Automatikus hangmagasság-korrekciós rendszer létrehozása

FIRTHA GERGELY

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

Kulcsszavak: hangmagasság-detektálás, hangmagasság-módosítás, kepsztrális analízis, phase vocoder

A cikk egy automatikus hangmagasság-korrekciós rendszer felépítését ismerteti. Bemutatja, milyen részfeladatok megoldása szükséges a cél eléréséhez, ismerteti ezen feladatok lehetséges megoldásait mind az idő-, mind a frekvenciatartományban és kitér az így létrehozott funkcionális blokkokból álló teljes rendszer működésére.

1. Bevezetés

Hangstúdiókban zenei felvétel készítése során, főként éneksávok esetén, gyakran a hangmagasság utólagos módosítása, pontosítása szükséges. Hangmagasság-korrekció során az énekhangot a hozzá legközelebb eső zenei hanghoz igazítjuk, de azt teljes zenei hangközökkel változtatva akár egyetlen éneksáv alapján kórus is létrehozható. Ennek megvalósításához egyértelmű, hogy elsőként a bemenő hangmagasság minél pontosabb meghatározása szükséges, amelyből meghatározható, mennyit kell azon változtatni a kívánt cél eléréséhez. Ezután a hangmagasság módosítása következik minél kisebb érzékelhető minőségi romlás mellett. Ezek közül egyik sem triviális megoldású, megvalósítható mindkettő a frekvencia- és időtartományban is. A cikkben ezeket az eljárásokat foglaljuk össze, röviden bemutatva azok működésének elméleti alapjait, előnyeiket, hátrányaikat, kitérve egy lehetséges, jól működő rendszer implementálására.

2. A hangmagasság detektálása

A hangmagasság észlelése igen összetett pszichofizikai folyamat, nem egyetlen paramétertől függő fizikai mennyiség. Az észlelt hangmagasság természetesen legszorosabban a spektrális összetevők frekvenciájával van összefüggésben, de függ az észlelt jel intenzitásától, a harmonikusokban való gazdagságtól, a hang időtartamától is. Összességében a feladat szempontjából elegendőnek tekinthető a harmonikus jel alapfrekvenciájának meghatározása. Erre mind az időtartományban, mind a frekvenciatartományban számos megoldás született.

Az időtartományban működő hangmagasság-detektáló algoritmusok általános jellemzője, hogy az eljárás hatékonyságát a bemenő jel harmonikusgazdagsága rontja, tehát hatékonyan csak néhány felharmonikust tartalmazó hangra alkalmazható. Épp ezért előzőleg aluláteresztő szűrés szükséges, amely optimális esetben csak az alapharmonikus összetevőt hagyja meg a jelben. Ez természetesen nagyon kevés esetben teljesíthető.

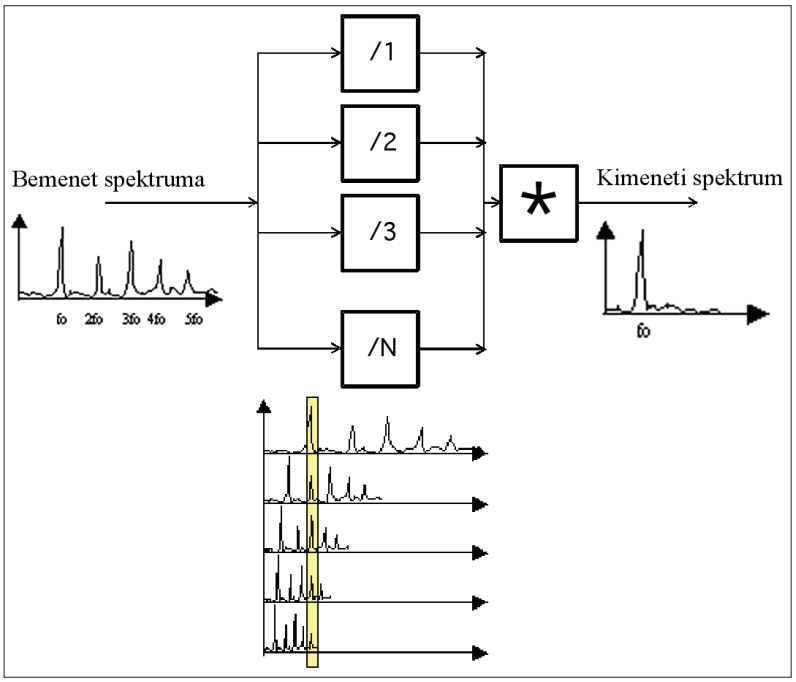
A hangmagasság detektálása az időtartományban legegyszerűbben nullátmenet-számlálással (zero crossing rate) történhet. Némileg szofisztikáltabb és manapság a legelterjedtebb időtartománybeli módszerek az autokorreláción alapuló algoritmusok (AMDF, ASMDF, MACF), melyek során a bementi jel egyes részei közötti hasonlóságot vizsgáljuk. Ennek során a jel valamely autokorrelációs függvényét számítjuk:

