Beszéd alapfrekvencia követés hatékony zöngésség detektálással

Bárdi Tamás

Pázmány Péter Katolikus Egyetem, Információs Technológia Kar
1083 Budapest, Práter u 50/A
bardi.tamas@itk.ppke.hu

Kivonat: A beszédjel alapfrekvenciát meghatározó algoritmusok, más néven pitch detektorok helyes működése csak úgy lehetséges, ha az automatikus zöngés-zöngétlen megkülönböztetés is megbízható. Az alábbiakban ismertetjük pitch detektorunkat, melyben a zöngésség detektálása a konkurens módszereknél alacsonyabb hiba százalékkal működik. Algoritmusunk a jól ismert autokorrelációs módszeren alapszik. Algoritmusunk zöngésség detektáló erejét egy olyan adatbázison vizsgáltuk, melyben a beszéddel szinkronban laryngográf jelet is rögzítettek.

1. Bevezetés

Az emberi hallás modern elméletei hitelt érdemlően megállapították, hogy a hangmagasság (pitch) észlelés nem mindig van egy-egy értelmű kapcsolatban az alapfrekvenciával (F0). Ennek ellenére a digitális beszéd-feldolgozásban az F0 becslő módszereket hagyományosan pitch detektor algoritmusoknak (PDA) nevezik. A tényleges beszéddallamot jól közelítő pitch kontúr sok alkalmazásban hasznosítható. Jelentős szerepe van a prozódikus elemzésekben. Ilyen például a mondat hangsúlyos helyeinek megtalálása a hanglejtés alapján, vagy a kérdő és kijelentő mondatok automatikus megkülönböztetése. A beszédfelismerés a tonális nyelveken, mint például a kínai vagy a vietnami, megoldhatatlan pitch detektor nélkül.

A szakirodalomban pitch detektor témában jó néhány módszer látott napvilágot az elmúlt évtizedekben [10], a legszélesebb körű áttekintésük Hess-nél olvasható [7]. A megoldások többsége mérsékelt teljesítményével elégedetlenségre adhat okot, de azért van néhány egészen jó is. Ilyen Bagshaw eSRPD [3, 4] módszere, amely kevesebb, mint 1%-ban becsli rosszul az alapfrekvenciát, ha zönge van a beszédben. De a zöngés gerjesztés meglétét vagy hiányát már 3-4% hibával detektálja.

Általánosságban elmondható, hogy nyelvtani jelentéssel bíró pitch csak a zöngés szegmentumokon figyelhető meg. Ezért pitch frekvencia meghatározásának feltétele a jó zöngésség detekció. A zöngés-zöngétlen megkülönböztetés (V/UV - voiced/unvoiced) szerepe a beszédfelismerésben is jelentős, hiszen számos olyan szópár van, pl. köt - köd, melyek kiejtésben csak egyik mássalhangzójuk zöngésségében különböznek.

Egy zöngésség meghatározására szolgáló algoritmus (VDA - voicing determination algorithm) gyakran implicit része egy PDA-nak vagy beszédfelismerőnek, de megvalósítható különállóan is. Számos VDA született [7] már különféle elméletek bevetésével, közülük néhány igazán figyelemre méltó, jó teljesítményt azonban csak nagyon kevés mutat. A pitch detektoroknál általában a V/UV tévesztések nagyobb százalékban fordulnak elő, mint az F0 becslési hibák. Atal és Rabiner [1, 2, 8] egy öt döntési paramétert használó VDA-val próbálkozott statisztikus mintázat-felismerési megközelítést alkalmazva. Módszerük 4%-os hibaarányt adott egy nehezebb feladat megoldásában, nevezetesen a zöngés/zöngétlen/csendes (nincs beszéd) (V/U/S - voiced/unvoiced/silent) osztályozásban az egyszerűbb zöngés/zöngétlen (V/UV) döntés helyett.

Egy PDA-t építettünk, melyben egy hatékony beépített zöngésség detektor működik. Algoritmusunk az autokorreláció függvényen (ACF) alapszik. A zönge detekcióban módszerünk 2%-hoz közeli hibaarányt ért el. Az algoritmus, ha az ACF számításához FFT-t alkalmazunk, kevesebb, mint 2 megaflop per szekundum processzszorigénnyel megvalósítható 8 kHz-es mintavételezés mellett.

