RECOGNITION AND SEARCH IN SKYPE CALLS

Pavel Tomášek
Bachelor Degree Programme (3), FIT BUT
E-mail: xtomas23@stud.fit.vutbr.cz

Supervised by: Jan Černocký
E-mail: cernocky@fit.vutbr.cz

ABSTRACT

The project aims at a system for recognition Skype calls. This work also focuses on finding out high-quality Skype recorder. Signal processing (there are used techniques of phoneme recognition and logical segmentation), speech recognition, and results presentation follows. Essential part of this work is also index and search implementation. This project is a proof-of-concept and shows one of the ways of speech recognition utilization.

1 ÚVOD

Žijeme v době, kdy informace mohou mít vysokou cenu. A máme-li možnost s takovými informacemi rychle a efektivně nakládat, úspěch nemusí být daleko.

Tento projekt nastiňuje budoucí možnosti práce s audio informacemi. Nástroj automaticky rozpoznávající texty hovorů nejen usnadní práci při zpětném procházení komunikace s lidmi přes komunikační program Skype v textové podobě, ale pomůže i při vyhledávání slov v těchto konverzacích.

V následujících odstavcích přiblížím možný způsob vyhotovení takovéhoto projektu, o jaké pilíře se projekt opírá. V závěru nastíním možné směrování budoucího vývoje.

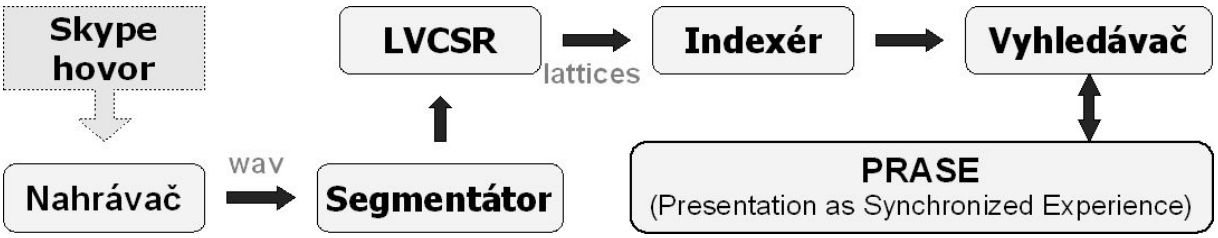
2 ROZBOR

Cílem projektu je rozpoznávání a vyhledávání ve Skype nahrávkách. Podstatnými částmi tohoto projektu jsou: aplikace zaznamenávající Skype hovory, mechanismy předzpracování signálu (Skype nahrávky) zahrnující fonémové rozpoznávání a logickou segmentaci. Dále samotný rozpoznávač řeči a indexace rozpoznané nahrávky (umožňuje následné vyhledávání v nahrávce). Jednotlivé prvky systému jsou vysvětleny níže.

Na obrázku 1 jsou vyobrazeny jednotlivé stavebními prvky projektu pro snazší orientaci.

2.1 NAHRÁVÁNÍ SKYPE HOVORŮ

Abychom mohli rozpoznávat, musíme nejdříve mít co rozpoznávat – je třeba umět kvalitně zaznamenat hovor. Za tímto účelem byl vybrán program *Skype Capture* (autorem je *Jiří Šimáček*).



Obrázek 1: Schéma projektu

Tato aplikace je v prostředí MS Windows schopna spolupracovat s programem Skype a zaznamenávat nahrávky v originální kvalitě. Výstupem je stereo záznam složený ze dvou kanálů, tedy místní a vzdálená strana konverzace.

2.2 PŘEDZPRACOVÁNÍ SIGNÁLU

Dále je důležité nahrávku připravit k rozpoznávání. Tato příprava zahrnuje úpravu vstupního signálu (nahrávky) převzorkováním na požadovanou frekvenci (8 kHz v případě implementovaného rozpoznávače). Dále je třeba rozpoznat jednotlivé fonémy v nahrávce. O to se stará fonémový rozpoznávač vyvinutý skupinou Speech@FIT [1], [2]. Pro potřeby zpracování je dále nutné nahrávku rozdělit na menší části (segmentovat). Nahrávka je *logicky* rozdělena na krátké úseky, které bude rozpoznávač postupně zpracovávat.

