Automatikus hangmagasság-korrekciós rendszer létrehozása frekvenciatartományban

Implementation of an automatic pitch correction system in the frequency domain

FIRTHA GERGELY
Budapesti Műszaki és Gazdaságtudományi Egyetem, Híradástechnikai Tanszék
gfirtha@gmail.com

Beérkezett: 2010.05.17., elfogadva: 2011.02.03

Kivonat – A cikk egy pusztán frekvenciatartománybeli módszereken alapuló hangmagasság korrekciós rendszer létrehozását ismerteti. Bemutatja, milyen részfeladatok megoldása szükséges a cél eléréséhez, ismerteti ezen feladatok lehetséges megoldásait a frekvenciatartományban, és kitér az így létrehozott funkcionális blokkokból álló rendszer működésére.

Abstract – The article delas with the implementation of an automatic pitch correction system, based only on the processing of signal in the frequency domain. It reveals what steps, tasks are needed in order to achieve the final aim, presents the possible solutions of these tasks in the frequency domain, and shows the operation of the complete system, built up from the implementation of these methods.

1. Bevezetés

Hangstúdiókban zenei felvétel készítése során, főleg éneksávok esetén jellemző, hogy a hangmagasság utólagos pontosítására van szükség. Míg a hangmagasság tetszőleges módosítása szűk korlátok között analóg módon is megoldott volt, addig a harmonikus jelek alapharmonikusának meghatározására a digitális jelfeldolgozás adta lehetőségek szolgáltak megoldásul, így lehetővé téve egy automatikusan működő hangmagasságkorrigáló rendszer létrehozását.

A hangmagasság korrekció folyamata három nagy részfeladatra bontható: elsőként a bemenő jelfolyam hangmagasságát kell megállapítanunk, melyből ezután meghatározható, mennyit kell azon változtatni, hogy a harmonikus hang valamely zenei skálába tartozzon, tehát a hangmagasság módosító algoritmus vezérlőjele előállítható. A feladatok egyike sem triviális megoldású, mind a hangmagasság felismerésre, mind a módosításra születtek idő- és frekvenciatartománybeli megoldások is. Cikkemben ezek közül a frekvenciatartománybeli eljárások lehetőségeit és felmerülő problémáit mutatom be, néhány esetben a problémák egy lehetséges megoldására is rámutatva.

2. A hangmagasság detektálása

A hangmagasság érzékelése összetett pszichofizikai folyamat, az érzékelt magasság elsősorban érzeti jellemző, nem pedig konkrét fizikai mennyiség: függ a hang intenzitásától, harmonikustartalmától, időtartamától, és a hullámforma egyéb fizikai jellemzőitől. A feladat szempontjából kielégítő eredményt ad, ha a hangmagasságot meghatározó tényezőként a harmonikus hang f_0 alapfrekvenciáját tekintjük, figyelembe véve a hallás azon fontos tulajdonságát, hogy az érzékelés alacsonyabb frekvenciákon jóval pontosabb, mint magasabb hangok esetén.