Az alábbi szakaszok az algoritmus moduláris szerkezetének megfelelően szerveződtek. A 2. szakasz az előfeldolgozó részt tárgyalja. Preprocesszorunkat úgy terveztük, hogy a V/UV megkülönböztetést a lehető legjobban segítse, az említett hibaarány elérésében nélkülözhetetlen szerepet játszik.

Az előfeldolgozás után a beszédjelből rövid időtartamú szakaszok kerülnek a basic extractor-nak nevezett egységhez. Itt számítjuk az ACF-et, majd ebből nyerjük a V/UV döntéshez és az F0 becsléshez szükséges paramétereket. Ebből a részből "halszálka" módszer alkalmazása érdemel említést, amely az "F0 a felső limiten" típusú hibákat csökkenti. Mindezeket a 3. szakasz tárgyalja.

Az egyszerű, de hatékony beépített VDA részletezése és kiértékelése a 4. szakasz és egyben cikkünk fő tárgya. A V/UV döntés két paraméteren alapszik, mindkettőt egy-egy küszöbbel hasonlítjuk össze. Ez a kétküszöbös módszer szintén hozzájárult a hibaszázalék csökkenéséhez. A szakirodalomban szokásos az előállított pitch kontúrok utólagos simítására egy posztprocesszort alkalmazni. Ilyet mi nem használtunk, mert a vizsgálatunk célja a beépített VDA képességének felmérése volt. A kiértékelésben a fókuszt a megbízható zöngésség detektálásra helyeztük.

2. A beszédjel előfeldolgozó

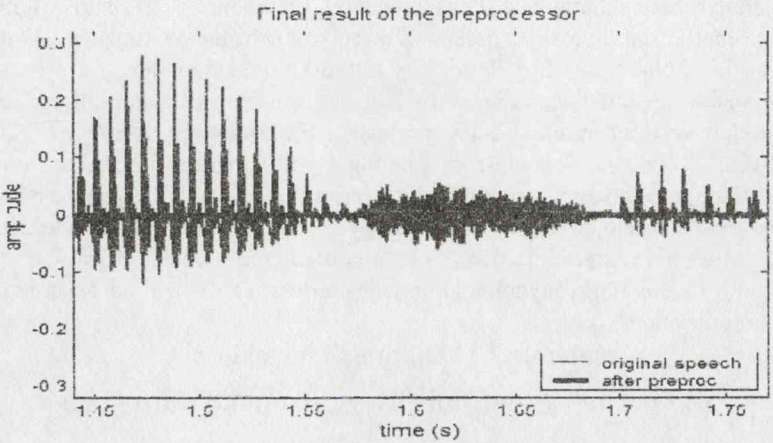
Általában egy PDA három fő komponensből épül fel: 1) preprocesszor, 2) basic extractor, 3) posztprocesszor. A preprocesszor fő feladata úgy transzformálni a beszédjelet, hogy utána az F0 becslés és a zönge detektálás könnyebb legyen.

A basic extractor rendszerint a beszédjelből vett tipikusan 20-50 milliszekundumos ablakokon dolgozik. A megkülönböztetés azonban, hogy mely műveletek tartoznak a preprocesszorhoz és melyek a basic extractor-hoz nagyon gyakran csak formális jelentőségű. Ha előbb kivesszük az ablakot a beszédjelből, majd azon futtatjuk a preprocesszort, akkor egyrészt fölöslegesen duplikálunk egy csomó számítást, ha az ablakok átfedik egymást, másrészt a preprocesszor és a basic extractor munkáját nehéz lesz külön-külön vizsgálni. Ha így teszünk, nem tudjuk például összefüggően meghallgatni a preprocesszorból kijövő jelet. A javaslatunk, hogy inkább futtassuk a

preprocesszort a beszédjel teljes hosszában, majd ebből vegyünk ablakokat és küldjük őket a basic extractor-hoz elemzésre. Ha így teszünk, érzékszervileg megfigyelhetővé válik a rendszer egy belső állapotában. Érzékszervi ellenőrző pontok elhelyezése egy összetett beszédfeldolgozó rendszer belsejében segítheti az empirikusan optimálandó paraméterek szerencsés megválasztását.

Preprocesszorunkban alul-áteresztő szűrést és ún. centerclip-et, magyarul középre vágást használunk. Mindkettő igen elterjedt a pitch detektorok szakirodalmában [6, 9, 11]. Az alul-átengedő szűrőnk Csebisev I-es típus, a levágási frekvencia 1200 Hz.