$$R_x(v) = \sum_{n=-\infty}^{\infty} x(n)x(n+v).$$

Periodikus $x(n)$ bemenő jel esetén az autokorrelációs függvény bizonyíthatóan periodikus és olyan v értékekre maximális, amely éppen $x(n)$ periódusideje. Ezáltal az eredeti függvény alapfrekvenciája $R_x(v)$ első maximumhelyével megkereshető. Problémát okoz azonban, hogy harmonikusokban gazdagabb jel esetén a korrelációs függvényben minden harmonikusnak megfelelően lokális maximumok jelennek meg. Ha az eredeti jel spektrumában nem az alapharmonikus hordozza a legnagyobb energiát, akkor az autokorrelációs függvényben nem a keresett alapfrekvenciához tartozik a függvény abszolút maximuma, így az algoritmus hibás eredményre vezet. Ez a probléma részben kiküszöbölhető az úgynevezett YIN-becslő alkalmazásával, ami a korrelációs szorzatmaximalizálással ellentétben a függvény részletei közötti különbségek minimumát keresi, így adva becslést az alapfrekvenciára.

Az időtartománybeli módszerek általános előnye a gyorsaság. A számításigénye olyan kicsi, hogy szinte minden valós idejű alkalmazásnál – így például a GSM beszédkódoló eljárások – a hangmagasság meghatározása az időtartományban történik.

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 triviális-



1. ábra A HPS algoritmus működése

nak tűnő spektrális maximumhely-keresésre, hiszen az esetek jelentős részében nem az alapharmonikus hordozza a legnagyobb energiát.

Az alapharmonikus keresés legegyszerűbb módja a Harmonic Spectrum Product (HPS) módszer, amelynek az alapötlete a következő: legalább N felharmonikust tartalmazó hang spektrumát N-ed részére összenyomva (ami újramintavételezéssel egyszerűen megvalósítható) az N. felharmonikus épp az alapharmonikus eredeti helyére – alapharmonikus „frekvencia-binjére" – 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. Nagyobb gondot jelent, hogy a módszer pontossága az egész spektrumon azonos: a hiba maximálisan két szomszédos frekvencia bin közötti távolság fele.

Másik megközelítése a problémának a kepsztrális analízis, 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):

$$Y(\omega) = |X(\mathrm{j}\omega)||H(\mathrm{j}\omega)|$$

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

$$C(\omega) = \mathcal{F}^{-1}\{\log|Y(\omega)|\} = \\ = \mathcal{F}^{-1}\{\log|X(\mathrm{j}\omega)|\} + \mathcal{F}^{-1}\{\log|H(\mathrm{j}\omega)|\}$$

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 HPS és kepsztrum módszerek együtt is alkalmazhatóak, ebben az esetben azonban a frekvenciafelbontás ismét romlik. Az FFT eredményeként kapott diszkrét spektrumban 50 ms hosszú blokkméretet alkalmazva a szomszédos frekvenciakomponensek közötti távolság akár 20 Hz fölött van, ami akár 10 Hz hibát jelenthet a detektálásban. Ez természetesen nem megengedhető, ezért a frekvenciafelbontás növelése szükséges, amelyre saját eljárás került kidolgozásra.

A felbontás növelésének alapja a frekvenciatartománybeli interpoláció parabolikus, vagy sinc függvényekkel, azaz a frekvenciacsúcsokra való parabolaillesztés úgy, hogy a parabolikus függvény értéke a szomszédos frekvenciakomponenseken nulla legyen. Ezek után az interpolált spektrum az így illesztett függvények összege, amelyben megkereshető a becsült eredeti spektrális csúcsérték. 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 és 44,1 kHz mintavételi frekvencia esetén a DFT frekvenciafelbontása. A *3. ábrán* az algoritmus eredménye látható állandóan növekedő frekvenciájú vizsgálójelre, becsléssel és becslés nélkül.

