

Každá operace může mít různou hodnotu, odtud můžeme rozlišovat dva druhy Levenshteinovy vzdálenosti, kdy každá operace má hodnotu 1, případně pozměněná Levenshteinova vzdálenost, kdy vložení a smazání má hodnotu 1 a náhrada hodnotu 2.

Minimum Edit Distance se vypočítává pomocí takzvaného Dynamického programování (Dynamic programming). Dynamické programování je název pro skupinu algoritmů, které ke svému výpočtu využívají ukládání výsledků podproblémů a kombinací těchto výsledků vypočítají původní problém.

- M – velikost zdrojového řetězce
- N – velikost cílového řetězce
- D(i,j), kde 0 < i < n a 0 < j < n

Inicializace:
D(i,0) = i
D(0,j) = j

Rekurzivní vztah:
For each i = 1..M
For each j = 1..N

$$D(i,j) = min \begin{cases} D(i-1,j)+1 \\ D(i,j-i)+1 \\ D(i-1,j-1)+1 \quad \text{if X(i) != Y(j) else 0} \end{cases}$$

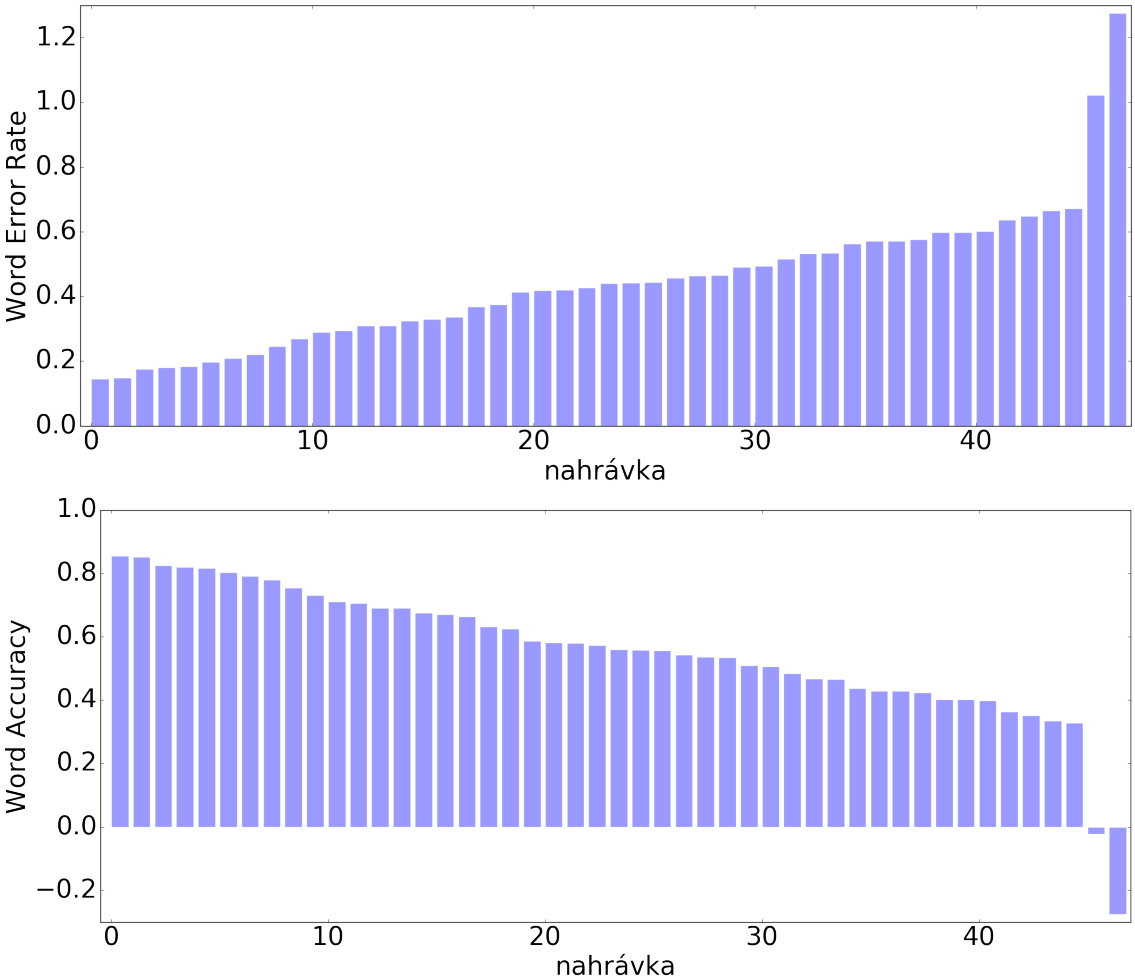
Výsledek:
MED = D(N,M)

Word Error Rate [7]

$$\text{WER} = \frac{S+D+I}{S+D+C} = \frac{S+D+I}{N} = \frac{\text{MED}}{N},$$

kde

- `S` je počet substitucí
- `D` je počet smazání



Obrázek 7.4: Word Accuracy a word error rate vypočítaný mezi automatickým přepisem a ručně vytvořeným přepisem.

- `I` je počet vložení
- `C` je počet správných slov
- `N` je počet slov

Word Accuracy je pak doplněk do 1[7]:

$$\mathrm{WAcc} = 1 - \mathrm{WER} = \frac{N - S - D - I}{N}$$

Vyhodnocení

Jak je vidět z obrázku 7.4, tak dvě nahrávky dokonce mají WER větší než 1 (respektive WAcc menší než 0), jedná se o sportovní nahrávky, jejichž ruční přepis a automatický přepis mají značně rozdílnou velikost. Ostatní nahrávky mají WER rovnoměrně rozmístěny mezi 16 % a 60 %.

7.3 Zarovnání

Dalším krokem bylo zarovnat textový přepis s fonémovým přepisem. K tomu byl využit skript `g2p_alignment.py`[1], který se při základním spuštění snaží pomocí forward-backward algoritmu iterativně najít optimální zarovnání grafému na foném, hledání je omezeno parametrem `maximum_beam`, který udává maximální relativní rozdíl mezi grafémem a fonémem. Kromě nalezeného zarovnání vrací i pravděpodobnosti, že se konkrétní grafém přepíše na určitý foném. Výsledné zarovnání je uloženo v souboru s koncovkou `alignment`. Kromě vypočítaného zarovnání se také ukládá získaný model. Model se skládá ze souborů s koncovkou `grapheme a costs`.

Získané zarovnání společně s původním fonémovým přepisem jsou vstupem pro skript `alignment2mlf.py`, který podle časových údajů z fonémového přepisu přiřadí jednotlivým zarovnáním jejich pravděpodobné časové značky. Výsledek uloží ve formátu `.mlf`, který z nich poskládá již hotová zarovnaná slova. Časové značky jednotlivých slov se ještě upraví, je-li na začátku nebo na konci slova přiřazen foném nemluveného zvuku a výsledek je uložen v souboru `timings.lab` ve formátu:

```
<slovo_přepisu> <čas_počátku_slova> <čas_konce_slova>
```

Celý proces zarovnání je znázorněn na obrázku 7.5.

Vyhodnocení

Z výsledného zarovnání je třeba vyrobit soubor typu `ctm`, kde každý řádek je tvaru:

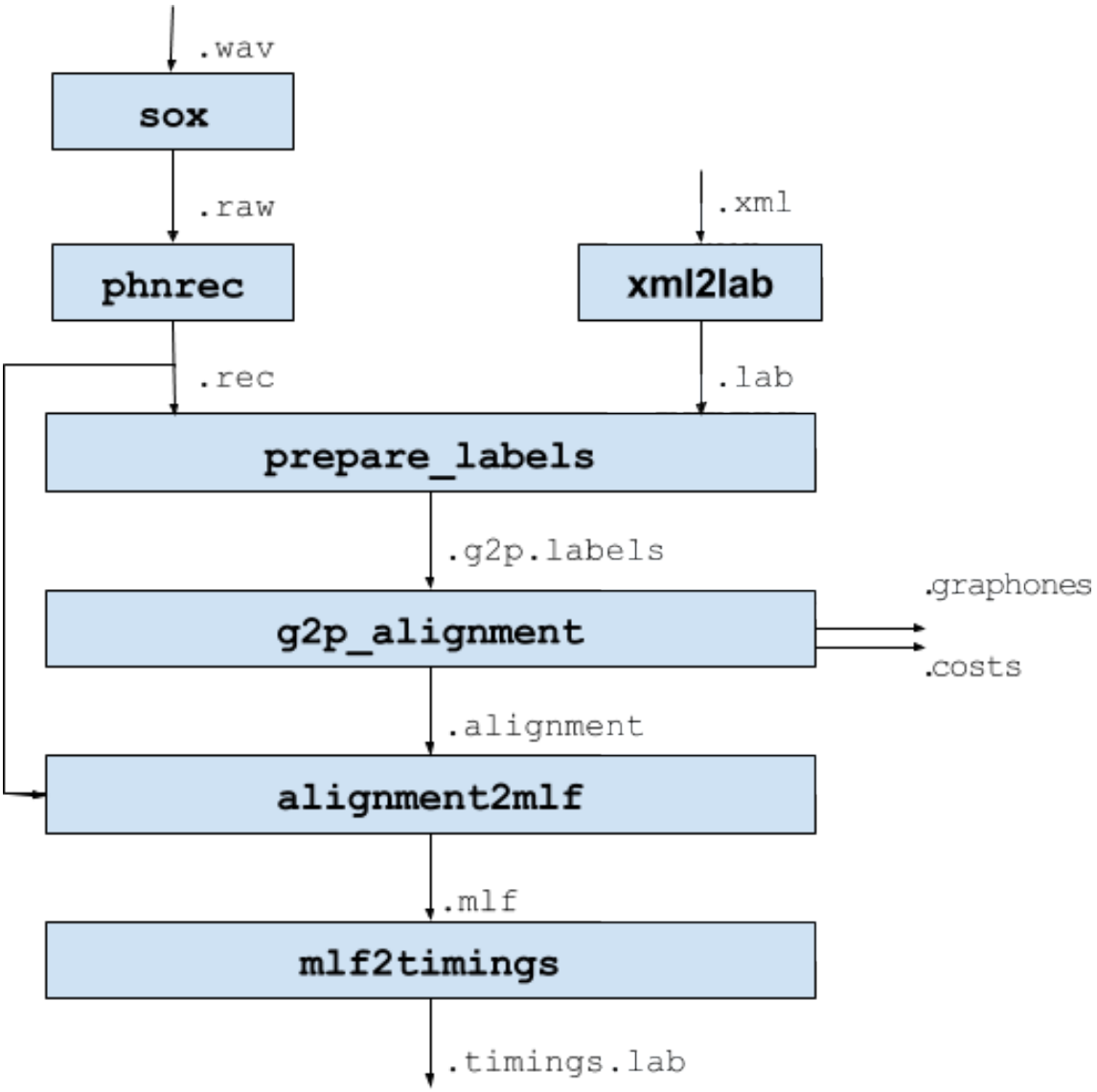
```
<název-nahrávky> 0 <čas začátku slova> <trvání> <zarovnané slovo>
```

Skórování probíhá pomocí skriptu `score-alignment.py`, který pomocí referenčního zarovnání spočítá přesnost, výtěžnost a f-hodnotu jednotlivých nahrávek a celkově za všechny nahrávky dohromady. Slovo se počítá jako správně zarovnané, pokud se začátek a konec neliší o více jak 100 milisekund od referenčních hodnot, tuto hodnotu je možno změnit pomocí parametru `--match_bound [hodnota]`.

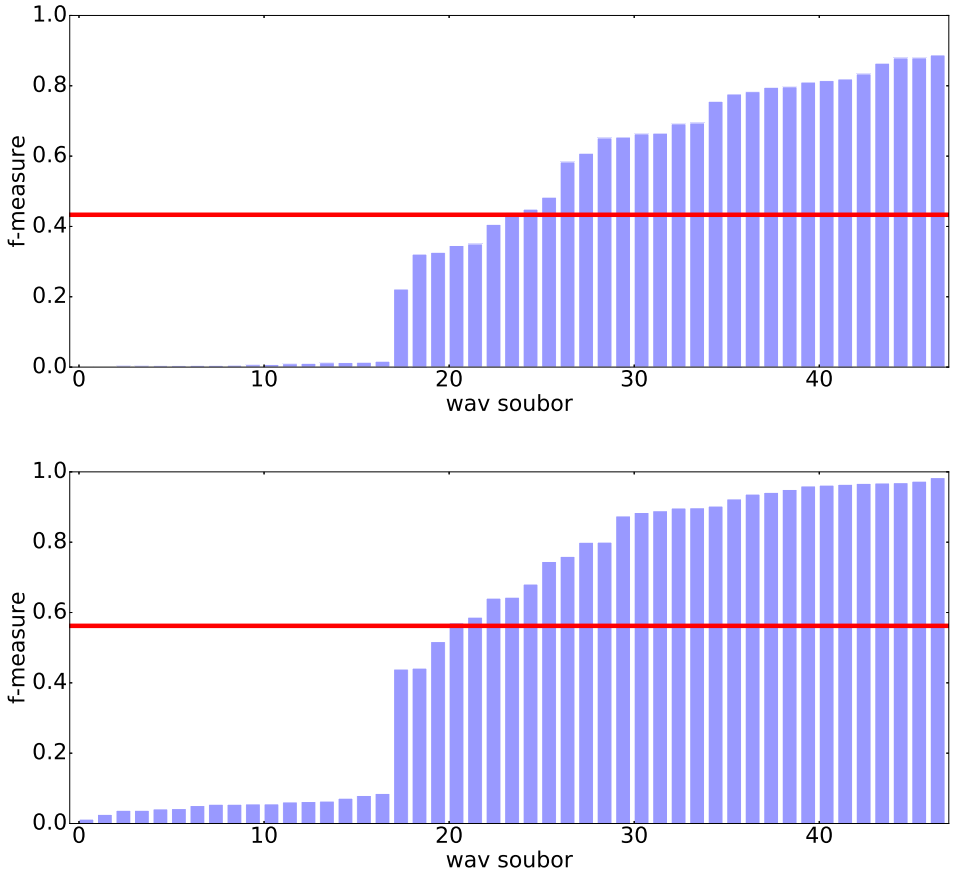
Výsledky

Vyhodnocení se provedlo s přesností na 100 ms a na jednu sekundu. Jak je vidět na obrázku 7.6, osm nejlepších nahrávek mělo úspěšnost větší než 80 % pro přesnost 100 ms a pro přesnost na 1s čtrnáct nahrávek nad 90 %, kde nejlepší nahrávka dosahovala téměř 97 %. Mezi nejlépe zarovnané nahrávky se řadili zpravodajské pořady a dokumenty. Naopak 17 z 47 nahrávek nedosáhlo ani 9 % úspěšnosti při přesnosti na 1 s. Celková úspěšnost za všechny nahrávky byla 43 % pro 100 ms, respektive 56 % pro 1 s.

[1] Copyright 2013-2015 Mirko Hannemann, BUT, mirko.hannemann@gmail.com



Obrázek 7.5: Celkový diagram zarovnání. Model (graphones, costs) se vypočítává pouze nad dodanými daty.

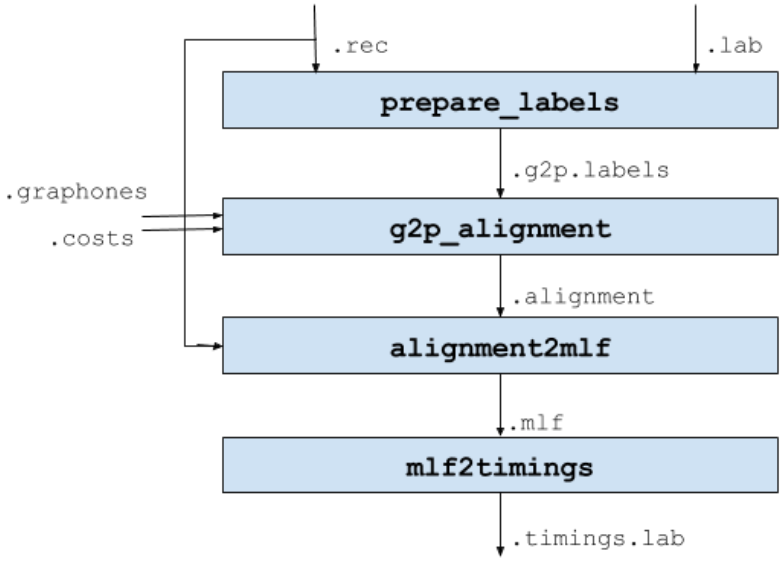


Obrázek 7.6: Horní graf ukazuje výsledky zarovnání, s přesností shody 100 ms, dolní s přesností na 1 s. Horizontální osa označuje jednotlivé soubory, vertikální f-hodnotu. Červená čára označuje celkovou f-hodnotu, pro všechny soubory.

7.4 Zarovnání podle modelu `n-té` nahrávky

Při tomto experimentu bylo zkoumáno, jak se mění zarovnání nebude-li se počítat model z právě zarovnávané nahrávky, ale využije se cizí model, jak znázorňuje obrázek 7.7.

Bylo využito získaných modelů z výše popsaného experimentu a podle těchto modelů byla zarovnána celá datová sada. Ne každý model však obsahuje všechny grafémy a fonémy, z tohoto důvodu byli grafémy (resp. fonémy) s velmi nízkým výskytem ze zarovnávaných nahrávek odstraněny. Bylo-li třeba zasáhnout více do zarovnávaných nahrávek, byl daný model z experimentu vyjmut.

