PITCH DETECTOR IN SPEECH PROCESSING

Dalibor PERNICA, Master Degree Programme (5)
Dept. of Computer Graphics and Multimedia, FIT, BUT
E-mail: xperni10@stud.fit.vutbr.cz

Supervised by: Dr. Petr Motlíček

ABSTRACT

The paper presents a novel method for detection of pitch in speech processing. The work focuses on a time domain algorithm for segmenting voiced speech that uses similarity of adjacent segments. The experimental results while testing this detector are presented and compared to OGIvox speech database.

1 ÚVOD

Základním tónem rozumíme základní kmitočet (*pitch*) na kterém kmitají hlasivky. Spolu s formantovými frekvencemi patří mezi základní fonetické charakteristiky řečového signálu. Využívá se zejména v syntetizátorech a kodérech řeči.

Existují dvě skupiny metod pracující v časové nebo ve frekvenční oblasti. V časové oblasti se využívá podobnosti řečových úseků a ve frekvenční např. vlastností lichých harmonických. Dále bude věnována pozornost detektoru pracujícím v časové oblasti.

2 SEGMENTACE V ČASOVÉ OBLASTI

Cílem segmentace je rozdělení řečového signálu na takové části, z nichž délka každé části přímo odpovídá periodě základního tónu ve zkoumaném úseku.

Mějme znělý řečový signál a v něm dva sousedící segmenty U_1 a U_2 začínající v místě, kde signál protíná časovou osu při vzestupu ze záporných hodnot do kladných. Nejjednodušším způsobem, jak ohodnotit podobnost segmentů, je určit energii rozdílového signálu mezi odpovídajícími si vzorky podle (1).

$$d(U_1,U_2) = \sum_{n=1}^{\min(n_1,n_2)} (x_1[n] - x_2[n])^2, \tag{1}$$

kde $U_i = (x_i[1], x_i[2], \ldots, x_i[n_i])$ pro $i = 1,2$ jsou segmenty signálu a n_i pro $i = 1,2$ jsou délky segmentů.

Nyní hledáme takovou posloupnost dělících bodů pro kterou je součet dílčích vzdáleností mezi sousedícími segmenty minimální, přičemž posloupnost musí začínat (končit) ve vzdálenosti d_{min} od začátku (konce) signálu a dále největší možná délka segmentů je d_{max}. Hodnoty udává (2).

$$d_{min} = \frac{F_s}{F_{max}}, \quad d_{max} = \frac{F_s}{F_{min}}, \tag{2}$$

kde F_s je vzorkovací frekvence, F_{min} a F_{max} je minimální a maximální hledaná frekvence základního tónu.

2.1 ALGORITMUS

Reprezentací každého segmentu jako uzlu v grafu s hranami směřujícími z předcházejících do následujících segmentů ohodnocených energií rozdílového signálu je možné problém segmentace převést na problém hledání minimální cesty grafem.

```
01 function Segmentation(signal, Fs, Fmin, Fmax, C)
02 dmin = Fs / Fmax; dmax = Fs / Fmin;
03 s = dc_removal(signal);
04 s = lowpass_filter(s, Fmax + 100);
05 P = find_eligible_points(s); pd = [];
06 foreach (pt in P) do {
07   L = get_left_segments(pt, dmin, dmax);
08   R = get_right_segments(pt, dmin, dmax);
09   foreach (l in L) and (r in R) do begin
10     if length(l) < C * length(r) then
11       pd = push(pd, [l, r, distance(l,r)]); }
12 tp = get_minimal_path(pd);
13 Cp = get_cutting_points(signal, tp);
```

Obrázek 1: Algoritmus segmentace

Po ustředění signálu (03) jsou dolní propustí (04) odstraněny neužitečné vyšší harmonické a následuje hledání dělících bodů (05). Pro každý dělící bod (06) jsou vyhledány předcházející (07) a následující (08) segmenty a uloženy do seznamu (11), přičemž se uvažují pouze segmenty s daným poměrem délek (10), kde `length(r)` představuje kratší délku z obou segmentů (konstanta `C` se zpravidla volí 1,2). Ze seznamu (grafu) je vybrána optimální cesta (12) a zjištěny jednotlivé segmenty (13).