2.3 ROZPOZNÁVÁNÍ ŘEČI

Nahrávka je zpracována systémem LVCSR (Large Vocabulary Conctinuous Speech Recognizer). Rozpoznávač využívá trénovacích modelů telefonních hovorů, čemuž se Skype komunikace podobá. Rozpoznávač byl převzat z projektu AMI(DA), na němž Speech@FIT spolupracuje. Použitá verze má ovšem omezený slovník i jazykové modely za účelem urychlení zpracování. Obsahuje pouze jednodušší akustické modely, bez adaptace mluvčího. K běhu potřebuje přibližně 200 MB operační paměti. Na průměrném počítači je asi šestkrát pomalejší než real-time. Je možné získat dva typy výstupu. Buď řetězec nejpravděpodobnějších slov (tzv. *1-best*), nebo slovní graf hypotéz (tzv. *lattice*) sloužící k indexaci, následně umožňující vyhledávání.

Úspěšnost rozpoznávání velmi závisí na dokonalosti angličtiny zaznamenané v nahrávce. U rodilých mluvčích se úspěšnost pohybuje kolem 70%. U lidí pokročilých v anglickém jazyce (ne-rodilých mluvčích) je v lepším případě správně rozpoznáno každé druhé slovo.

2.4 INDEXACE

Výstup z rozpoznávače je indexován. Slova jsou transformována na číselné identifikátory. Je vytvořen invertovaný index a výstup je dále uložen v binární formě, což urychluje přístup.

2.5 VYHLEDÁVÁNÍ

Jakmile uživatel zadá požadavek k vyhledání, zadaný řetězec je vyhledáván v invertovaném indexu. Poté je vygenerován seznam možných kandidátů. Konečným výstupem je tak seznam hypotéz seřazený dle pravděpodobností.

2.6 PREZENTAČNÍ SOFTWARE

Veškeré dosavadní snažení by mohlo přijít vniveč, nebylo-li by možné vhodně prezentovat výsledky rozpoznávání. Proto byla použita aplikace *PRASE* (Presentation as Synchronized Experience) vyvinutá skupinou Speech@FIT. Je to multiplatformní multimediální přehrávač využívající prostředí wxWidgets schopný přehrávat záznam a zobrazovat odpovídající rozpoznanou řeč (1-best výstup rozpoznávače). Umožňuje také zadávat požadavky k vyhledávání.

2.7 ČASOVÁ NÁROČNOST

V následující tabulce je uveden čas, jaký je v současnosti potřebný ke kompletnímu zpracování nahrávky (předzpracování, rozpoznání, indexace).

Výpočetní stroj	**Délka nahrávky** → **délka zpracování** (sekundy)	
Laptop Compaq Evo N600c, Pentium III 1.2GHz, 512MB RAM	*00:00:50* → *00:18:30*	*00:04:06* → *02:02:30*
Laptop ASUS A6000, Centrino Duo T2300 1.6GHz, 1024MB RAM	*00:00:50* → *00:08:18*	*00:04:06* → *00:53:20*

2.8 BUDOUCÍ VÝVOJ

Ve snaze zlepšit úspěšnost rozpoznávání bude implementován novější rozpoznávač. Bude aktualizován i fonémový rozpoznávač – zde dojde ke zlepšení z časového hlediska.

Neustálý vývoj Multi-media Browseru se do budoucna projeví i v tomto projektu.

Plánuje se umístění projektu na Linuxový server s možností uložení Skype hovorů a následným zpracováním. Výsledkem bude stáhnutelný balíček spustitelný pomocí Multi-media Browseru.

Implementovaný systém je schopen zpracovat jen obvyklý Skype hovor mezi dvěma lidmi. Dalším krokem tedy může být i zpracovávání konferenčních hovorů (identifikací mluvčího).

3 ZÁVĚR

Vzhledem k tomu, že v projektu je použit pouze základní rozpoznávač (bez adaptivních technik), je zpracování nahrávky rychlé, avšak úspěšnost rozpoznávání je tím výrazně snížena.

Tento projekt demonstruje široké možnosti uplatnění technik vyvinutých skupinou Speech@FIT pro rozpoznávání řeči. Skype budiž pouze názorným příkladem. Podobným způsobem se dají zpracovávat téměř jakékoli (kvalitnější) zvukové záznamy mluvených projevů.

REFERENCE

[1] Schwarz Petr, Matějka Pavel, Černocký Jan: Towards Lower Error Rates in Phoneme Recognition, In: Proceedings of 7th International Conference Text,Speech and Dialoque 2004, Brno, CZ, Springer, 2004, ISBN 3-540-23049-1

[2] Schwarz Petr, Matějka Pavel, Černocký Jan: Hierarchical structures of neural networks for phoneme recognition, In: Proceedings of ICASSP 2006, Toulouse, FR, 2006