Az időtartománybeli és frekvenciatartománybeli módszereken kívül az utóbbi időben kezdenek elterjedni a különböző wavelet-transzformáción alapuló hangmagasságot meghatározó algoritmusok. Ekkor a jelet egyes anya-wavelet bázisfüggvények által kifeszített térbe transzformáljuk lineáris transzformációval. Az anya-waveleteket megfelelően megválasztva információ nyerhető például a jel nullátmeneteiről, amelyből a hangmagasság meghatározható.

Az alaphangi felismerő algoritmusok hatékonyságát napjainkan legjobban a MIDOMI és Shazam programok mutatják, amelyek akár lényeges időbeli és frekvencia-

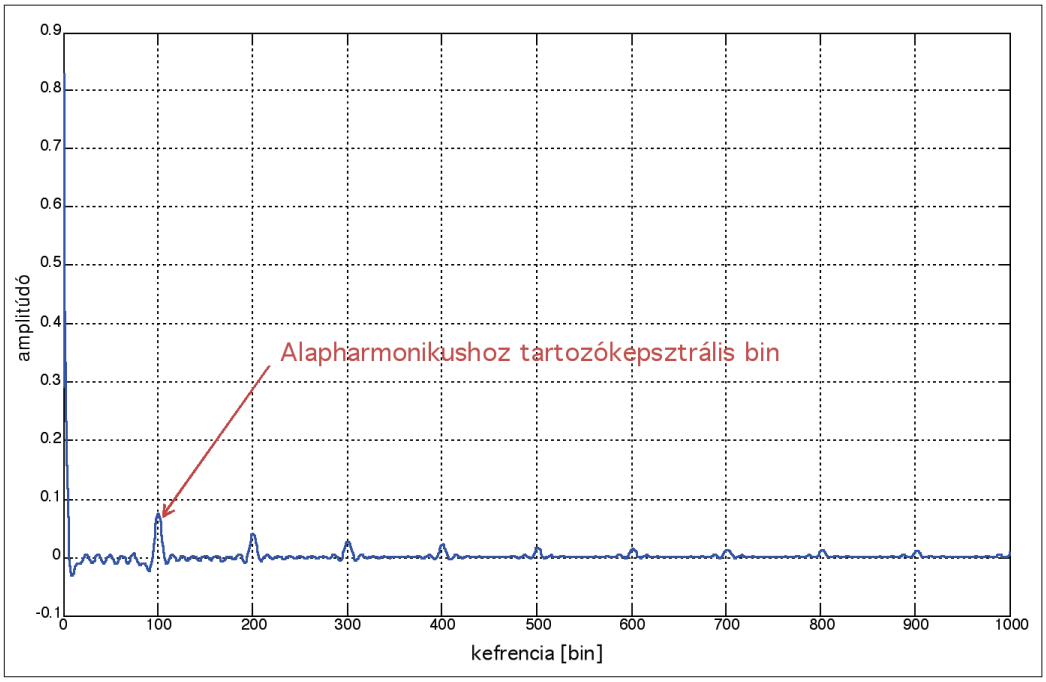
beli tévedések mellett is képesek dúdolásból, éneklésből, fütyülésből az adatbázisban szereplő dallamokra ráismerni. Ezen programok működésének alapjai, algoritmusaik azonban természetesen nem ismertek.