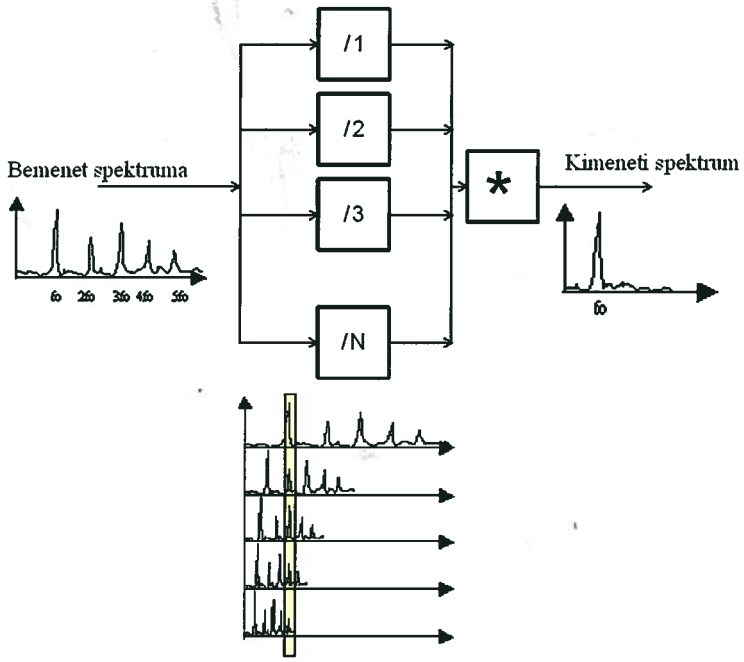
A frekvenciatartománybeli módszerek abból a tényből indulnak ki, hogy harmonikus hangok esetén a spektrum főként az alapharmonikus egészszámú többszörösein tartalmaz összetevőket, épp ezért a következő algoritmusok monofonikus hangokra nem működnek. A feladat természetesen nem vezethető vissza a triviálisnak tűnő spektrális maximumhely-keresésre, hiszen az esetek jelentős részében nem az alapharmonikus hordozza a legnagyobb energiát.

Az alapharmonikus keresés legegyszerűbb módja az Harmonic Spectrum Product módszer, amelynek az alapötlete a következő [1]: legalább N felharmonikust tartalmazó hang spektrumát N-ed részére összenyomva (ami újramintavételezéssel egyszerűen megvalósítható) az N-edik felharmonikus épp az alapharmonikus eredeti helyére – alapharmonikus frekvenciabinjére – kerül. Ezután az eredeti és az összenyomott spektrumokat összeszorozva ideális esetben az alapharmonikuson kívüli összetevők elhanyagolhatóvá válnak, és az alapharmonikus egyszerű maximumhely kereséssel meghatározható (1. ábra).

A módszer fő előnye a gyorsasága, kis számításigénye, valamint az, hogy mind additív, mind multiplikatív zajra érzéketlen. Hátránya azonban a lehetséges oktávhiba, azaz nem az alapharmonikusra, hanem valamely felharmonikusára való döntés, nagyon sok felharmonikust tartalmazó jelek estén. A probléma kiküszöbölhető úgy, ha a spektrumban az alaphangként detektált frekvencia felénél található komponenst vizsgáljuk. Nagyobb gondot jelent azonban, hogy a módszer pontossága az egész spektrumon azonos: a hiba maximálisan két szomszédos frekvenciabin közötti távolság fele.

Más megközelítése a problémának a kepsztrális analízis [2], melynek alapja az a tény, hogy egy több felharmonikust tartalmazó jel spektruma maga is részben periodikusnak tekinthető, és ez a periodicitás egy újbóli Fourier-transzformációval detektálható. Kepsztrális analízis során az énekhangot a hangszálak által keltett tisztán harmonikus rezgés a vokális traktus hangalakító szervei által szűrt változatának tekintjük, azaz felírható a gerjesztőjel és a vokális traktust reprezentáló szűrő spektrumainak szorzataként (ez az emberi hangképzés forrás-szűrő modellje):

$$Y(\omega) = |X(j\omega)||X(j\omega)| \,. \tag{1}$$



1. ábra: A HPS algoritmus működése – Operation of HPS algorithm

Az egyenlet mindkét oldalának logaritmusát véve a szorzás összeadássá alakul, amelyet ezután tagonként inverz Fourier-transzformálhatunk:

$$C(\omega) = \mathfrak{F}^{-1}\{\log|Y(\omega)|\} = \mathfrak{F}^{-1}\{|X(j\omega)|\} * \mathfrak{F}^{-1}\{|X(j\omega)|\} . \quad (2)$$

Az egyenlet bal oldala a hang kepsztruma, amelyben a 0 ms környezete a szűrő jellemzője, míg a gerjesztésből származó csúcsok kb. 5 ms után jelennek meg. Ezután a kepsztrumban való csúcskereséssel az eredeti hang alapfrekvenciája könnyen meghatározható (2. ábra).