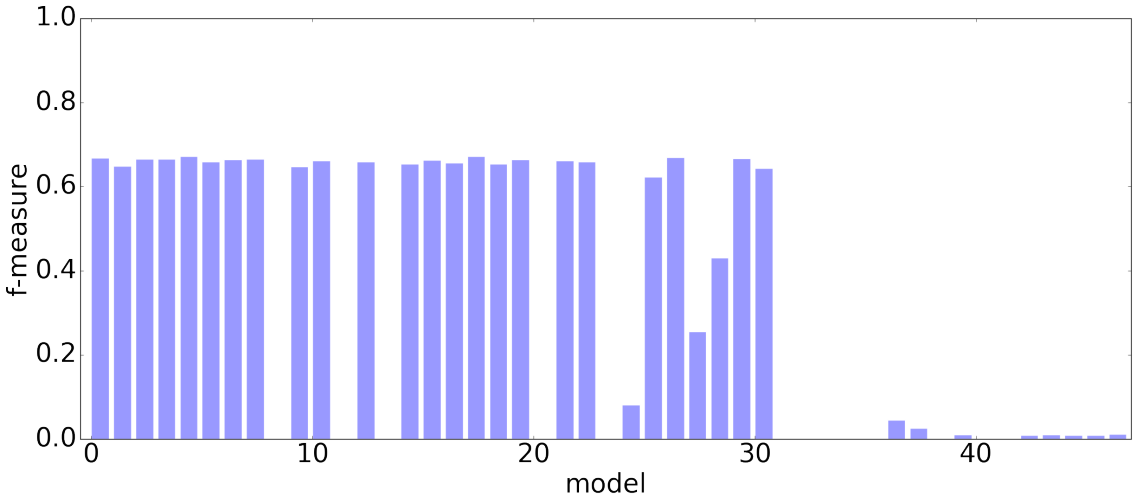


Obrázek 7.7: Schéma zarovnání při využití cizího modelu.

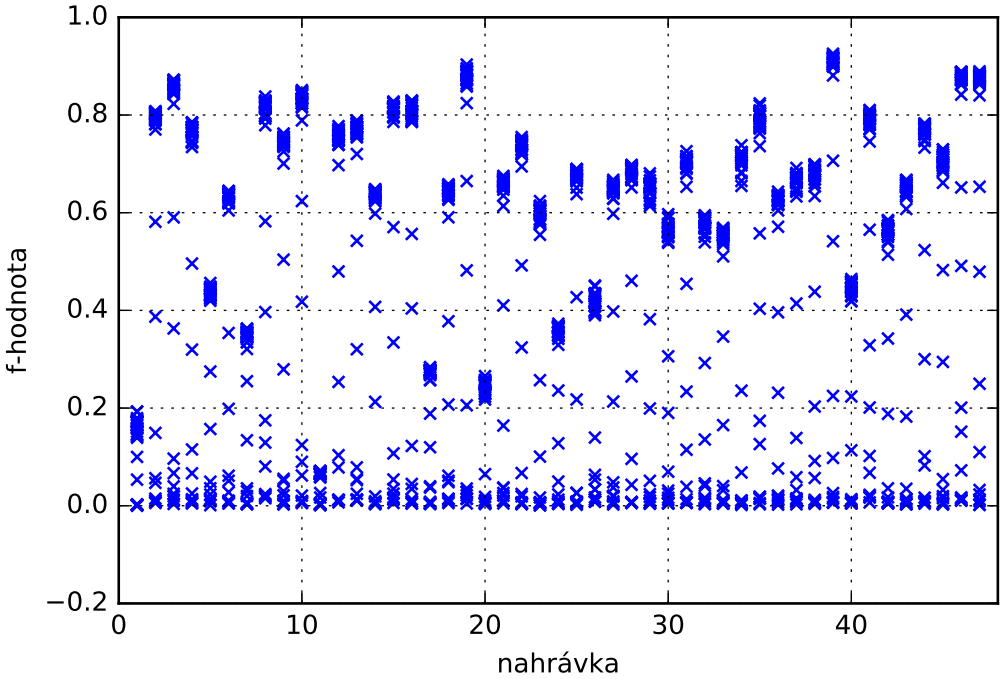
Výsledky

Jak je vidět z obrázku 7.8, tak zarovnání podle většiny modelů, které měli f-hodnotu větší jak 30 % získalo celkovou f-hodnotu mezi 60–65 %. Naopak zarovnání podle modelu s nižší f-hodnotou nedosáhlo téměř žádné přesnosti.

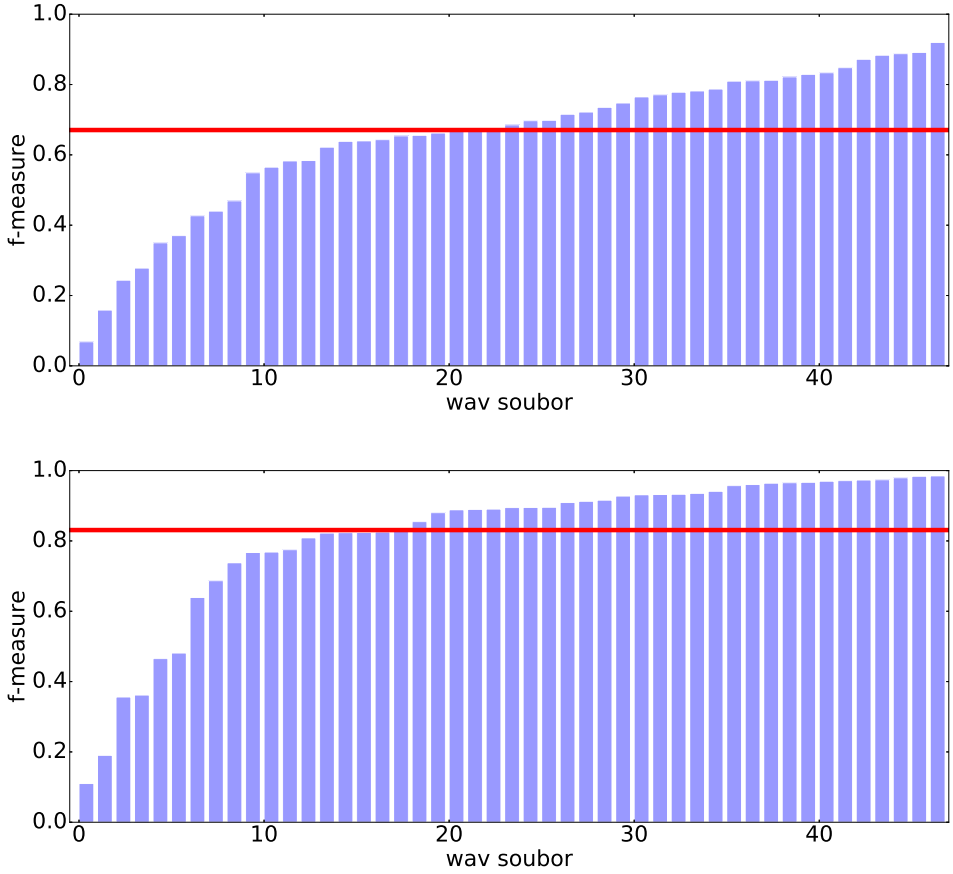
Neexistuje závislost mezi kvalitou zarovnání podle modelu nahrávky, která získala f-hodnotu větší jak 80 % a modelu nahrávky s f-hodnotou okolo 50 %. Zarovnání podle modelu z nahrávky s f-hodnotou mezi 30–50 % je kvalita zarovnání nepředvídatelná.



Obrázek 7.8: Celková f-hodnota pro celou datovou sadu, při zarovnání podle modelu n-té nahrávky. X-ová osa udává pořadí modelu získaného ze zarovnání v 7.3.



Obrázek 7.9: F-hodnota jednotlivých nahrávek



Obrázek 7.10: Horní graf ukazuje výsledky zarovnání, podle modelu tří nejlepších nahrávek z experimentu 7.3 s přesností shody 100 ms, dolní s přeností na 1 s. Horizontální osa označuje jednotlivé soubory, vertikální f-hodnotu. Červená čára označuje celkovou f-hodnotu pro všechny soubory.

7.5 Spojení nahrávek

Dále bylo zkoumáno, zda se kvalita zarovnání zlepší, bude-li se zarovnání provádět podle modelu získaného z nejlépe zarovnaných nahrávek v experimentu 7.3. V tomto experimentu byly spojeny tři nejlepší nahrávky a natrénovaný model byl použit pro zarovnání celé datové sady.