3. A hangmagasság módosítása

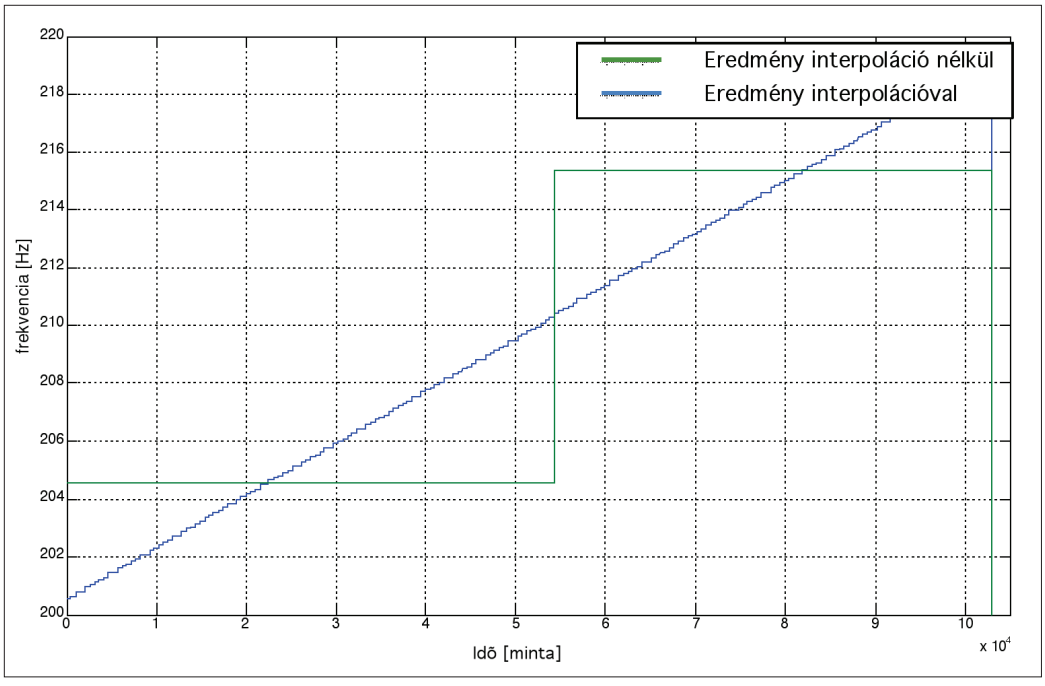
A hangmagasság ismeretében a következő feladat a lehető legkisebb minőségi romlás mellett a hangmagasság módosítása. Erre természetesen ismét mind a frekvenciatartományban, mind az időtartományban születtek eljárások.

Az időtartománybeli módszerek általánosan ugyanazon az elven működnek: a bemenő jel hosszának nyújtása változatlan hangmagasság mellett, majd a lejátszási sebességet növelve a hangmagasság megváltoztatása. Erre a két legelterjedtebb módszerek a *Synchronous OverLap and Add* (SOLA) és *Pitch-Synchronous OverLap and Add* (PSOLA) algoritmusok.

A SOLA algoritmus során a jelet, egymást átfedő blokkokra osztjuk egymástól egyenlő távolságra. Ezután a cél a blokkok egymástól távolabbra való „széthúzása", azaz az átfedő intervallumok hosszának csökkentése. Ahhoz azonban, hogy ez ne okozzon hallható változást a jelben, előzetesen meg kell keresni egy maximálisan hasonló részt az átfedési intervallumban. Ez lehetséges például egy autokorrelációs függvény vizsgálatával. Ezek után a blokkokat egymáshoz képest az így meghatározott helyre eltolva elérhető a bemenő jel



2. ábra
Harmonikus jel kepsztruma



3. ábra
A hangmagasság detektálás eredménye lineárisan növekedő frekvenciájú bemenő jelre, interpolációval és interpoláció nélkül

hosszának nyújtása változatlan hangmagasság mellett. A lejátszási sebességet újramintavételezéssel változtatva az eredeti lejátszási idő visszaállítható, így a hangmagasság tetszőlegesen módosítható. A PSOLA algoritmus a SOLA algoritmus kiegészítése többek között dinamikusan változtatható blokkmérettel, amely blokkméretet egy előzetes alapperiódusidő-becslés alapján választ az algoritmus.

Az időtartománybeli módszerek előnye a hangmagasság detektáláshoz hasonlóan ismét gyorsaságukban rejlik, azonban nagy hátrányuk, hogy a kimenő jelben általában hallható torzítás keletkezik, ezért a professzionális, stúdiókörnyezetben is alkalmazott szoftverekben szinte mindig a következő frekvenciatartománybeli eljárást alkalmazzák.