A módszer előnye, hogy a felbontása alacsonyabb frekvenciákon pontosabb, azaz jobban illeszkedik az emberi hallásmodellhez. Hátránya azonban hogy magasabb hangokon a szűrő és a gerjesztés kepsztruma már nehezebben különválasztható, így pontosan leginkább mélyebb hangokon működik, emellett a HPS-nél jelentősen zajérzékenyebb.

A kepsztrum frekvenciatartományba való visszaskálázásával (az N. kepsztrum bin az f_s/N frekvenciabinbe kerül), majd a HPS által szolgáltatott eredménnyel való összeszorzásával a HPS algoritmus oktávhibája és a kepsztrum analízis zavarérzékenysége jelentősen csökenthető, azonban ezzel a frekvenciafelbontás ismét az egész frekvenciatartományban állandó lesz, azaz épp a kepsztrális analízis előnyét vesztjük el.

Vizsgáljuk meg, általános esetben mekkora a frekvenciafelbontás! Ahhoz, hogy a pillanatnyi hangmagasság jól követhető legyen, azaz a vizsgált időtartamon belül a spektrum szerkezetének változása minimális legyen általában az énekdallam kb. 50 ms hosszú (2048 minta, f_s = 44,1 kHz) blokkjait vizsgáljuk. Az FFT felbontása ekkor, mint ismeretes:

$$\Delta f = \frac{1}{t_{\text{mintavétel}}} = \frac{f_s}{N} = 21{,}5 \text{ Hz} . \quad (3)$$

A frekvenciahiba tehát maximálisan ennek a fele lehet. Az átlagos énekhangok alapfrekvenciája a 200–500 Hz tartományban mozog. 500 Hz környezetében két szomszédos félhang frekvenciája 30 Hz. Egyértelmű, hogy ebben a tartományban 11 Hz pontatlanság nem megengedhető, a felbontás növelése szükséges. Erre lehetséges megoldás az FFT pontszámának növelése pl. nullmintákkal sűrítéssel, majd két szomszédos frekvenciabin alapján becsülni kell az eredeti spektrumcsúcs helyét. Erre a következő eljárást dolgoztam ki (3. ábra):

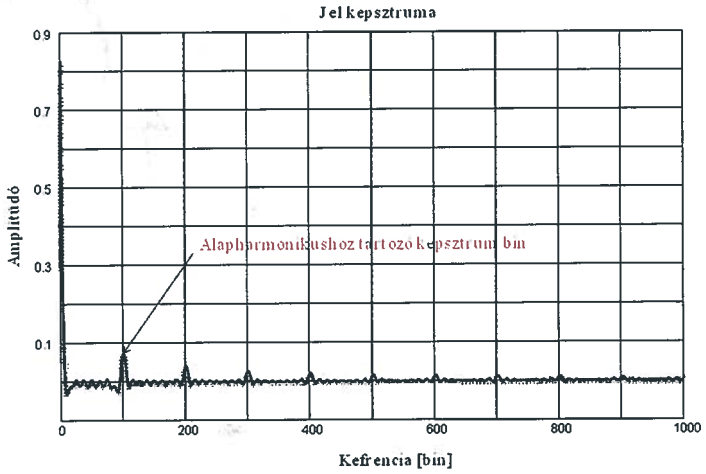
- Illesszünk parabolát, vagy sinc függvényt az eredetileg kapott k_{max} maximális frekvenciacsúcsra és két szomszédjára, úgy, hogy az illesztett függvény a szomszédos frekvenciabinen zérus értéket vegyen fel
- Az illesztett függvényeket összegezzük, majd vegyük az így kapott két frekvenciacsúcs helyének számtani közepét: $k_{\text{átlag}}$
- Mivel az így számított középérték nem járja be a teljes frekvenciabinek közötti szakasz felét, ezért mozgását terjesszük ki $k_{max} \pm \Delta f/2$-re, az így kapott frekvenciabin a becsült alapfrekvencia.