Az adaptív középre vágás technikája időben változó vágási szintet alkalmaz, mely a jel amplitúdójának függvényében változik. Általában ez a változó középre vágási szint a beszédjel valamilyen burkolójának egy rögzített százaléka. A módszerünkben az újítás, hogy kombinálja a két lépést, az alul áteresztő szűrést és a középre vágást. A burkolót az eredeti beszédjel amplitúdójából számítjuk, majd ennek 40%-át alkalmazzuk változó középre vágási szintként, de már a szűrt jelen. Mivel a tisztán sztohasztikus gerjesztésű beszéd szegmentumokon általában ennél nagyobb a nagy frekvenciás komponensek részaránya, a módszerünk a zöngétlen mássalhangzókat gyakorlatilag mindenütt nullára redukálja (1.ábra). Ez az effektus jelentősen javítja az automatikus V/UV döntés esélyeit.



1.ábra: Az eredeti beszédjel és a preprocesszor kimenete.

3. A basic extractor

A PDA-nak ez a része először a beszéd ablak autokorreláció függvényét számítja ki, majd az algoritmus az ACF "legjobb" csúcsát keresi meg. Az ACF értéke a kiválasztott csúcsnál, mint a periodicitás egy mértéke a zöngésség detektálására szolgál, a csúcs eltolási ideje pedig a periódus időt becsli. De hogy találjuk meg a "legjobb" csúcsot? Amint azt a későbbiekben látni fogjuk, a "legjobb" lokális maximum koránt sem feltétlenül globális is egyben.

Elöljáróban megjegyezzük, hogy az összes itt leírt képletben az idő dimenziójú változók és konstansok (τ, t, u, W) másodpercben értendők, a beszédjel kezelése analóg: integrálokkal, folytonos idővel és amplitúdóval. Az amplitúdót a rendszerben feldol-

gozható maximális amplitúdó arányában jelöljük: $-1.0 \le x(t) \le 1.0$. A fenti jelölésekkel biztosítjuk a tárgyalás függetlenségét a mintavételi frekvenciától és bitmélységtől. Konkrét alkalmazásban a mintavételi frekvencia és a minták számábrázolása ismeretében formuláink könnyen a megfelelő digitális változatra konvertálhatók.

A rövid távú autokorrelációnak a jelfeldolgozásban gyakran használt "rézsútos" (biased) definíciója helyett de Cheveigné [5] javaslata alapján annak "egyenes" (unbiased) definícióját használjuk, majd az ACF-et mesterségesen lejtősítjük. (W az ablak hossza, a vizsgálat során 32 ms-t használtunk)

$$r_t(\tau) = \frac{\int_{t-W/2}^{t+W/2} x(u)x(u-\tau)du}{\int_{t-W/2}^{t+W/2} x(u)^2 du} \quad (\tau, t, u, W \text{ szekundumban}) \quad (1)$$

és a mesterséges lejtés (a *gr* tényezővel szabályozhatjuk az erősségét):

$$r_t^{biased}(\tau) = r_t(\tau) \cdot (1 - gr \cdot \tau) \quad (2)$$

Az ACF lejtése oktáv tévesztés elkerülése miatt fontos, így a tényleges alapperiódusnak előnyt biztosíthatunk a többszöröseivel szemben. A "rézsútos" definíció a lejtést automatikusan biztosítja, de ennek mértéke kizárólag W-től függ. A mesterséges lejtéssel az ablak hossz és a "lejtőszög" külön-külön hangolható.

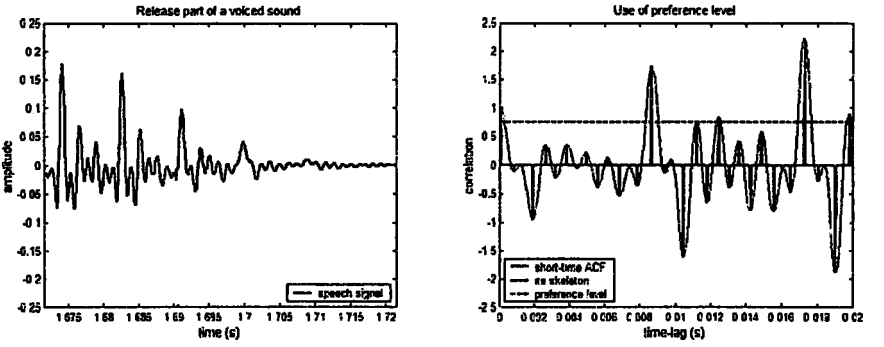
Mélyhangok kezdeti szakaszán az ACF gyakran a keresési intervallum szélén nagyobb értéket vesz fel, mint az alapperiódusnál. Ez a jelenség okozza az "F0 a felső limiten" típusú hibákat. Megoldási javaslatunk a problémára a "halszálka" módszer, a szkeleton függvény alkalmazása. Egy függvény szkeletonja a függvény értékét veszi fel annak lokális szélső értékeinél és nullát egyébként. Itt a céljainknak a lokális szélső érték szigorú és nem szigorú definíciói közötti átmenet felel meg.