Výsledky

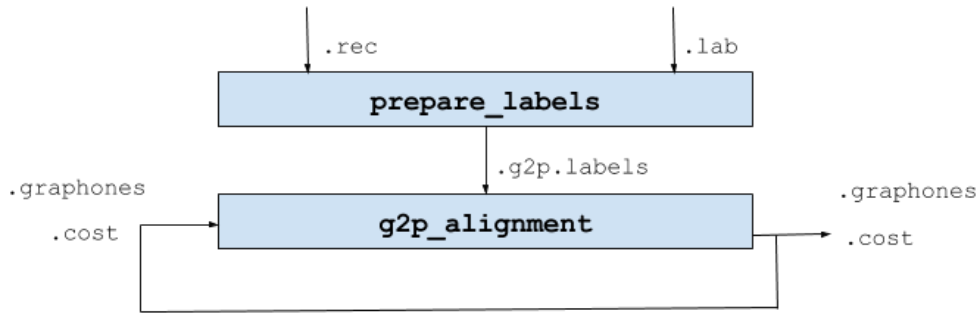
Vyhodnocení se zase provedlo s přesností na 100 ms a na 1 s, viz obrázek 7.10. Při přesností na 100 ms dosáhlo více ja 80 % 13 nahrávek a pouze 6 nahrávek mělo méně jak 40 %. Pří přesnosti na 1 s mělo 31 nahrávek více jak 80 % z toho 10 nahrávek dosáhlo více jak 95 % a 8 nahrávek nedosáhlo ani 15 %. Celková úspěšnost za všechny nahrávky byla 65 % pro 100 ms, respektive 83 % pro 1 s.

7.6 Zarovnání podle modelu `n-nejlepších` nahrávek

Při tomto experimentu na rozdíl od předchozího bylo využito k zarovnání modelu natrénovaného na `n-nejlepších` nahrávkách. Model byl trénován podle obrázku 7.11, kdy iterativně byl z modelu na `n-nejlepších` nahrávkách natrénován model `(n+1)-nejlepších` nahrávkách.

Stejně jako u předchozího experimentu 7.4 byly nahrávky, jejichž model nebyl schopný zarovnat celou testovací sadu (ani po menších úpravách), z experimentu vyjmuty.

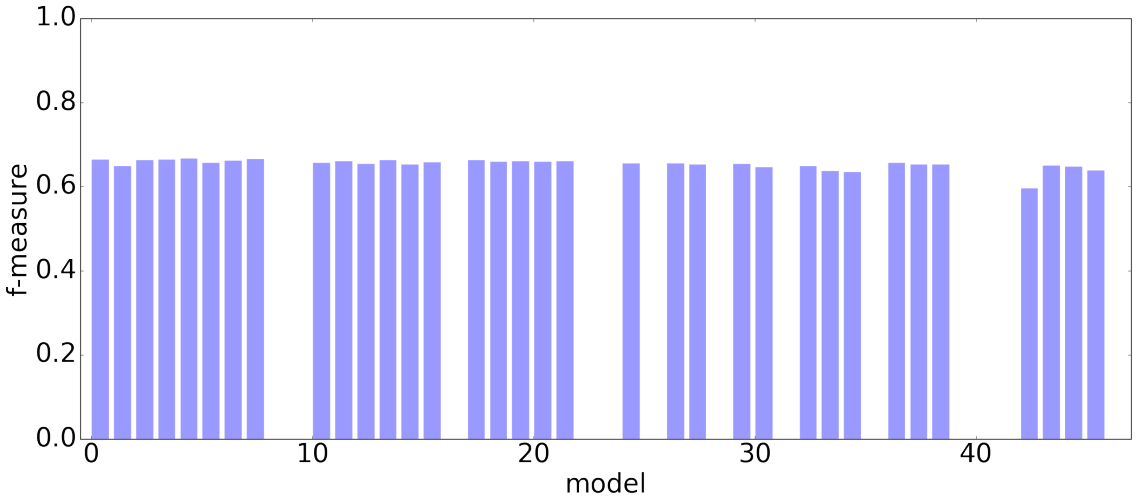
$model(n) = \text{g2p-alignment}(\text{n-tá nahrávka}, model(n-1))$



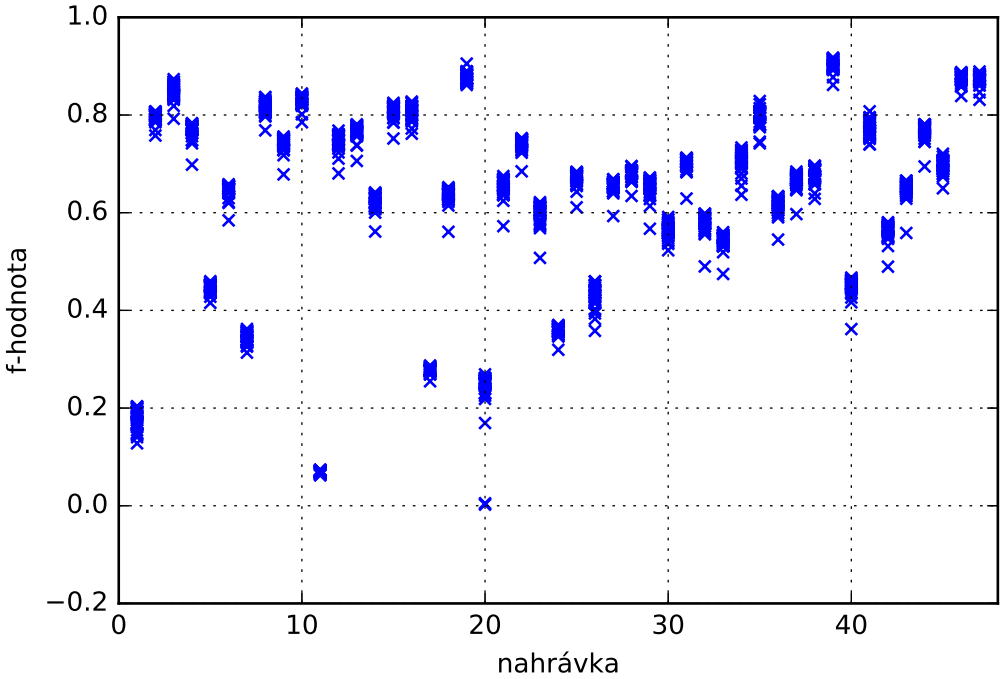
Obrázek 7.11: Schéma získání modelu na více nahrávkách

Výsledky

Výsledky jsou zobrazeny na obrázku 7.17. Na rozdíl od experimentu 7.4, kdy zarovnání podle modelů nahrávek s nízkou f hodnotou z 7.3 nedosáhlo téměř žádné úspěšnosti, zde neměly tyto nahrávky velký vliv na kvalitu zarovnání. Zarovnání podle všech modelů dosáhlo f-hodnoty mezi 60 % a 67 %.



Obrázek 7.12: Celková f-hodnota pro celou datovou sadu, při zarovnání podle modelu n-nejlepších nahrávek



Obrázek 7.13: F-hodnota jednotlivých nahrávek

7.7 Kombinace dobrých a špatných nahrávek

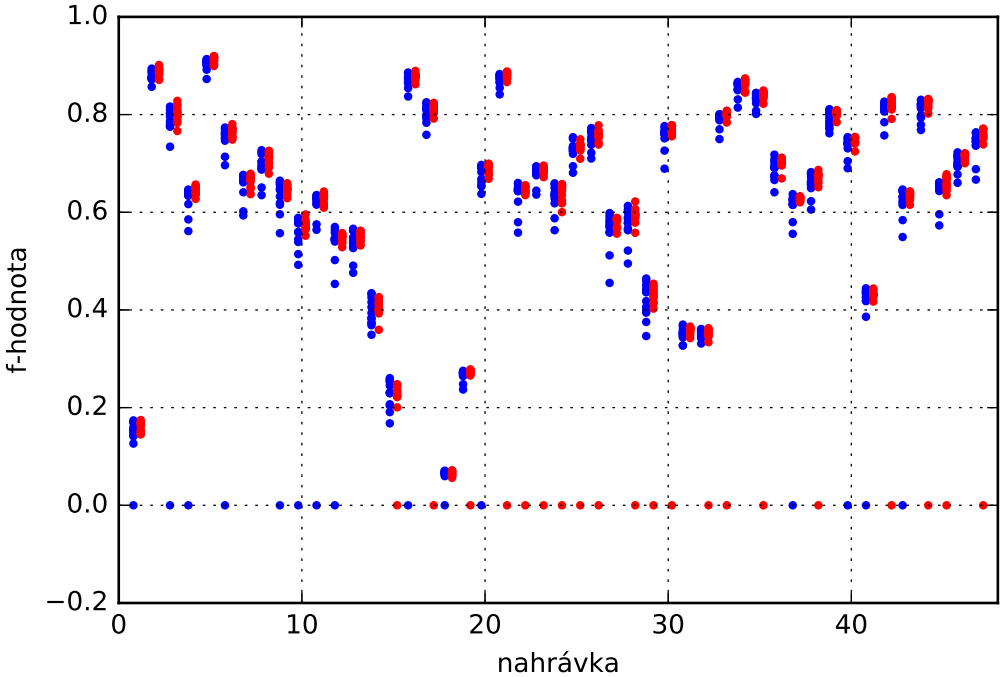
Tento experiment měl za cíl zjistit, jak se mění kvalita zarovnání, je-li model natrénován na kombinaci dobrých a špatných nahrávek. Bylo vybráno osm nahrávek, z toho čtyři dobré

a čtyři špatné. Modely byli natrénovány na sekvencí všech dvojic, kde jedna nahrávka je špatná a druhá dobrá. Model byl trénován iterativně podle obrázku 7.11 v experimentu 7.6.