Az így kapott módszerrel a frekvenciafelbontás pontossága 4096 mintányi blokkok esetén 0,1 Hz körül van, amely már bőven elegendő pontosság az eredeti feladat megoldására a 10,75 Hz pontossággal szemben, amely 4096 minta esetén a DFT frekvenciafelbontása. A 4. ábrán az algoritmus eredménye látható állandóan növekedő frekvenciájú vizsgálójelre, becsléssel és becslés nélkül.

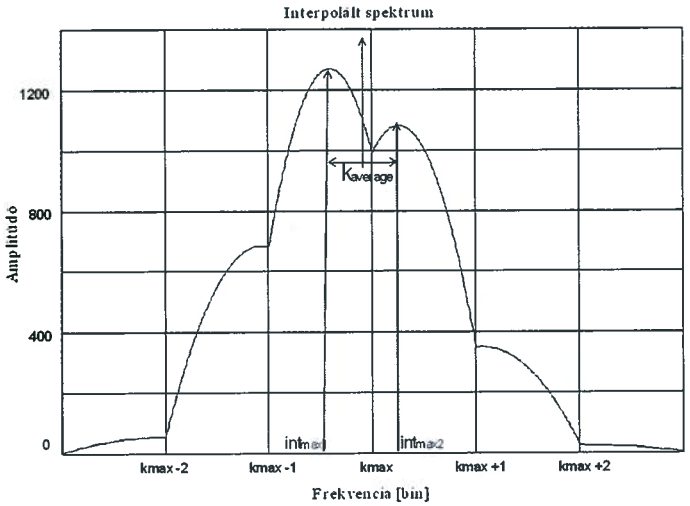
Énekdallamok esetén fontos, hogy az algoritmus különbséget tegyen a zöngés, és a nem harmonikus zöngétlen hangok között. A zöngés-zöngétlen különválasztására lehetőséget nyújt a spektrum tömegközéppontjának számítása, mivel a zöngétlen hangok spektrális tömegközéppontja tapasztalat alapján sokkal nagyobb frekvencián van, mint a harmonikus jeleké. Ez alapján egy egyszerű küszöbérték állítással a zöngés és zöngétlen hangok egyszerűen különválaszthatók.

3. A hangmagasság módosítása

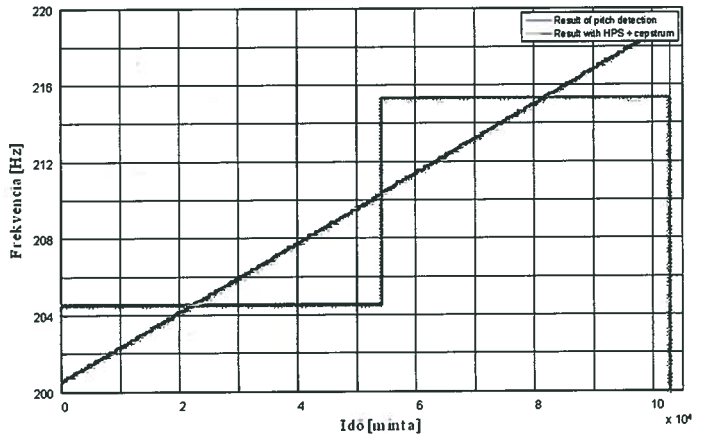
A hangmagasság ismeretében a következő feladat a lehető legkisebb minőségi romlás mellett a hangmagasság módosítása. Erre napjainkban a legszofisztikáltabb eljárás a phase vocoder



2. ábra: Harmonikus jel kepsztruma – Cepstrum of harmonic input signal



3. ábra: A detektált alapfrekvencia környezete parabolikus interpoláció után – The detected pitch frequency bin after interpolation



4. ábra: A hangmagasság detektálás eredménye lineárisan növekedő frekvenciájú bemenő jelre interpolációval és interpoláció nélkül – Result of pitch detection for sweep sinus input signal with and without interpolation

technika [3, 4], amelyet a kereskedelemben is kapható hangmagasság módosító szoftverek – így az Antares Auto-Tune – is alkalmaz.