Definíció: *f* valós függvénynek lokális szélső értéke van *x*-ben, ha *x*-ben nem szigorúan monoton és nem sík.

Definíció: $g = skeleton(f)$ akkor és csak akkor

$$g(x) = \begin{cases} f(x) & \text{ha } f\text{-nek lokális szélső értéke van } x\text{-ben} \\ 0 & \text{egyébként} \end{cases} \quad (3)$$

A mesterséges lejtés ellenére a tisztán zöngés hangok elhalkuló végein az ACF hajlamos a tényleges alapperiódus idő többszöröseinél egyre növekvő csúcsokat mutatni, amint az a 2. ábrán látható.



2. ábra: Egy magánhangzó elhalkuló vége és annak autokorrelációja.

Ez a jelenség csak olyankor fordulhat elő, ha az ACF a periódus időnél 1-hez közeli vagy afölötti értéket vesz fel. Ezért a probléma megoldására egy preferencia szint bevezetését javasoljuk. Az algoritmus válassza az első csúcsot, ami a preferencia szintet meghaladja. Ha ilyen nincs, akkor a legmagasabb csúcsot. Mi tapasztalati alapon 0.75-öt használtunk preferencia szintként.

Összegezve a basic extractor algoritmusunk lépései a korrekt végrehajtási sorrendben a következők:

Step 1: Az ACF kiszámítása (2) szerint.

Step 2: Szálkásítás:

$$sr_t(\tau) = skeleton(r_t(\tau))$$

Step 3: A keresési tartomány korlátozása (limited skeleton):

$$\text{Legyen } [F0_{\min}; F0_{\max}] \text{ a keresési intervallum,}$$

$$srl_t(\tau) = \begin{cases} -0.5 & \text{ha } \tau < 1/F0_{\max} \\ sr_t(\tau) & \text{ha } 1/F0_{\max} \le \tau \le 1/F0_{\min} \\ -0.5 & \text{ha } \tau > 1/F0_{\min} \end{cases} \tag{4}$$

Step 4: Mesterséges lejtés:

$$srl_t^{biased}(\tau) = (1 - gr \cdot \tau) \cdot srl_t(\tau); \quad \text{ahol gr=1.75} \tag{5}$$

Step 5: F0 becslés.

Step 5/A: Preferencia szint alkalmazása:

$$\tau^* = \min\{\tau : srl_t^{biased}(\tau) \ge 0.75\} \tag{6}$$

Step 5/B: Ha 5/A sikertelen, válasszuk a legmagasabb csúcsot:

$$\tau^* = \arg\max_{\tau}\{srl_t^{biased}(\tau)\} \tag{7}$$

és ekkor az alapfrekvencia:

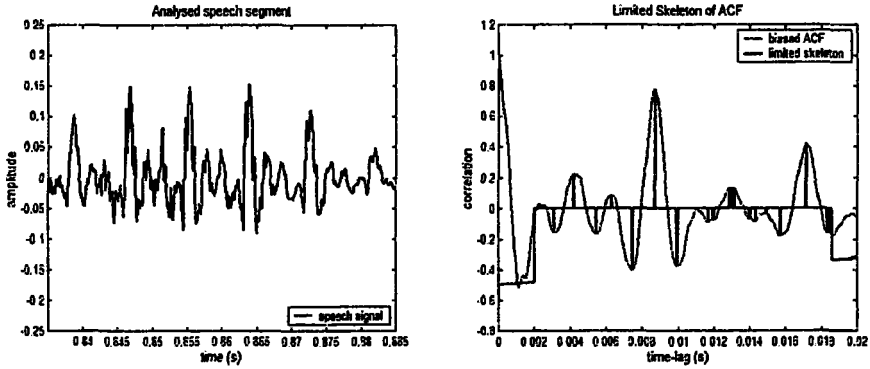
$$F0^* = \frac{1}{\tau^*}. \tag{8}$$

Step 6: A V/UV döntési paraméter:

$$rm_t = srl_t(\tau^*) \tag{9}$$

az "egyenes" (unbiased) korlátozott (limited) szkeletonból.

