AUTOMATIC SPEAKER DEPENDENT KEYWORD SPOTTING

František VOSTÁL, Master Degree Programme (5)
Dept. of Radio Electronics, FEEC, BUT
E-mail: xvosta01@stud.feec.vutbr.cz

Supervised by: Doc. Ing. Milan Sigmund, CSc.

ABSTRACT

This article describes new developed method for speaker dependent keyword spotting in audio files based on mel-frequency cepstral coefficients as features of speech waveforms. The used search technique can be called “sliding keyword”. It means that a keyword slides along the test speech and in each position a distance between the corresponding part of test speech and keyword is computed. Dynamic time-warping algorithm is used to provide an efficient time alignment.

1 ÚVOD

Problematika vyhledávání klíčových slov ve zvukových souborech se rozvíjí s množstvím dostupných zvukových dat a potřebou jejich prohledávání. Oproti zápisu textu má zvuková forma výhodu v rychlosti záznamu, ale dochází k problému při potřebě rychlého prohledávání zvukových souborů a získávání informací z pořízených nahrávek.

Metoda, použitá v této práci, je založena na postupném přikládání klíčového slova k úseku testované promluvy a výpočtu jejich vzdálenosti. Pokud tato vzdálenost, představující míru shody, klesne pod určitou hranici, je v daném místě promluvy detekováno klíčové slovo. Rozdíly v délce klíčového slova a slova obsaženého v úseku prohledávané promluvy se přizpůsobují metodou dynamického borcení času (Dynamic Time Warping).

2 PARAMETRIZACE ŘEČOVÉHO SIGNÁLU

Při zpracování je nutno řečový signál určitým způsobem popsat. Nepracuje se tedy přímo se zvukovým signálem reprezentovaným tvarem vlny, ale s jeho parametrizovanou formou. U řečového signálu se předpokládá jeho stacionarita v úseku délky 10 až 20 ms, což odpovídá rychlosti změny artikulačního traktu. Řeč tedy rozdělíme na úseky této délky a z každého segmentu spočítáme vektor parametrů, který bude daný úsek popisovat. Jedním z nejpoužívanějších způsobů parametrizace řečového signálu jsou mel frekvenční kepstrální koeficienty (Mel Frequency Cepstral Coefficients – MFCC). Jejich výpočet spočívá

v určení spektra (DFT) segmentu, jeho umocnění a váhování bankou filtrů. Tím získáme energii v jednotlivých frekvenčních pásmech. Ta je pak logaritmována a zpětnou Fourierovou transformací dostaneme kepstrální koeficienty. Podrobnější popis lze nalézt v [1],[3].

3 MÍRA VZDÁLENOSTI ŘEČOVÝCH RÁMCŮ

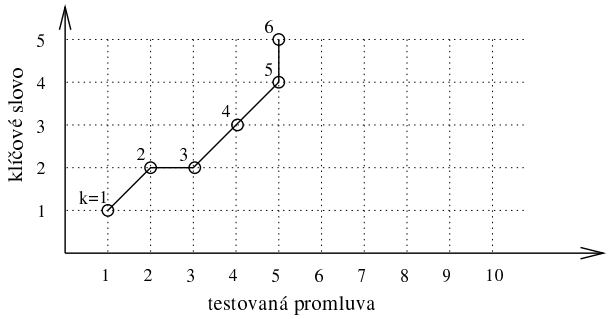
Prohledávání zvukového souboru spočívá v postupném posouvání klíčového slova podél testované promluvy a určení jeho podobnosti s aktuálním úsekem testované promluvy. Tato podobnost je označována jako *míra vzdálenosti* nebo *míra zkreslení*. Pro kepstrální koeficienty se nejčastěji používá *kepstrální míra*:

$$d_{CEP}(t,r) = \sum_{i=1}^{P} [c_t(i) - c_r(i)]^2 \tag{1}$$

kde $d_{CEP}(t,r)$ je vzdálenost mezi segmentem t testovaného signálu a segmentem r referenčního signálu, $c_t(i)$ a $c_r(i)$ jsou parametrizované signály a P určuje počet kepstrálních koeficientů. Při výpočtu se navíc koeficienty váhují (viz. [1]), protože ne všechny mají stejnou důležitost pro určení vzdálenosti.