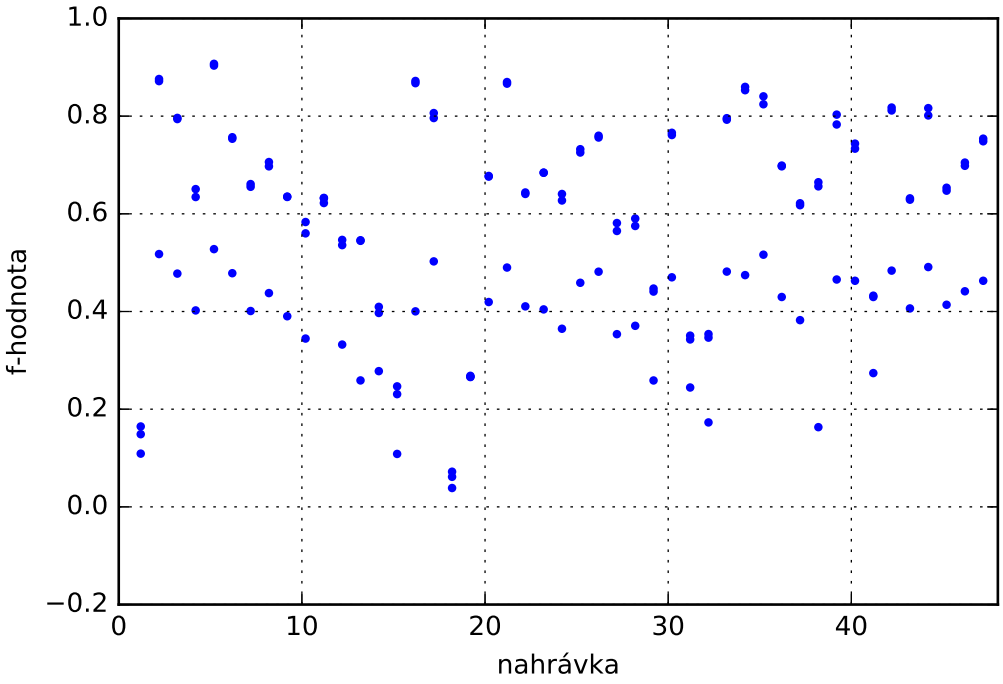
Výsledky

Při kombinaci dvou nahrávek se celková hodnota zarovnání pohybovala stejně jako u předchozího experimentu 7.6 okolo 65 %. Jak ukazuje obrázek 7.14, tak zarovnání, kdy se prvně model natrénoval na dobré a potom na špatné nahrávce (modrá), v některých případech má o několik procent menší úspěšnost. V obou případech však existovaly nahrávky, které se nepovedly s daným modelem zarovnat

Druhý obrázek 7.15 ukazuje zarovnání podle čtyřech špatných a následně na jedné dobré nahrávce. V tomto případě dva modely dosáhly podobných výsledků okolo 65 % a jeden model 40 % a pro jednu nahrávku se nepodařilo natrénovat model.



Obrázek 7.14: Výsledky zarovnání jednotlivých nahrávek. Červená označuje zarovnání, kdy prvně byl model natrénován na dobré nahrávce a poté na špatné, modrá obráceně.



Obrázek 7.15: Výsledky zarovnání jednotlivých nahrávek. Model byl postupně natrénován na čtyřech špatných nahrávkách a poté na jedné dobré.

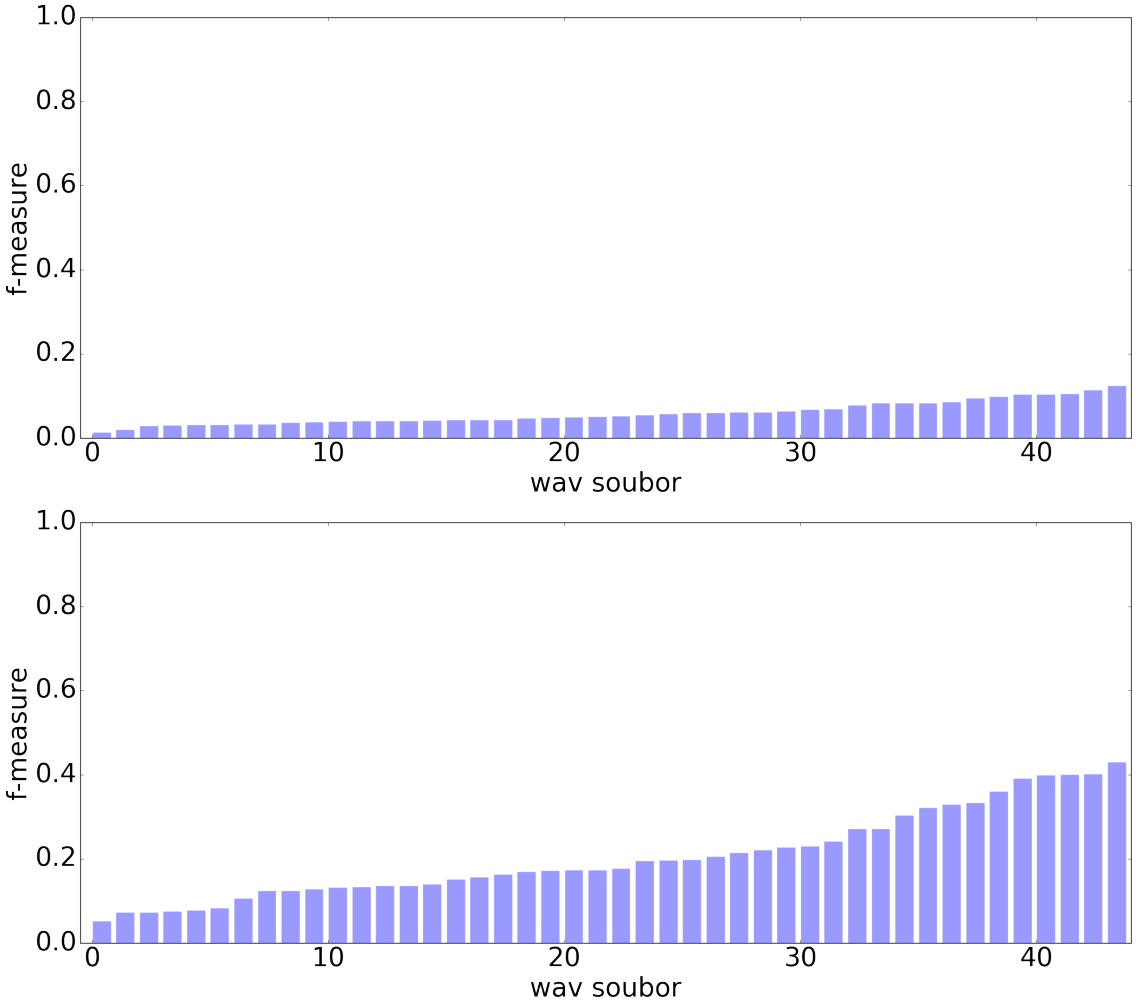
7.8 Zarovnání podle Maďarského fonémového přepisu

Při tomto experimentu nebyly použity původní anglické fonémové přepisy, ale byly vygenerované nové fonémové přepisy, pomocí fonémového rozpoznávače phnrec, vyvinutého na Fakultě informačních technologií v Brně[2]. Tímto rozpoznávačem byly získány maďarské fonémové přepisy. S těmito přepisy pak bylo provedeno zarovnání podle obrázku 7.5.

Výsledky

Jak ukazuje obrázek 7.16, tento experiment nedosáhl dobrých výsledků. Při přesnosti 100ms dosáhly nejlepší nahrávky pouze 2 %. Při zvýšení přesnosti na 1 s a 10 s se nejlepší nahrávka pohybovala okolo 13 % respektive 44 %. Nutno podotknout, že určení zarovnání s přesností na 10 sekund je většině případů nedostatečné. Tři nahrávky se dokonce ani nepodařilo zarovnat, z důvodu numerické chyby při zarovnání, jednalo se o sportovní přenosy, které obsahují velké množství okolního hluku.

[2] http://www.fit.vutbr.cz/research/groups/speech/sw/phnrec.html



Obrázek 7.16: Horní graf ukazuje výsledky experimentu 7.8, s přesností shody 1 s, dolní s přeností na 10 s. Horizontální osa označuje jednotlivé soubory, vertikální osa f-hodnotu.

7.9 Zarovnání podle modelu `n-té` nahrávky s maďarským fonémovým přepisem

Tento experiment měl za cíl zjistit jak se chová zarovnání maďarského fonémového přepisu podle modelu natrénovaného na jiné nahrávce, podle obr 7.7. Pro tento účel byly využity modely z experimentu 7.8.

Výsledky

Při tomto experimentu zarovnání při využití většiny modelů selhalo. Ve většině případů byl použitý model schopný zarovnat pouze tu nahrávku, na které byl natrénován. Pokud se podařilo nahrávku zarovnat, kvalita zarovnání byla na podobné úrovni jako při experimentu 7.8.