A 3. ábra mutatja az algoritmus működését.



3. ábra: Az *srl* (limited skeleton) maximuma mutatja a beszéd ablak alapperiódusát.

4. Zöngés-zöngétlen megkülönböztetés

Zöngésség detektorunk rm_t paramétert (9) használja döntése meghozatalában, valamint a jel energia logaritmusát:

$$p_t = 10 \cdot \log_{10}\left(\frac{1}{W}\int_{t-W/2}^{t+W/2} x(u)^2\,du\right) \quad \text{(decibel)} \tag{10}$$

A definícióból következik, hogy a maximális amplitúdójú négyszögjelre $p_t = 0\,\text{dB}$.

Ezek után a VDA egyszerűen összehasonlítja a paramétereket egy-egy küszöbbel. A zöngésség indikátor függvény pedig:

$$voicing(t) = \begin{cases} 1 & \text{ha } (rm_t > rmth)\,\&\,(p_t > pth) \\ 0 & \text{minden más esetben} \end{cases} \tag{11}$$

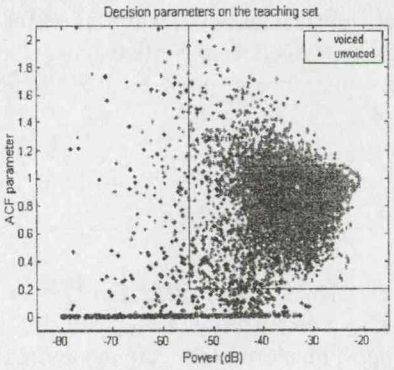
Ahol *rmth* és *pth* a küszöbök.

A kulcskérdés a továbbiakban a küszöbök optimális megválasztása. A hangolási folyamatot egybe kötöttük a döntési hibaarány kiértékelésével. A kiértékelésre szolgáló adatbázist két részre osztottuk: az egyik felén a betanítást, a másik felén az ellenőrzést végezzük. Tanításkor a küszöböket optimáljuk az adatbázis első felén, a másik felén pedig ellenőrizzük a VDA-t az optimált küszöbökkel. Természetesen az adatbázis két fele nem tartalmazhat közös részt, ez meghamisítaná a kiértékelést. A tanító és a teszt halmazba vegyesen tettük a női és férfi beszéd felvételeket, hogy az optimalizáció lehető legnagyobb beszélőfüggetlenséget biztosítsa.

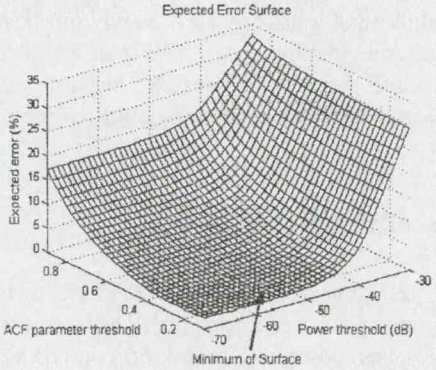
A döntési paraméterek kinyerése a teszt során W=32 ms ablakhosszal történt. Az $F0$ keresési tartomány 55 és 480 Hz között volt. A 4/a. ábra mutatja a paraméterek eloszlását a tanító halmazon. A világos pontok jelölik a zöngés, a sötétek a zöngétlen

szakaszokból származó paraméter párokat. A köztük haladó egyenes vonalak a kétküszöbös döntési módszert (11) ábrázolják. A vonalakon túlra tévedt sötét és világos pöttyök mutatják, hogy ez a módszer sem tökéletes.

A kétváltozós várható hibaarány felület az eloszlásokból származik. A felület értéke az *(x,y)* pontban azt jelenti, hogy *rmth=x* és *pth=y* küszöböket választva ennyi a V/UV tévesztés aránya a tanító halmazon. A felület mélypontja jelöli az optimális küszöböket. A 4/b ábrán látható a várható hibaarány felület.



4/a ábra: A döntési paraméterek eloszlása.



4/b ábra: Várható hibaarány felület.