Napjainkban a legszofisztikáltabb eljárás a frekvenciatartományban működő phase vocoder technika, 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 úgynevezett ugrásméret. 50%-os átlapolódás mellett például $R_a=N/2$. Ekkor levezethető, hogy az *s*. blokk esetén a jel STFT-je és DFT-je, illetve a fázisuk között az alábbi kapcsolat van:

$$X(sR_a,k)_{STFT} = W_N^{sR_ak}\tilde{X}(sR_a,k)_{DFT}\ , ahol\ W_N = e^{-j2\pi/N}$$

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

A feldolgozás folyamán tehát a fázist folyamatosan korrigálni kell. Hogy a fázis ne tartson végtelenbe, célszerű a]-π; π] 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 fázisfüggvény idő szerinti 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:

$$d\varphi = \frac{\Delta\varphi_{STFT}}{R_a} = \frac{\Delta\tilde{\varphi}_{DFT}}{R_a} - \frac{2\pi k}{N}$$

ahol a korrekciós tag fázis-korrekciót leíró egyenletet az *s*. és (s+1). blokkra felírva egyszerűen kijön. A pillanatnyi frekvenciák segítségével a hangmagasság módosítása egyszerűen végrehajtható a szinuszos összetevők közvetlen szintetizálásával az alábbi módon.

- Az analízis során kiszámoljuk a fázisváltozást egy minta alatt: $d\varphi(k) = \frac{\Delta\varphi(k)}{R_a}$

  Hasonlóan az egyes összetevők egy minta alatti amplitúdó változását:

  $$dA(k) = \frac{A\big((s+1)Ra,k\big) - A(sRa,k)}{R_a}$$

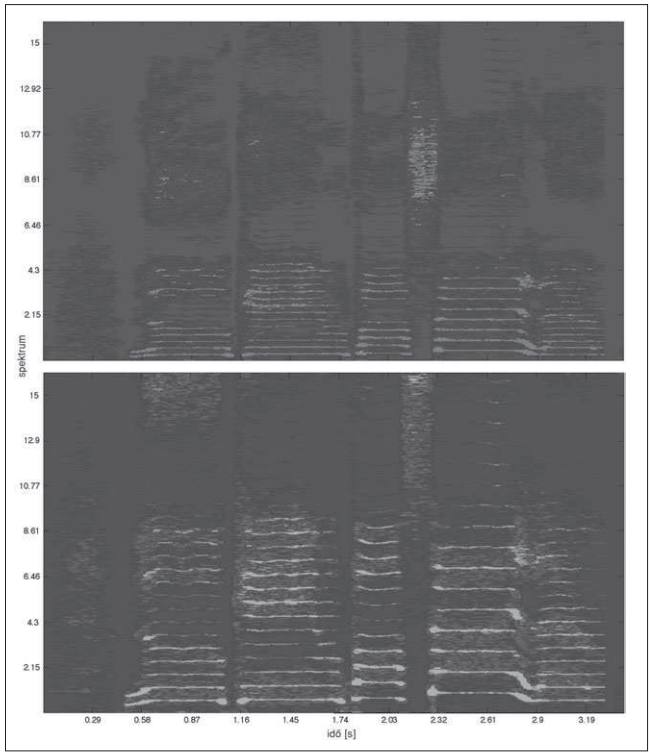
- A fázisváltozást megszorozzuk a transzpozíciós tényezővel (*transpo*), így végrehajtva a frekvenciamódosítást és integráljuk a módosított fázist az időtartományon úgy, hogy a blokkon belül az adott *X*(*n*,*k*) minta értéke:

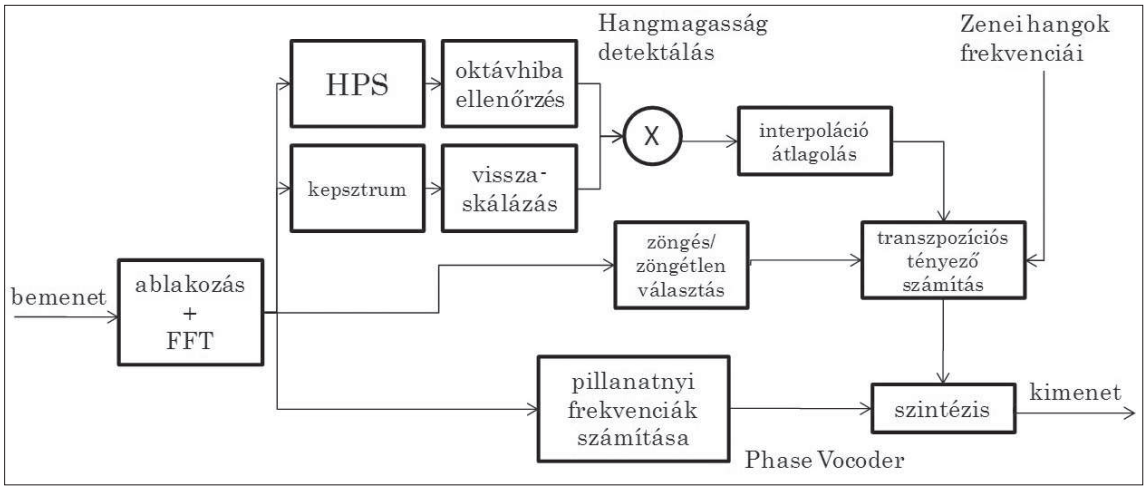
$$\tilde{\varphi}(n+1,k) = \tilde{\varphi}(n,k) + transpo\ \cdot d\varphi(k)$$
$$A(n+1,k) = A(n,k) + dA(k)$$
$$X(n,k) = A(n,k)\sin(\tilde{\varphi}(n,k))$$