4 APLIKACE METODY DYNAMICKÉHO PROGRAMOVÁNÍ

Metoda dynamického borcení času slouží k nelineární transformaci časové osy jednoho signálu tak, aby měl stejnou délku jako signál druhý a svým průběhem se mu co nejvíce podobal. Princip metody je naznačen na obrázku 1. Výpočet vzdálenosti začíná



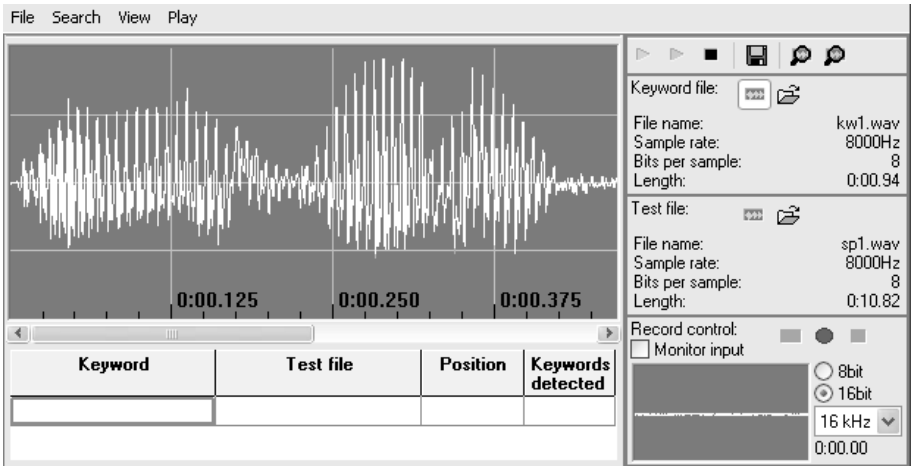
Obrázek 1: DTW cesta pro srovnání klíčového slova a úseku testované promluvy.

prvním segmentem obrazu klíčového slova a testované promluvy a postupuje se ve směrech daných tzv. *lokálním omezení cesty* (popis typů viz. [1]). V příkladu je použit typ I. Ten umožňuje pouze tři cesty a to o jeden segment vpravo, nahoru nebo diagonálně. V každém bodě se vypočítá vzdálenost podél cest, po kterých se lze do daného místa dostat a uvažuje se nejmenší dosažená hodnota. Tak se postupuje až do koncového bodu, kde dostaneme výslednou minimální vzdálenost. Počáteční bod se posune o jeden segment a postup se opakuje. Pokud hodnota vzdálenosti klesne pod určitou prahovou hodnotu, je v daném místě detekováno klíčové slovo.

5 VYTVOŘENÝ PROGRAM

Program je napsán v jazyce C++ s využitím knihoven *wxWindows* pro vytvoření grafického uživatelského rozhraní. To umožňuje překlad programu na různých platformách (Windows, Linux, Macintosh) beze změny zdrojového kódu. Primárně byl program vyvíjen v systému *Windows* a testován byl také v *Linuxu*. Blokové schéma hlednání klíčového slova v promluvě je na obrázku 2.

Obrázek 2: Blokové schéma hledání klíčových slov.



Obrázek 3: Výsledný program.

Vytvořený program umožňuje načíst soubor s testovanou promluvou a klíčovým slovem z disku nebo ho nahrát přes mikrofon. Volitelná je kvalita záznamu. V testované promluvě se potom vyhledává buď přesně celé klíčové slovo nebo jeho označená podstatná část (podobně jako hledání v textu např. ve Wordu). Výsledkem je tabulka s nalezenými časovými pozicemi klíčových slov.

REFERENCE

[1] Psutka J. *Komunikace s počítačem mluvenou řečí*. Academia. Praha, 1995

[2] Sigmund M. *Analýza řečových signálů*. Skripta FEI VUT v Brně. Brno: MJ Servis, 2000

[3] Černocký J. *Zpracování řečových signálů – texty k přednáškám*. Dokument dostupný na URL http://www.fee.vutbr.cz/˜cernocky/Students (říjen 2001).