Az eljárás során a tetszőleges hosszú folytonos bemenő jelből tetszőleges kezdőpontú N hosszú blokkját dolgozzuk fel, úgy hogy az idő abszolút origója a 0. idő bin, azaz a feldolgozott blokk helye az időtengelyen ismert. Ez elméletben a jel egy olyan ablakfüggvénnyel való szorzásával, amely csak a vizsgált blokk helyén nem nulla, majd a teljes jel Fourier-transzformációjával érhető el. Ez a rövid idejű Fourier-transzformáció (STFT). A gyakorlatban azonban a jelből közvetlenül az aktuálisan vizsgált N hosszú blokk DFT-je számítható, amely során az időtengely origója a vizsgált blokk kezdete: ez a különbség a feldolgozás során hibás fázisértelmezéshez vezet. A blokkok átlapolódnak, kezdőpontjuk egymástól R_a távolságban van, ez az ún. ugrásméret. 50%-os átlapolódás mellett pl. $R_a = N/2$. Ekkor levezethető, hogy az s-edik blokk esetén a jel STFT-je és DFT-je, illetve a fázisuk között az alábbi kapcsolat van:

$$X(sR_a,k))_{STFT} = W_N^{sR_a k}\tilde{X}(sR_a,k)_{DFT}, \tag{4}$$

$$\varphi(sR_a,k)_{STFT} = \tilde{\varphi}(sR_a,k)_{DFT} - \frac{2\pi k}{N}sR_a\,, \tag{5}$$

ahol $W_N = e^{j2\pi/N}$.

A feldolgozás folyamán tehát a fázist folyamatosan korrigálni kell. Hogy a fázis ne tartson végtelenbe célszerűen a $]-\pi;\pi]$ tartományba transzformálni minden blokk feldolgozása során.

A phase vocoder-hez kapcsolódó talán legfontosabb fogalom a pillanatnyi frekvencia. Ez az egyes frekvenciákon való fázisváltozás egy mintavételi idő alatt, azaz a fázis első deriváltja. Értelemszerűen, ha két egymást követő blokk kezdetén a fázis ismert, akkor a fáziskülönbség és a blokkok közötti időbeli távolság hányadosa épp az adott k. pillanatnyi frekvencia:

$$\mathrm{d}\varphi = \frac{\Delta\varphi_{STFT}}{R_a} = \frac{\Delta\tilde{\varphi}_{STFT}}{R_a} - \frac{2\pi k}{N}\,, \tag{6}$$

ahol a korrekciós tag az (5) egyenletet az s-edik és $(s+1)$-edik blokkra felírva egyszerűen kijön.

A pillanatnyi frekvenciák segítségével a hangmagasság módosítás egyszerűen végrehajtható, akár a szinusz összetevők közvetlen szintetizálásával a következő módon:

- Analízis során kiszámoljuk a fázisváltozást egy minta alatt:

$$\mathrm{d}\varphi(k) = \frac{\Delta\varphi(k)}{R_a}\,, \tag{7}$$

  hasonlóan az egyes összetevők egy minta alatti amplitúdóváltozását:

$$\mathrm{d}A(k) = \frac{A((s+1)R_a,k) - A(sR_a,k)}{R_a}\,. \tag{8}$$

- A fázisváltozást megszorozzuk a transzpozíciós tényezővel (*transpo*), így végrehajtva a frekvenciamódosítást, és integráljuk a módosított fázist az időtartományon úgy, hogy a blokkon belül az adott $X(n,k)$ minta értéke:

$$\tilde{\varphi}(n+1,k) = \tilde{\varphi}(n,k) + transpo\cdot\mathrm{d}\varphi(k) \tag{9}$$

$$A(n+1,k) = A(n,k) + \mathrm{d}A(k) \tag{10}$$

$$X(n,k) = A(n,k)\sin(\tilde{\varphi}(n,k)) \tag{11}$$

- Az algoritmust minden frekvenciára elvégezve tehát közvetlenül, inverz Fourier-transzformáció nélkül újraszintetizálhatjuk az immár frekvenciában módosított összetevőket, végezve el a feladatot: Az algoritmus eredménye az 5. ábrán látható a hangmagasság oktávnyi emelése mellett.