7.10 Průnik zarovnaných nahrávek

Tento experiment měl za cíl najít v zarovnání správně zarovnaná slova, tedy zvýšit přesnost (precision). Pro tento účel byl napsán skript `prunik-ctm.py`, který bere na vstupu dvě nahrávky a na výstup vypíše pouze ta slova, která se shodují. Dvě zarovnaná slova se berou jako shodná pouze pokud se čas začátku a konce liší maximálně o zadaný parametr.

Bylo zkoumáno, jak se mění přesnost (precision) zarovnání, pokud se bude iterativně aplikovat skript `prunik-ctm` na zarovnání z experimentů 7.4 a 7.6.

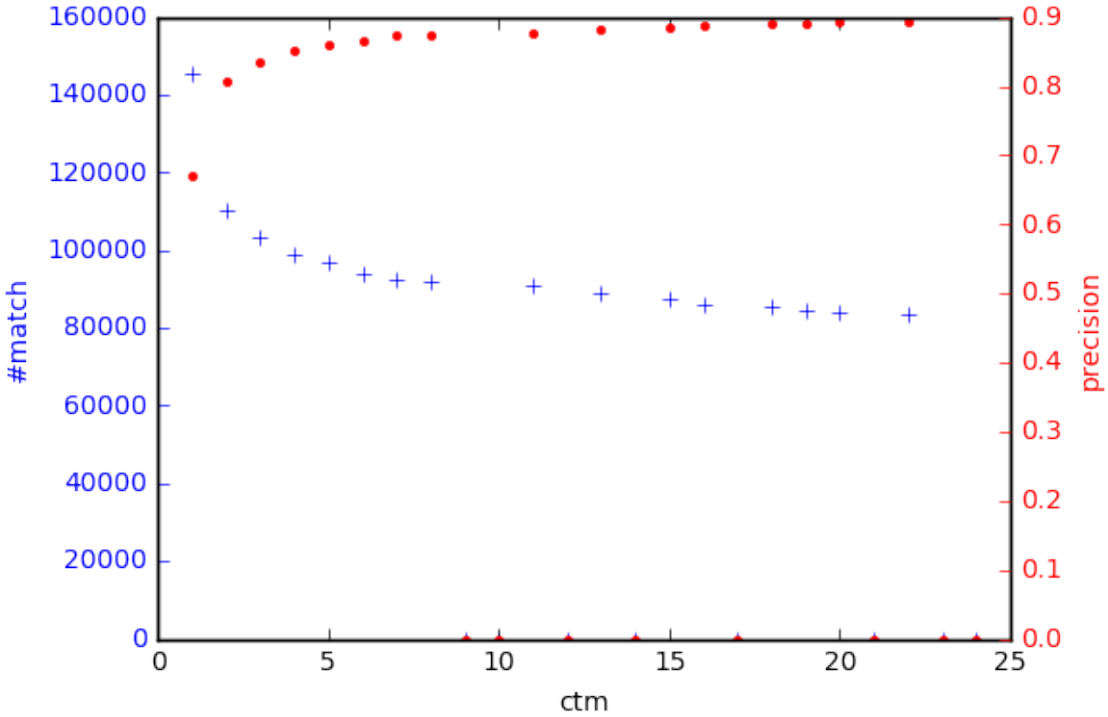
Dále byl zkoumán průnik zarovnání z 7.6, kdy pro každé dvě zarovnání byl proveden průnik.

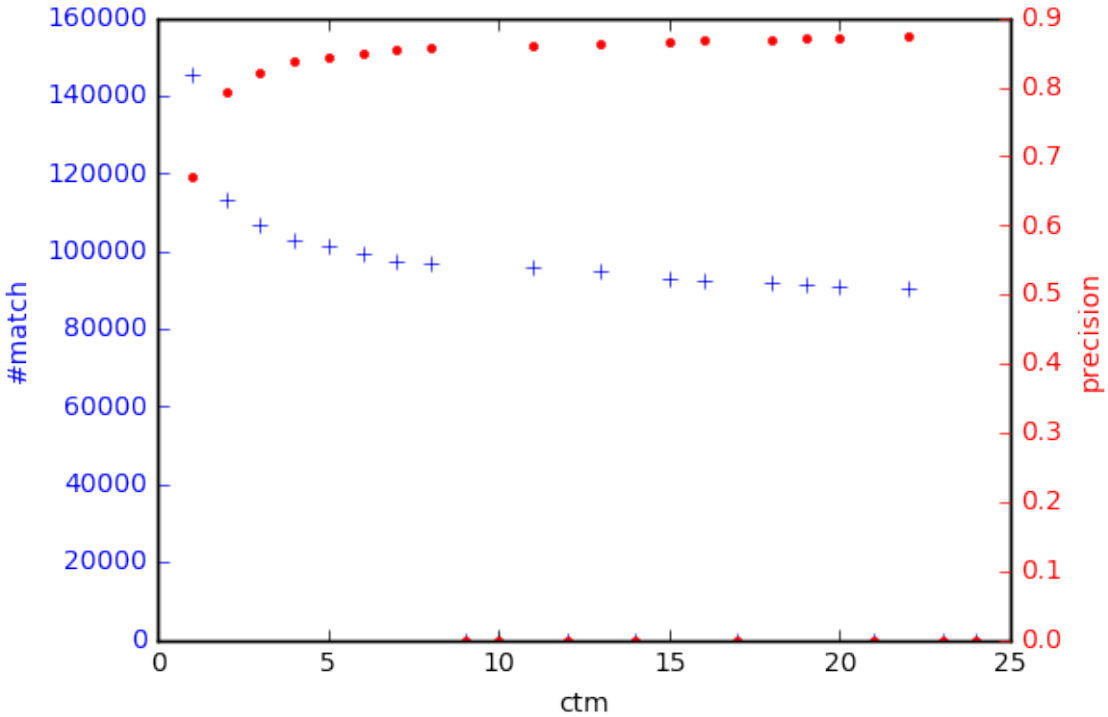
Výsledky

Obrázek 7.17 ukazuje, jak se mění přesnost a počet slov, které jsou brány jako shodné při postupném průniku zarovnání. Při průniku zarovnání z experimentu 7.4 se přesnost zvýšila téměř až k 90 % a počet slov se zmenšil o 45 %, u experimentu 7.6 87 % a 39 %.

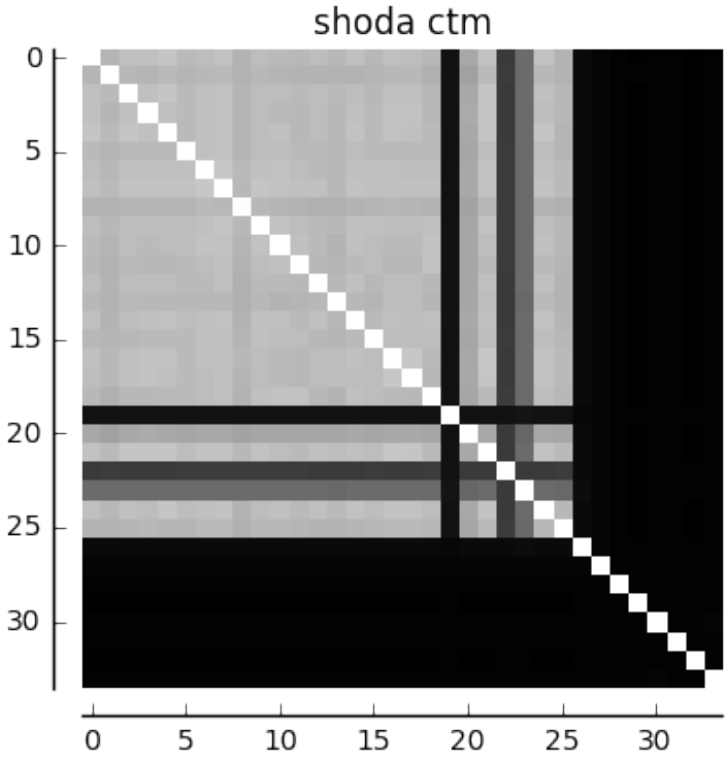
Dále bylo zjištěno, že jednotlivé zarovnání se shodují převážně u slov, která jsou zarovnána správně, tedy zarovnání, která jsou, špatná nemají vysokou shodu, viz černá oblast na obrázku 7.18.

Průnikem dvou kvalitních zarovnání bylo dosaženo nejlepších výsledků 7.19, v mnoha případech výsledná f-hodnota přesáhla 70 %, přesnost se pohybovala okolo 78–82 %.

