Detekce hlasivkových pulsů v řečových signálech

Jakub Kopřiva[1]

1 Úvod

Tato práce se zabývá problémem automatické detekce hlasivkových pulzů ("automatic pitch-marking") v řečových signálech. Algoritmy řešící tuto úlohu jsou jedním ze základních procedur systémů pro syntézu řeči, zvláště pak těch na bázi konkatenačního přístupu. V takových systémech syntézy řeči dochází k řetězení velmi krátkých řečových segmentů. Rozdělení databázového řečového signálu na mikrosegmenty se provádí právě pomocí co nejpřesněji detekovaných hlasivkových pulzů.

Cílem práce bylo realizovat některé algoritmy pro detekci hlasivkových pulzů v řeči a výsledky porovnat s výsledky známých fonetických nástrojů (PRAAT, GLOAT). Snahou bylo se úspěšností co nejvíce přiblížit výsledkům algoritmu MPA[2] („Multi-phase algorithm"), který byl vyvinut na pracovišti FAV/KKY. Ten používá i hlasivkový (EGG) signál a je proto mezi porovnávanými algoritmy považován za pomyslný „strop". Vyrovnáním se úspěšnosti algoritmu MPA by zanikla potřeba získávání EGG signálu.

Ukázalo se, že pro přesnou detekci hlasivkových pulzů bývá zapotřebí co nejpřesnější kontura základního hlasivkového tónu. V této práci bylo proto zahrnuto i porovnání algoritmů detekce základního hlasivkového tónu ("pitch-tracking").

2 Porovnání algoritmů výpočtu základního hlasivkového tónu

Přístupů pro získání kontury základního hlasivkového tónu je větší množství. Základní stále používaný přístup je autokorelace hledající maxima krátkodobé autokorelační funkce. Složitější přístupy mohou například využívat hledání optimální cesty v kepstrogramu, jak je uvedeno v [1].

V této části práce bylo porovnáno celkem 5 dostupných metod (RAPT, AMDF, EWENDER, PRAAT, GLOAT) podle jejich přesnosti ve smyslu RMSE i podle vzájemné Pearsonovy korelace:

$$RMSE(R,T) = \sqrt{E((R-T)^2)} \tag{1}$$

$$COR(E,T) = \frac{E(RT) - E(R)E(T)}{\sqrt{E(R^2) - E^2(R)}\sqrt{E(T^2) - E^2(T)}} \tag{2}$$

Tabulka 1 obsahuje dosažené výsledky. V obou případech bylo nejlepších výsledků dosaženo pomocí metody EWENDER, která byla doplněna informací o (ne)znělosti z metody RAPT.

	RAPT	AMDF	EWENDER	PRAAT	GLOAT
RMSE [Hz]	20,68	41,93	**15,39**	22,33	24,32
COR	0,771	0,615	**0,783**	0,733	0,699

Tabulka 1: Dosažené přesnosti algoritmů detekce základního hlasivkového tónu

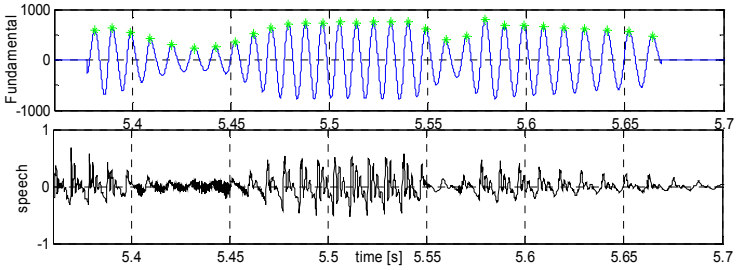
[1] student navazujícího studijního programu Aplikované vědy a informatika, obor Kybernetika a řídicí technika, specializace Umělá inteligence, e-mail: kopr@students.zcu.cz

[2] http://www.kky.zcu.cz/cs/publications/LegatM_2011_Onthedetectionof

3 Porovnání algoritmů pro detekci pitch marků

V hlavní části práce byly implementovány algoritmy prezentované v článcích [2] a [3]. V článku [2] se jedná o pitch markovací algoritmus, který byl realizován neuronovou sítí. Počet neuronů ve skryté vrstvě byl zvolen na hodnotu 20. Z klasifikovaného signálu se vytvoří sada parametrů obsahující informace o sousedních maximech a minimech. V trénovací fázi je použit algoritmus "Backpropagation". Neuronové síti se předkládají vstupy, a z referenčních ručně určených pitch marků požadované hodnoty výstupů. Tím dojde k natrénování a síť je připravena k použití ve fázi pracovní.

Základem algoritmu v [3] jsou dvě funkce a procedura pro hledání vrcholků. První funkcí je krátkodobý průběh energie, jehož lokální maxima podle předpokladu odpovídají pitch markům. Tento signál však musí být použit spolu s fundamentální vlnou, která se počítá pomocí konvoluce pouze řečového signálu s Hammingovým okénkem. Vznikne tak hladká vlna s frekvencí odpovídající základnímu hlasivkovému tónu. Detekce pitch marků je realizována kombinací obou funkcí.



Obrázek 1: Fundamentální vlna a řečový signál

Vzhledem k podobnosti fundamentální vlny (F vlna) a EGG signálu byl proveden experiment, kdy v algoritmu MPA byl EGG signál nahrazen F vlnou. Byly tak překvapivě dosaženy jen o málo horší výsledky. Výsledky byly zaznamenány do tabulky 2.

metoda	přesnost	metoda	přesnost
MPA + EGG signál	93,51%	PRAAT (cc)	89,91%
MPA + F vlna	92,64%	PRAAT (ac)	89,71%
Neuronová síť [2]	91,41%	GLOAT	87,34%
MPA + F vlna (spojitá)	**91,23%**	Energie a fundamentální vlna [3]	86,63%

Tabulka 2: Dosažené přesnosti algoritmů detekce hlasivkových pulsů

4 Závěr

Použitím F vlny místo EGG signálu byly dosaženy uspokojivé výsledky. V případech, kde není k dispozici EGG, může být použita simulace pomocí F vlny se ztrátou přesnosti 2,27%.

Poděkování: Příspěvek byl podpořen grantovým projektem ZČU v Plzni č. SGS-2010-054.

Literatura:

[1] EWENDER, T. - HOFFMAN, S. - PFISTER, B. Nearly Perfect Detection of Continous F0 Contour and Frame Classification for TTS Synthesis. In *Proceedings of Interspeech 2009*. Brighton, UK, 2009, pp. 100-103

[2] BARNARD, E. - COLE, R. A. - VEA, M. P. - ALLEVA, F. A. Pitch detection with a neural-net classifier. *IEEE Transactions on Signal Processing*. Vol. 39, No. 2, February 1991, pp. 298-307.

[3] EWENDER, T. - HOFFMAN, S. - PFISTER, B. Nearly Perfect Detection of Continous F0 Contour and Frame Classification for TTS Synthesis. In *Proceedings of Interspeech 2009*. Brighton, UK, 2009, pp. 100-103.