- Az algoritmust minden frekvencián elvégezve közvetlenül, inverz Fourier-transzformáció nélkül újraszintetizálhatjuk az immár frekvenciában módosított összetevőket.

Az algoritmus eredménye a *4. ábrán* látható a hangmagasság oktávnyi emelése mellett. A kimeneti jelet meghallgatva elmondható, hogy torzítás egyáltalán nem hallható a hangmagasság-módosítás után. A bemutatotton kívül phase vocoder-t alkalmazva a hangmagasságváltoztatás akár a SOLA-hoz hasonló időnyújtás, majd újramintavételezés alapon is végrehajtható, hiszen, mint látható volt, a folytonos fázis a jelben biztosítható, így elkerülve a hallható torzítást, amely a blokkhatárokon történő ugrásból származik.

4. ábra
Az eredeti énekdallam és frekvenciakétszerezett dallam spektrumának változása az idő függvényében





5. ábra A teljes hangmagasság korrekciós rendszer blokkvázlata

4. A teljes rendszer felépítése és működése

A bemutatott algoritmusok segítségével egy teljesen automatikusan működő hangmagasság-korrekciós rendszer létrehozása lehetséges. A hangmagasság megváltoztatására a legjobb minőséget a phase vocoder nyújtja, amely a frekvenciatartományban működik. Ezt alkalmazva azonban eleve rendelkezésünkre áll a jel spektruma, így kézenfekvő, hogy a hangmagasság detektálását is a frekvenciatartományban hajtsuk végre. A következőkben egy pusztán a frekvenciatartományban működő rendszer működését mutatjuk be az előzőekben bemutatott módszereket alkalmazva.

A rendszer feladata, hogy meghatározza a bemenő hang alapfrekvenciáját, majd ez alapján végrehajtsa a hangmagasság módosítását úgy, hogy a végeredményként kapott énekhang a zenei skála hangjai közé tartozzon. 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:

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

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

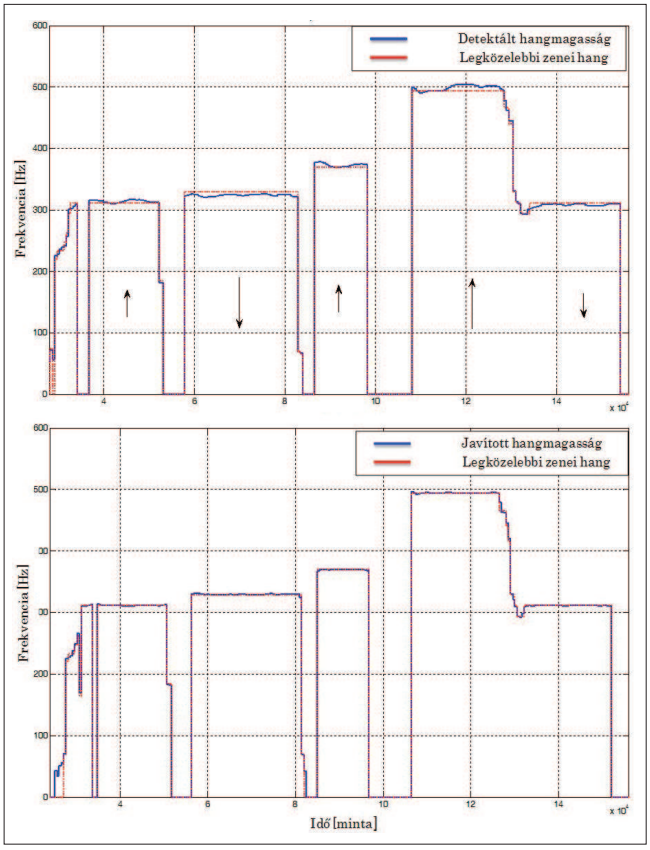
Énekdallamok esetén zöngétlen hangok esetén a hangmagasság nem értelmezhető, így fontos, hogy az algoritmus különbséget tegyen a zöngés, és a nem harmonikus, zaj jellegű, zöngétlen hangok között. A zöngés-zöngétlen különválasztására lehetőséget nyújt a spektrum tömegközéppontjának számítása, mivel a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