Az optimált küszöbök: $pth = -55.2dB$ és $rmth = 0.23$. A hibafelület értéke ebben a pontban 1.95%. A kapott küszöböket teszteltük az adatbázis másik felén és a V/UV tévesztési arány **2.13%.** Ezt a hibaszázalékot, mint végeredményt tekinthetjük, ez az algoritmusunk teljesítménye.

5. Összegzés

Áttekintve az algoritmusunkat úgy látjuk, három jó részmegoldás játszott kulcsszerepet a 2.13%-os hibaarány elérésében. Az első az alul-áteresztő szűrés kombinálása a center clip-pel, a másik szkeleton függvény használata a basic extractor-ban, a harmadik pedig a jel energia figyelembe vétele a zöngésség meghatározásban. A jel energia sokkal jobban jelzi a zöngét, ha azt az előfeldolgozó után mérjük, mint ha az eredeti beszéden. Az algoritmus precíz megfogalmazása és a korrekt végrehajtási sorrend szintén lényeges.

Algoritmusunk implementálható valós idejű alkalmazásban is. Ekkor az algoritmusból fakadó (nem kiküszöbölhető) késés elsősorban az ablakszélességtől függ. 32 ms-os ablakot használtunk, ez 16 ms késést okoz. Ehhez még a burkoló számítás és az alul áteresztő szűrés legfeljebb 5 ms-ot tesz hozzá. A közeli jövőben elkészítünk egy PC-n futó valós idejű demo alkalmazást.

6. A kiértékelés adatbázisa

Algoritmusunkat a Fundamental Frequency Determination Algorithm Evaluation Database (FDA) elnevezésű beszéd adatbázison ellenőriztük. Ezt a University of Edinburgh egyetem Centre for Speech Technology Research intézetében készítették. A szerzője Paul Christopher Bagshaw. Az adatbázis letölthető az Internetről, az URL: http://www.cstr.ed.ac.uk/~pcb/fda-eval.tar.gz. 7 percnyi beszédet tartalmaz. 50 angol mondat, mindegyik egy férfi és egy női beszélő elmondásában. A tejes idő 37%-ában zöngés szegmentumok és 63%-ban zönge nélküliek (zöngétlen mássalhangzó és beszédszünet együtt). A beszédet laryngográf jellel szinkronban vették fel. Ez alapján címkézték a zöngés és zönge nélküli szegmentumokat.

Köszönetnyilvánítás

A szerző szeretné köszönetét kifejezni témavezetőjének, Dr. Takács Györgynek a iránymutatásáért és segítségéért, valamint a Pázmány Péter Katolikus Egyetem Információs Technológiai Kar doktori iskolája vezetőinek a bizalomért és a támogatásért.

Bibliográfia

1. B. S. Atal and L. R. Rabiner "A Pattern Recognition Approach to Voiced—Unvoiced—Silence Classification with Applications to Speech Recognition" IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, pp. 201—212 (1976)
2. B. S. Atal and L. R. Rabiner: "Voiced-unvoice decision without pitch detection" J Acoust. Soc. Am., Vol. 58 (1975)
3. P. C. Bagshaw Automatic prosodic analysis for computer aided pronunciation teaching PhD Thesis, Univ. Edinburgh (1994)
4. P. C. Bagshaw, S. M. Hiller and M. A. Jack "Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching" Proc. 3rd European Conf. on Speech Comm. and Technology, Vol. 2, pp. 1003—1006, Berlin (1993)
5. A. de Cheveigné and H. Kawahara: "YIN, a fundamental frequency estimator for speech and music" J Acoust. Soc. Am., Vol. 111, Apr (2002)
6. J. R. Deller, J. H. L. Hansen and J. G. Proakis Discrete-Time Processing of Speech Signals, Macmillan, New York (1993)
7. W. A. Hess Pitch Determination of Speech Signals, Berlin, Springer-Verlag (1983)
8. L. R. Rabiner "Evaluation of a statistical approach to voiced-unvoiced-silence analysis for telephone quality speech" Bell Syst. Tech. J., Vol. 56, pp. 455—482 (1977)
9. L. R. Rabiner "On the Use of Autocorrelation Analysis for Pitch Detection" IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-25, pp. 24—33 (1977)
10. L. R. Rabiner, M. J. Cheng, A. E. Rosenberg and C. A. McGonegal "A Comparative Performance Study of Several Pitch Detection Algorithms" IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-24, pp. 399—418 (1976)
11. L. R. Rabiner and R. W. Schafer Digital Processing of Speech Signals, Prentice Hall, Engelwood Cliffs NJ (1978)