Doba zpracování závisí na vzorkovací frekvenci F_s, na rozpětí minimální a maximální frekvence $[F_{min}, F_{max}]$ a na povaze zkoumaného signálu.

Frekvence základního tónu se vypočítá podle (3) ze vzdáleností d ve vzorcích jednotlivých dělících bodů nalezené minimální cesty.

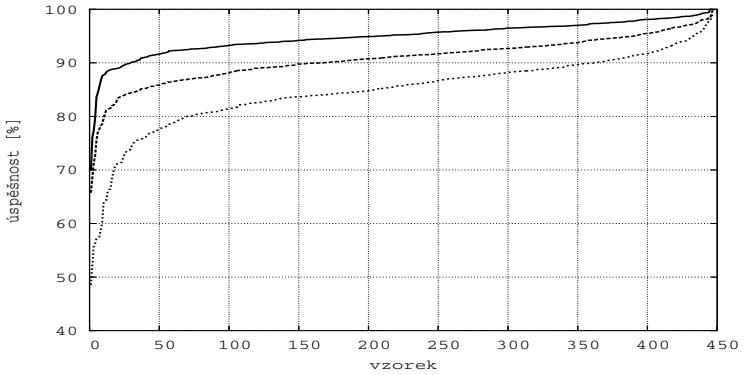
$$F_0 = \frac{F_s}{d} \tag{3}$$

Průměrná chyba určení vzdálenosti d je 1 vzorek, která po převodu na frekvenci může dosáhnout 40 Hz při $F_s = 8000$ Hz a pro základní tón 600 Hz. Proto je vhodné zvýšit přesnost výpočtu vzdálenosti lineární interpolací polohy dělících bodů.

3 VÝSLEDKY

Výsledky byly testovány na 450 řečových nahrávkách z referenční databáze OGIvox a pro každý vzorek určena úspěšnost jako poměr správně určených frekvencí vůči počtu rámců v nichž se mluví. Tolerance byla zvolena ± 15 Hz. Úspěšnost ilustruje obrázek 2.

Na svislé ose je úspěšnost v procentech a na vodorovné ose jsou referenční nahrávky. Hodnoty úspěšnosti byly seřazeny vzestupně. Horní průběh byl získán zanedbáním rámců s detekovanou nulovou frekvencí (způsobenou pozdním rozběhem segmentace). Dolní průběh zachycuje úspěšnost bez zanedbání nulových rámců. Prostřední průběh byl získán po nasazení mediánového filtru (bez zanedbání nulových rámců).



Obrázek 2: Úspěšnost segmentace (popis v textu)

Podle prostředního (v praxi očekávatelného) průběhu se podařilo 300 souborů určit s úspěšností vyší než 90 % a 410 souborů s úspěšností nad 85 %. Nízká úspěšnost pod 80 % je způsobena především detekcí poloviční a násobné frekvence.

4 ZÁVĚR

Metoda vyžaduje kvalitní průběhy signálu. Malá amplituda a/nebo prudká změna tvaru a frekvence signálu má v řadě případů za následek detekci poloviční nebo násobné frekvence. Chyby tohoto typu lze omezit zpracováním signálu po částech doplněné dvoupruchodovou segmentací pro normální a zrcadlově otočený signál. Díky zpracování signálu po částech se zvyšuje rychlost detekce (průchodu grafem) a je možné uvažovat o nasazení detektoru pro zpracování v reálném čase.

REFERENCE

[1] Petrushin, V. A.: Pitch-Synchronous Speech Signal Segmentation and Its Applications, Proceedings of TSD 2003, České Budějovice, 2003.

[2] Psutka, J.: Komunikace s počítačem mluvenou řečí, Academia Praha, 1995.

[3] Gold B., Morgan N.: Speech and Audio Signal Processing, John Wiley & Sons, Inc. 1999.