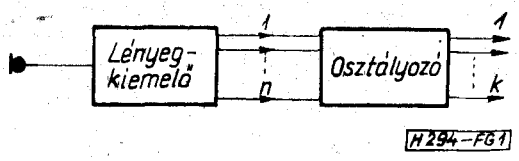
DR. FÖLDVÁRI RUDOLF – DR. GORDOS GÉZA
Híradástechnikai Elektronika Intézet

Az emberi hangmagasság-felismerés új hipotetikus modellje

ETO 534.784.072

A beszédfelismerési eljárások során szinte minden esetben felhasználásra kerülnek a hangmagasságot meghatározó adatok. A felismerés folyamata két fő részre, lényegkiemelésre és osztályozásra osztható (1. ábra).



1. ábra

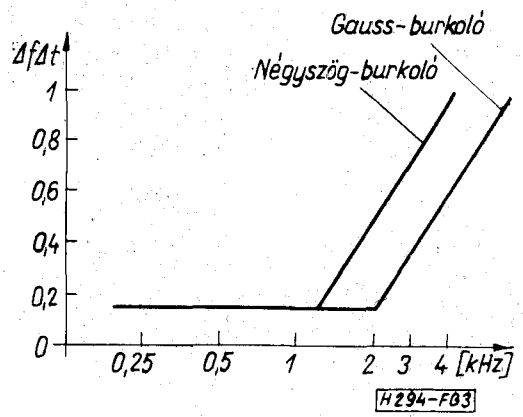
A lényegkiemelő feladata a beszédjel kb. 30000 illetve 50 000 bit/sec információsebességének lecsökkentése, és ezzel együtt olyan paraméterek előállítása, melyek az akusztikai alakzatra (pl. szó vagy fonéna) a legjellemzőbbek, illetve a különböző alakzatok szempontjából a legkülönbözőbbek. Az így előállított n paraméter alapján az osztályozónak kell az alakzatot a k osztály valamelyikébe sorolni. A lényegkiemelés elvégzésére egyenes eljárás nem ismeretes, csak ad hoc úton valósítható meg, azonban hasznos lényegkiemelési módszerek már kialakultak. A lényegkiemelő által előállított paraméterek közül az egyik legfontosabb paraméter a hangmagasságot meghatározó adat. A hangmagasság-paraméter előállítására több módszer került kidolgozásra, azonban az emberi hangmagasság meghatározó képesség egyikkel sem magyarázható maradéktalanul.

Az emberi hallás hangmagasság meghatározására az első tudományos hipotézis 1843-ból Ohmtól származik, és akusztikai Ohm-törvény néven ismeretes [1]. Ohm feltételezi, hogy a fül Fourier-analízist végez, és a hangmagasságot a legalacsonyabb frekvenciájú Fourier-komponens határozza meg. Helmholtz is csatlakozik Ohm elméletéhez, és úgy képzeli, hogy a Corti-szerv igen sok rezonátort tartalmaz, melyeknek mindegyike meghatározott frekvenciájú, és meghatározott idegszálat ingerel [2]. Békésy György kísérletileg kimutatta, hogy ilyen független rezonátorok a belső fülben nincsenek. A belső fület követő idegi feldolgozás sem képzelhető el a Helmholtz-féle rezonanciaelmélet, azaz kizárólag frekvenciatartományban történő feldolgozás alapján, ugyanis meglepően rövid idejű jelek is határozott hangmagasságérzetet alakítanak ki, és az emberi hallásnak ez a tulajdonsága nem modellezhető egy egyszerű sávszűrő rendszerrel.

Az elmondottak könnyen beláthatók a 2. ábra alapján, melyen egy Δt hosszúságú és f_0 frekvenciájú jelet, valamint Fourier-transzformáltját láthatjuk. Ha f_0 értékét annak alapján határozzuk meg, hogy a spektrum energiájának zöme a Δf sávon belül helyezkedik el, tehát f_0 is ebben a sávban van, akkor a meghatározás bizonytalansága a

$$\Delta f \Delta t \geq 1 \qquad (1)$$

Gábor-féle összefüggésből számítható. A $\Delta f \Delta t$ szorzat konstans, és értéke Δt és Δf definiálásától, valamint az időfüggvény burkolójától függ. Az emberi hallás idő- és frekvenciatartományban való viselkedését, azaz a $\Delta f \Delta t$ szorzat értékét a 3. ábra mutatja [3].