Természetesen mind a hangmagasság detektálás, mind a hangmagasság módosítás elvégezhető időtartományban is. Módosításra időtartományban gyakori módszer pl. SOLA, vagy PSOLA algoritmus, azonban ezek a módszerek minősége messze elmarad a phase vocoder által produkált minőség mellett, amellyel a bemutatotton kívül más módon is elérhető a hangmagasság módosítása. A hangmagasság detektálására jónéhány időtartománybeli megoldás létezik – nullátmenet vizsgálat, csúcsérték-számlálás, autokorrelációs algoritmusok, YIN becslés alkalmazása csak néhányat említve –, sőt ún. fázistérbeli vizsgálattal és különböző wavelet transzformációkkal is kinyerhető a keresett alapfrekvencia érték. Ezen módszerek hatékonyságban gyakran felérnek a frekvenciatartománybeli módszerekkel, sok esetben gyorsabbak is, azonban a phase vocoder eleve frekvenciatartományban működik, így az FFT számítás nem jelent számításigény többletet. Az alaphang felismerő algoritmusok hatékonyságát napjainkan legjobban a MIDOMI és Shazam programok mutatják, amelyek akár lényeges időbeli és frekvenciabeli tévedések mellet is képesek dúdolásból, éneklésből, fütyülésből az adatbázisban szereplő dallamokra ráismerni. Ezen programok működésének alapjai, algoritmusaik azonban természetesen nem ismertek.

4. A teljes rendszer felépítése és működése

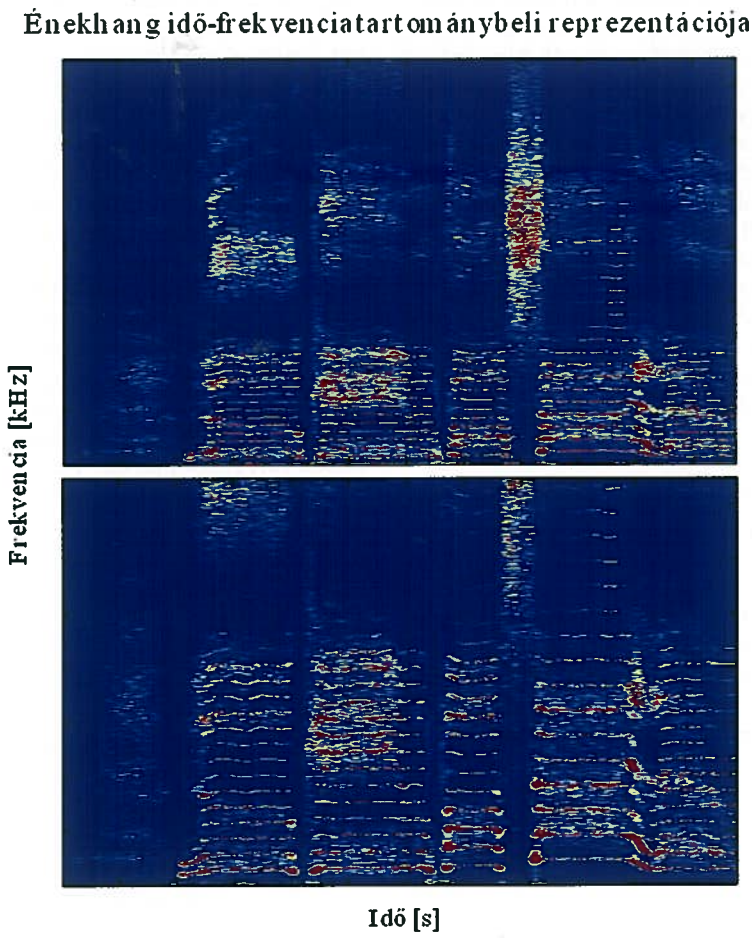
Hatékony módszerrel a hangmagasság detektálására és módosítására már csak a szükséges módosítás mértékének meghatározása szükséges. Ez legegyszerűbben a bemenő jel detektált alapfrekvenciájához legközelebb eső zenei hang megkeresésével, majd a bemenő frekvencia erre való igazításával érhető el: A zenei skálába tartozó frekvenciák:

$$f_n = 55\cdot(\sqrt[12]{2})^n\,, \text{ahol } n\in N. \tag{12}$$

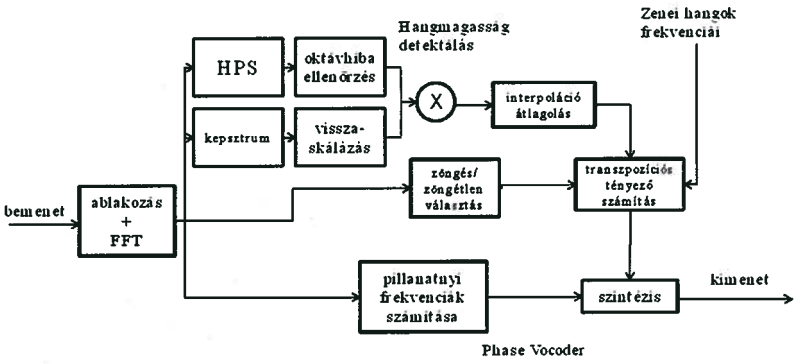
Megvizsgálva, hogy a detektált frekvencia melyik f_n frekvenciához van legközelebb, majd az arányukat kiszámítva a transzpozíciós tényező így egyszerűen meghatározható.

A teljes hangmagasság korrekciós rendszer felépítése a 6. ábrán látható: A bemenő jelet átlapolódó blokkonként ablakozzuk és Fourier-transzformáljuk. Meghatározzuk az alapharmonikus frekvenciáját – a pontos eredmény érdekében frekvenciatartománybeli interpoláció segítségével – majd megvizsgáljuk, az mely zenei hanghoz van legközelebb. Ha a bemenő hang zöngétlen mássalhangzó, nincs szükség hangmagasság módosítására: a transzpozíciós tényező értéke egységnyi. A kimenetet eközben a phase vocoder algoritmus segítségével folyamatosan szintetizáljuk a pillanatnyi frekvenciák segítségével, amelyeket szükség esetén folyamatosan módosíthatunk.

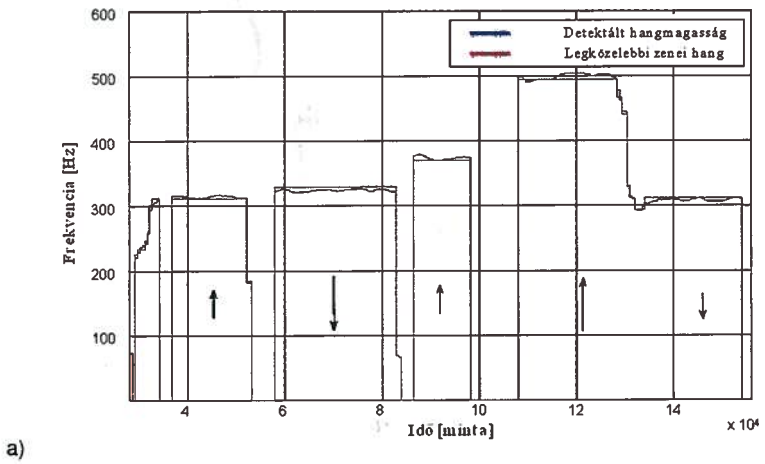
A rendszer működésének vizsgálata a 7.a) és b) ábrákon látható: Női énekhangot stúdióprogram beágyazott hangmagasság módosító algoritmusával hamissá téve – a módosítás „iránya" az ábrán nyilakkal jelölve – majd rendszer hangmagasság felismerő algoritmusán átfuttatva a detektált hangmagasság a 7.a) ábrán kék színnel, a hozzá legközelebb eső zenei hang frekvenciája pedig szaggatott piros vonallal látható. A frekvencia egyes időpillanatokban zérusértékű, ezeken a helyeken a rendszer zöngétlen mássalhangzót, vagy csendet detektált. A 7.b) ábrán a teljes rendszer kimenete látható a hangmagasság detektáláson ismét átfuttatva: Látható, hogy a rendszer a kitűzött célt tökéletesen végrehajtotta, a szintetizált énekben már minden hang zenei skálába esik. A bemenetet és kimenetet meghallgatva, és összehasonlítva ez be is bizonyosodik, a jól hallható hamis hangokat a rendszer kijavította.