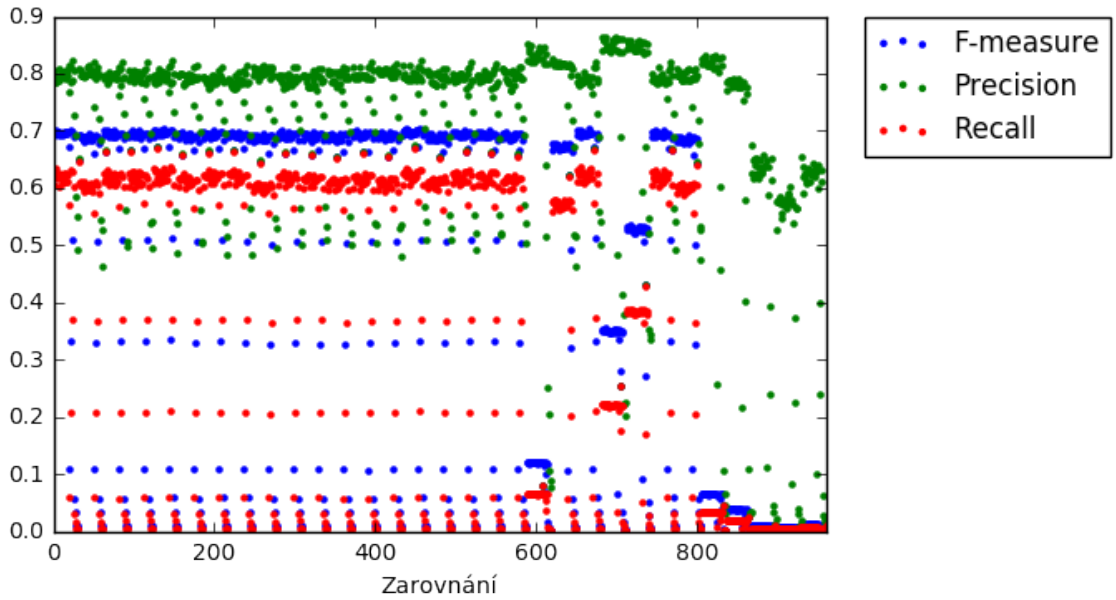




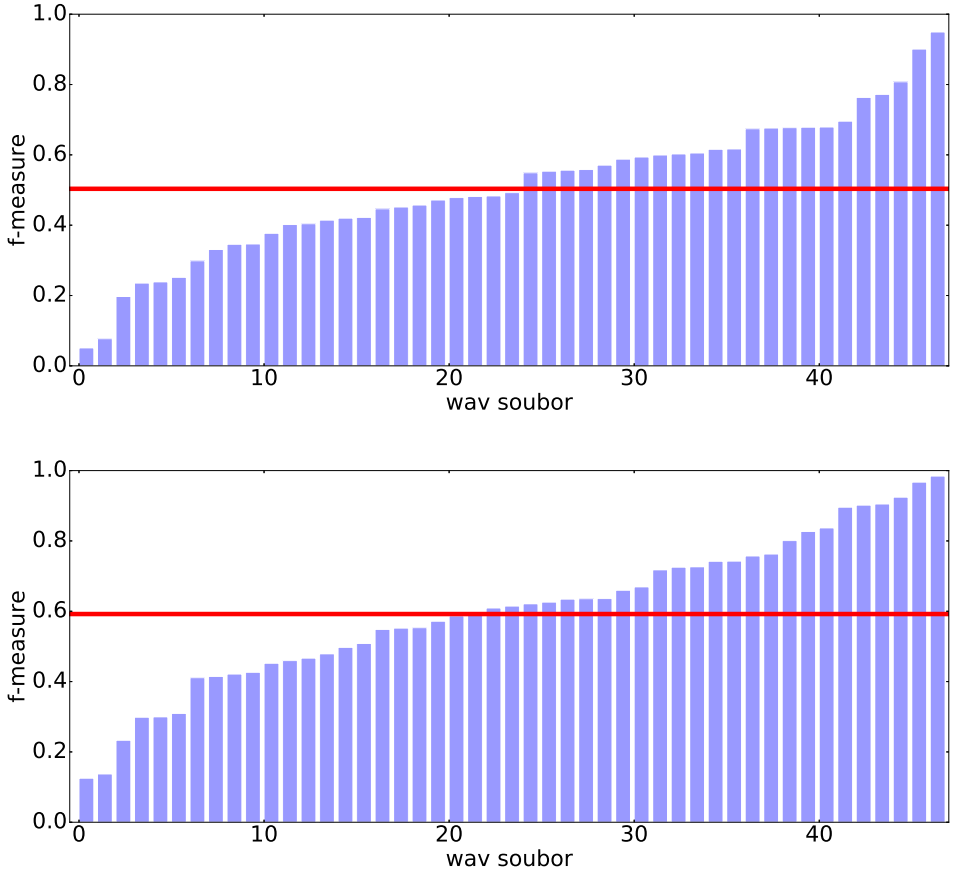
Obrázek 7.17: Počet shod – modře a přesnost(precision) – červeně, při průniku jednotlivých zarovnání. Horní graf průnik zarovnání 7.8, dolní graf průnik zarovnání 7.17. Byly použity pouze zarovnání s f-measure větší jak 60 %.



Obrázek 7.18: Shoda jednotlivých zarovnání z experimentu 7.8. Přesnost shody 100 ms. Bílá – 100 % shody, černá – 0 %.



Obrázek 7.19: Vyhodnocení jednotlivých průniků z 7.18 (řazeno po řádcích).



Obrázek 7.20: Výsledky zarovnání pomocí přepisu získaného automatickým rozpoznávačem řeči. Vrchní graf ukazuje zarovnání s přesností shody na 100ms, dolní graf s přesností shody na 1000ms. Horizontální osa označuje jednotlivé soubory, vertikální f-hodnotu. Červená přímka označuje celkovou f-hodnotu pro všechny soubory.

7.11 Zarovnání grafému na grafém

Ve všech předchozích experimentech byla nahrávka zarovnána tak, že se vytvořil fonémový přepis a ten byl zarovnán s přepisem nahrávky. Při tomto experimentu se místo fonémového přepisu využívá přepis získaný pomocí automatického rozpoznávače řeči, můžeme tedy nazvat tento postup jako zarovnání grafému na grafém (G2G – Graphem-to-Graphem). Časové značky k jednotlivým grafémům byly vypočteny rovnoměrně z časových údajů z automatického rozpoznávače.

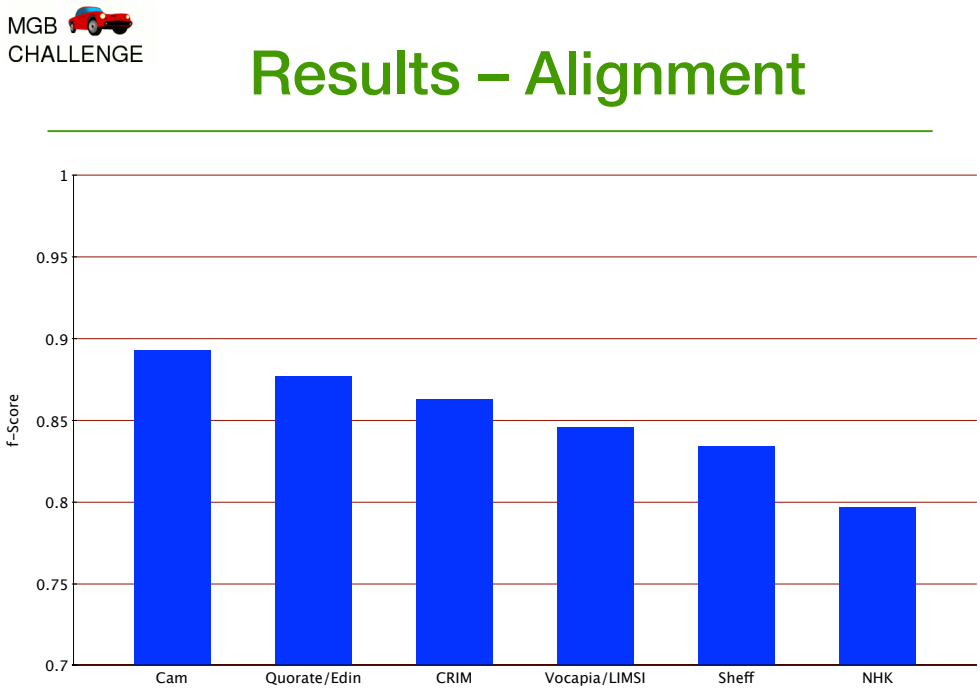
Výsledky

Jak ukazuje obrázek 7.20, tak bylo dosaženo 50 % úspěšnosti u přesnosti zarovnání na 100 ms, respektive 60 % u přesnosti na 1 s. F-hodnota jednotlivých nahrávek je v téměř celém rozmezí 0–1 a do značné míry koreluje s word accuracy 7.4.