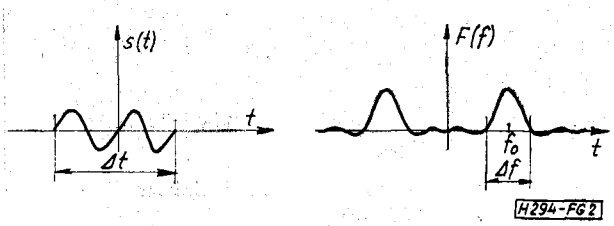


3. ábra

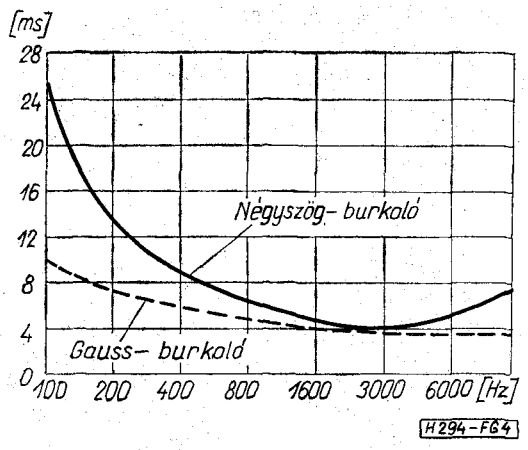
Mint az a 3. ábrából látható, Gauss-burkoló esetén kb. 2000 Hz-ig $\Delta f \Delta t \cong 0{,}13$. Ez az összefüggés érvényes, ha $\Delta t < 100$ ms, azonban testszőlegesen kicsi sem lehet, mert a hangmagasságérzet kialakulásához szükséges a 4. ábrán látható minimális idő [4]. Meglepő, hogy Gauss-burkoló esetén 1 kHz-nél 4 periódus, 100 Hz-nél pedig mindössze 1 periódus elegendő a hangmagasságérzet kialakulásához.

Hosszú idejű hangok hangmagasság-érzékelése sem magyarázható Fourier-analízissel, ugyanis a szubjektív hangmagasság nem mindig egyezik meg a hangspektrum legmélyebb összetevőjével, sőt lehetséges, hogy egyik komponenssel sem. Ha egy gazdag harmónikustartalmú hangból kiszűrjük az alapfrekvenviát, akkor továbbra is ezt a hiányzó alapfrekvenciát fogjuk hangmagasságnak hallani. Ez a jelenség a fül nonlinearitásával nem magyaráz-

Beérkezett: 1974. V. 21.



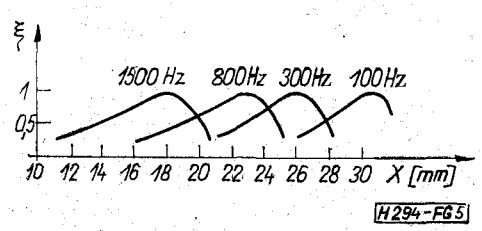
2. ábra



4. ábra

ható meg megnyugtatóan, mert léteznek olyan hangok, melyeknél a komponensekből adódó kombinációs termékek sem esnek egybe a szubjektív hangmagassággal. Ebben az esetben az időfüggvény burkolójának kváziperiódikusságát halljuk hangmagasságnak [5].