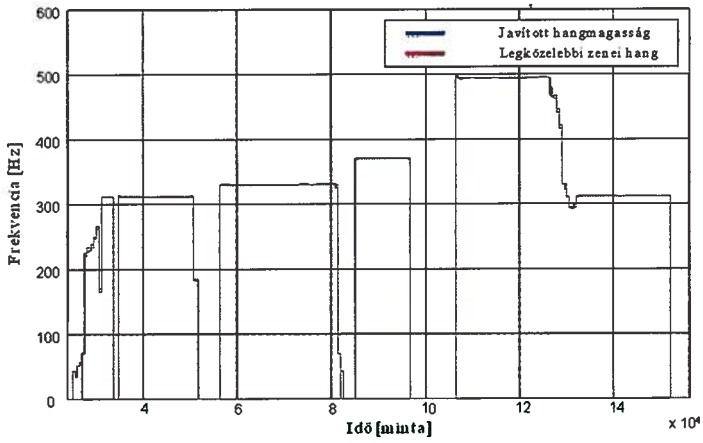
5. ábra: Az eredeti énekdallam és frekvenciakétszerezett énekdallam spektrumának a változása az időben – The spectrum of original and frequency doubled tune versus time



6. ábra: A teljes hangmagasság korrekciós rendszer blokkvázlata – Block diagram of the complete pitch correction system



a)



b)

7. ábra: Hamis női énekdallam automatikus hangmagasság javítás a) előtt és b) után - Detected pitch of false tune a) before and b) after automatic pitch correction

5. Összefoglalás

A cél, tehát egy pusztán frekvenciatartománybeli feldolgozáson alapuló automatikusan működő hangmagasság korrekció a bemutatott módszerekkel elérhető. Az DFT felbontása interpolációval jelentősen növelhető, akár tized Hertz pontosságig is, amely már felül is múlja a feladathoz szükséges felbontást. A phase vocoder eljárás hallható minőségi romlás nélkül képes a hangmagasságot módosítani, épp ezért a kereskedelmi forgalomban kapható szoftverek ezt az algoritmust használják. A bemutatott rendszer sok továbbfejlesztési lehetőséget rejt magában: a zöngés-zöngétlen különválasztás az esetek nagy részében helyesen működik, azonban a H hangot, amelynek spektrális súlypontja az alacsonyabb frekvenciákon helyezkedik el, nem képes detektálni. Szintén problémát okoz a vezérlőjel előállítás statikus jellege, amely miatt a rendszer a hajlításokat és a negyedhangot meghaladó amplitúdójú vibrátókat nem képes kezelni. Működéséből eredően a HPS algoritmus képes lenne polifonikus hangok összetevőinek alapfrekvenciájának felismerésére is. Épp emiatt frekvenciafüggő vezérlőjellel lehetséges lenne a rendszer működését akár polifonikus hangokra is kiterjeszteni.

Hivatkozások

[1] G. Middleton. Pitch detection algorithms, connexions module. `http://cnx.org/content/m11714/latest/`.

[2] V.B. Osdol. Cepstrum, use and application of the cepstrum domain for speech analysis, connexions module. `http://cnx.org/content/m12469/latest/`.

[3] U. Zölzer. *Digital Audio Effects*. JohnWiley & Sons Ltd, 2002.

[4] N. Bernardini, A.D. Götzen, and D. Arfib. Traditional implementations of a phase-vocoder: The tricks of the trade. In *Proceedings of the COST G-6 Conference on Digital Audio Effects (DAFX-00)*, 2000.