Ezek alapján a teljes rendszer felépítése az *5. ábrán* látható: A bemenő jelet átlapolódó blokkonként ablakozzuk és Fourier-transzformáljuk. Meghatározzuk az alapharmonikus frekvenciáját – a pontosabb eredmény érdekében frekvenciatartománybeli interpolációt alkalmazva – majd megvizsgáljuk, mely zenei hanghoz van legközelebb. Ha a bemenő hang zöngétlen mássalhangzó, nincs szükség a 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 *6/a. és b. ábrákon* látható: Női énekhangot stúdióprogram beágyazott hangmagasság-módosító algoritmusával hamissá téve – a módosítás „irányát” az ábrán nyilakkal jelölve – majd a rendszer hangmagasság-felismerő algoritmusán átfuttatva a detektált hangmagasság a 6/a. ábrán folytonos vonallal, a hozzá legközelebb eső zenei hang frekvenciája pedig szaggatott vonallal látható. A frekvencia egyes időpillanatokban zérusértékű, ezeken a helyeken a rendszer zöngétlen mássalhangzót, vagy csendet detektált. A 6/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. Összefoglalás

A kitűzött feladat – tehát egy teljesen automatikusan működő hangmagasság korrekciós rendszer létrehozása – a bemutatott módszerekkel végrehajtható. Az egyes feladatok lehetséges megoldásainak megismerése után egyértelművé vált, hogy a megfelelő minőség érdekében célszerű a hangmagasság-módosítást a frekvenciatartományban elvégezni, így végül az egész rendszer pusztán a frekvenciatartományban működik. Az ehhez szükséges DFT felbontása – amely a legfőbb limitáló tényező – interpolációval jelentősen növelhető, a munka során kidolgozott módszerrel akár tized Hz pontosságig is, amely már felül is múlja a feladathoz szükséges felbontást.



6/a. és 6/b. ábra
Hamis női énekdallam
automatikus hangmagasság javítás előtt és után

A phase vocoder eljárás hallható minőségi romlás nélkül képes a hangmagasságot módosítani, épp ezért a kereskedelmi forgalomban kapható szoftverek ezt az algoritmust használják. Az itt bemutatott rendszer számos 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. Bár hangmagasság-javításnál nincsenek félhangnyinál nagyobb transzpozíciós tényezők, a hangmagasságot tetszőlegesen változtatva az énekhang elváltozása érzékelhetővé válna. Ez elkerülhető formáns megőrzést alkalmazásával: mivel az egyéni énekhangot leginkább a spektrális burkológörbe jellemzi, azt megőrizve a hangmagasság-módosítás az énekhang változása nélkül hajtható végre. 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, így frekvenciafüggő vezérlőjellel lehetséges a rendszer működését akár polifonikus hangokra is kiterjeszteni. Ezeket a kiegészítéseket a rendszerbe integrálva a bemutatott módszerekkel akár a jelenleg forgalomban kapható, hasonló célú programok képességein túlmutató szoftvert lehet létrehozni.

A szerzőről

FIRTHA GERGELY 2010-ben szerzett BSc diplomát a Budapesti Műszaki és Gazdaságtudományi Egyetem Villamosmérnöki és Informatikai Karán. Jelenleg a BME MSc képzését végzi Médiatechnológiák és médiakommunikáció szakirányon. Kutatási területe főként a hangtér fizikai reprodukciója sokcsatornás hangrendszer segítségével, a hangtérszintézis.