Kapitola 8

Závěr

Cílem této práce bylo seznámit se s metodami zarovnání textu a audia a s pracemi, které se daným problémem zabývaly. Vyzkoušet jednotlivé postupy a případně uvažovat i metody nezávislé na jazyku. Veškeré experimenty byly provedeny na datech z MGB Challenge 2015. Nejlepší dosažené zarovnání v této práci dosáhlo f-hodnoty přes 0,7. Jak je vidět na obrázku 8.1, tak výsledky systémů přihlášených do MGB Challenge dosáhla f-hodnoty mezi 0,8–0,9. Mezi nejlépe zarovnané nahrávky patřily zpravodajské a dokumentární pořady. Nejhorších výsledků naopak dosáhly sportovní pořady. Zarovnání Bylo provedeno pomocí nástroje G2P-alignment. Bylo vyzkoušeno zarovnání pomocí přepisu automatického rozpoznávače řeči, anglického fonémového přepisu a maďarského fonémového přepisu získaného pomocí fonémového rozpoznávače phnrec.



Obrázek 8.1: Výsledky úkolu 2 – MGB Challenge 2015. Převzato z [2].

Byl spočítán word error rate (WER) mezi automatickým přepisem a ručně vytvořeným přepisem a spočítáno zarovnání mezi těmito přepisy, kdy výsledky WER a zarovnání do značné míry mezi sebou korelují. Dále bylo provedeno zarovnání mezi anglickým fonémovým přepisem a ručním přepisem, kdy každá nahrávka byla zarovnána na modelu spočítaném na dané nahrávce. Zarovnání dosáhlo celkového výsledku 0,43 f-hodnoty, kde více jak jedna třetina nahrávek nedosáhla téměř žádné úspěšnosti.

V dalších experimentech bylo zkoumáno jak se změní kvalita zarovnání pomocí jiných modelů. Jako první byl natrénován model na spojených třech nejlepších nahrávkách z předchozího experimentu. Při tomto experimentu se celková úspěšnost zarovnání zvedla na 0,65 f-hodnoty, kde až na dvě nahrávky všechny dosáhly alespoň 0,2 f-hodnoty. Dále bylo zkoušeno jak dopadne zarovnání celé sady podle modelů jednotlivých nahrávek. Zde by se výsledná zarovnání mohla řadit do tří skupin, zarovnání podle modelu nahrávky která dosáhla získala více jak 0,50, 0,2–0,5 a méně jak 0,2 f-hodnoty. V první skupině dosáhla všechna zarovnání podobné úspěšnosti jako při spojení nejlepších nahrávek, ve druhé byl výsledek nepředvídatelný a ve třetí výsledné zarovnání nedosáhlo téměř žádné úspěšnosti. Při dalším experimentu byl postupně natrénován model na n–nejlepších nahrávkách, kde všechny zarovnání oproti předchozímu experimentu dosáhlo více jak 0,60 f-hodnoty, tedy špatné nahrávky na výsledné zarovnání neměly téměř žádný vliv. Dále bylo zkoušeno jak se mění zarovnání bude-li model natrénován různé kombinaci dobrých a špatných nahrávek. Většina výsledků se zase pohybovala okolo f-hodnoty 0,65, ale několik modelů, které byly nakonec trénovány špatnou nahrávkou dosáhly o něco horších výsledků (0,4–0,6 f-hodnoty). Horší výsledky byly pravděpodobně způsobený tím, že na rozdíl od modelů trénovaných na n-nejlepších nahrávkách nebyly modely natolik ustáleny. Dále bylo vyzkoušeno spojit jednotlivá zarovnání tak, aby výsledné zarovná obsahovalo pouze ta slova, které jsou zarovnána s určitou odchylkou stejně. Při spojení pouze dvou zarovnání bylo zjištěno, že výsledné zarovnání koreluje s podobností jednotlivých zarovnání. Tedy dvě špatná zarovnání si nejsou téměř podobná. Postupným spojením zarovnání, která měla f-hodnotu okolo 0,65, se zvedla přesnost z původních 0,65 až k 0,9 a f-hodnotou stále okolo 0,65. Nejlepší f-hodnoty této práce bylo dosaženo právě spojením dvou kvalitních zarovnání, výsledná f-hodnota byla vyšší jak 0,7.

Jako jazykově nezávislou metodu bylo zvoleno použít fonémový přepis jiného jazyka, než dané nahrávky. Konkrétně byl použit maďarský fonémový přepis. Následné zarovnání se provádělo stejně jako s anglickým fonémovým přepisem. Zarovnání bylo provedeno na modelu, který byl natrénován na každé nahrávce zvlášť a podle modelu získaného pouze z jedné nahrávky. Tyto experimenty nedosáhly dobrých výsledků, celková f-hodnota se pohybovala pouze okolo 0,02. Takto špatné výsledky mohou být způsobeny například tím, že hláska `th`, která se velmi často vyskytuje v anglickém jazyce, je v maďarském fonémovém přepisu identifikována jako zvuk pozadí. Dalším faktorem je, že se jedná o značně složitá data, která obsahují mnoho zvuků na pozadí a překrývající se hovory.

Při dalším pokračování v této práci, by se mohly brát v úvahu věrohodnosti fonémů a pravděpodobnosti u získaného zarovnání. Pro zarovnání nezávislá na jazyku by mohly být vyzkoušeny fonémové přepisy dalších jazyků, případně vyzkoušet zde zkoumané metody na jednodušších datech, například na audio knihách, které jsou většinou čtené jedním mluvčím, nahrávka odpovídá přepisu a obsahuje minimum okolního hluku.

Literatura

[1] *Forced Alignment*. [Online; navštíveno 12.01.2017].
URL https://www.isip.piconepress.com/projects/speech/software/tutorials/production/fundamentals/v1.0/section_04/s04_04_p01.html

[2] MGB overview. 2015, [Online; navštíveno 5.7.2017].
URL http://www.asru2015.org/Slides/AS15_Challenge3.pdf

[3] Edit distance. 2017, [Online; navštíveno 20.6.2017].
URL https://en.wikipedia.org/wiki/Edit_distance

[4] Grapheme. 2017, [Online; navštíveno 10.1.2017].
URL https://en.wikipedia.org/wiki/Grapheme

[5] Phoneme. 2017, [Online; navštíveno 10.1.2017].
URL https://en.wikipedia.org/wiki/Phoneme

[6] Precision and recall. 2017, [Online; navštíveno 15.1.2017].
URL https://en.wikipedia.org/wiki/Precision_and_recall

[7] Word error rate. 2017, [Online; navštíveno 5.1.2017].
URL https://en.wikipedia.org/wiki/Word_error_rate

[8] Bell, P.; Gales, M. J. F.; Hain, T.; aj.: The MGB challenge. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, IEEE, 2015, ISBN 9781479972913, s. 687–693, doi:10.1109/ASRU.2015.7404863.
URL http://ieeexplore.ieee.org/document/7404863/

[9] Bell, P.; Renals, S.: A system for automatic alignment of broadcast media captions using weighted finite-state transducers. *ASRU 2015*, 2013: s. 675 – 680.
URL http://homepages.inf.ed.ac.uk/srenals/pb-align-asru2015.pdf

[10] Bisani, M.; Ney, H.: Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, ročník 50, č. 5, 2008: s. 434 – 451, ISSN 0167-6393, doi:http://dx.doi.org/10.1016/j.specom.2008.01.002.
URL http://www.sciencedirect.com/science/article/pii/S0167639308000046

[11] Boháč, M.; Blavka, K.: Text-to-Speech Alignment for Imperfect Transcriptions. *Lecture Notes in Computer Science*, ročník vol 8082, 2013.
URL https://link.springer.com/chapter/10.1007/978-3-642-40585-3_67

[12] Hazen, T.: Automatic alignment and error correction of human generated transcripts for long speech recordings. *INTERSPEECH*, 2006,: s. 1606 – 1609.
URL https://groups.csail.mit.edu/sls/publications/2006/IS061258.pdf

[13] Hoffmann, S.; Pfister, B.: Text-to-speech alignment of long recordings using universal phone. *INTERSPEECH*, 2015: s. 1520 – 1524.
URL http://www.tik.ee.ethz.ch/spr/publications/Hoffmann:13a.pdf

[14] Moreno, P. J.; Joerg, C.; van Thong, J. M.; aj.: A Recursive Algorithm for the Forced Alignment of Very Long Audio Segments In 5th Int. Conf. on Spoken Language Processing. 1998.
URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.649.6346&rep=rep1&type=pdf

[15] Robert-Ribes, J.; Mukhtar, R.: Automatic generation of hyperlinks between audio and transcript. *Eurospeech*, 1997: s. 903 – 906.
URL http://mirlab.org/conference_papers/International_Conference/Eurospeech%201997/pdf/t3d/a1238.pdf

[16] Schwarz, P.: *Rozpoznávání fonémů z dlouhého časového kontextu*. Dizertační práce, Vysoké učení technické v Brně, Fakulta informačních technologií, 2009.
URL http://www.fit.vutbr.cz/study/DP/PD.php?id=109

Přílohy

Příloha A

Obsah přiloženého paměťového média

- projekt.pdf – tato technická zpráva
- sablona2017d – zdrojové soubory pro vytvoření projekt.pdf
- result – výsledky provedených experimentů
- run – skripty pro jednotlivé experimenty
- video.mp4
- plakat.pdf