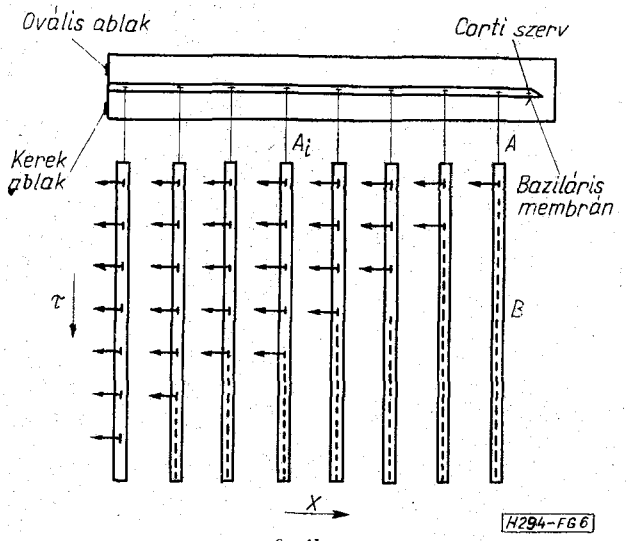
A különböző modellek az emberi hallás eddig tárgyalt tulajdonságait csak részben közelítik. Stacioner szinuszos jel esetén sem indokolható a hallás frekvencia megkülönböztető képessége a belső fülben kialakuló hullámformával. A hanginger a külső fülön keresztül a középfülben levő hallócsontok, majd az ovális ablak közvetítésével jut a belső fülbe, azaz tulajdonképpen a csigába. A csigában levő baziláris membránon — a membránon található Corti szervvel együtt — a hanginger hatására haladó hullám jön létre [6]. A membrán rezgésének amplitúdója a hely függvénye, melynek maximuma igen lapos (5. ábra). Az ábrán X az ovális ablaktól való távolság mm-ben, ξ a membrán kitérésének relatív amplitúdója, a görbéken található paraméter pedig a gerjesztés frekvenciája. A baziláris membrán mozgásából nem következik a fül éles analizálási képessége, és a fiziológiai vizsgálatokból megállapítható, hogy a Corti-szervből az agy felé haladó idegpályák között keresztirányú összeköttetések is léteznek, melyek a gyengébben ingerelt idegpályákat gátolják (laterális inhibitió) [7]. Továbbá megállapították, hogy minél közelebb fekszik a megfigyelési hely a központi idegrendszerhez, annál élesebben koncentrálódik egy szűk frekvenciatartományra az egyes idegek ingerelhetősége [8]. Ezt a jelenséget Zwicker úgy modellezte, hogy a jelet paralel sávszűrőkkel analizálta, majd a szűrők után szelektivitásnövelést alkalmazott, azaz a jelszegény csatornában az erősítést csökkentette [9]. A modell felépítését tekintve követi eddigi fiziológiai ismereteinket, azonban teljesen azonosan működik akár



5. ábra

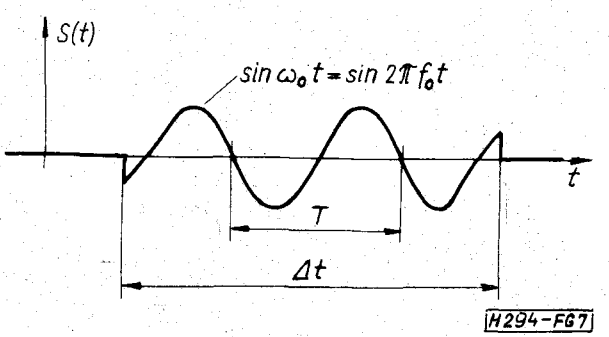
koherens akár inkoherens komponensekből áll az analizálandó jel.

Licklider egy lényegesen bonyolultabb modellt készített, mely azon a feltételezésen alapul, hogy a hangmagasság-megkülönböztetés a hallás idegi részében autokorrelátorokkal történik. Ez a feltételezés lehetővé teszi a burkoló periodicitásának felismerését is [10]. A modell a 6. ábrán látható. A B autokorrelátorok mindegyike a Corti-szerv X_i helyéről kiinduló A_i idegszál akciópotenciálját korrelálja. A hangkép teljesítményspektruma az X irányban, a periodicitása pedig a τ irányban ismerhető fel. A modell szemléletes, a hangkép egy hálózatra képződik le. Ennek ellenére ezzel a modellel sem lehet a fül idő-frekvencia felbontóképességét indokolni. Az X irányú teljesítménysűrűség spektrum pontosan ugyanolyan tulajdonságokkal rendelkezik, mint az a 2. ábrán látható, továbbá az autokorrelációs függvényből számítható teljesítménysűrűség spektrum is azonos tulajdonságú, azaz továbbra is érvényes a $\Delta f \Delta t \geq 1$ bizonytalansági reláció [11].

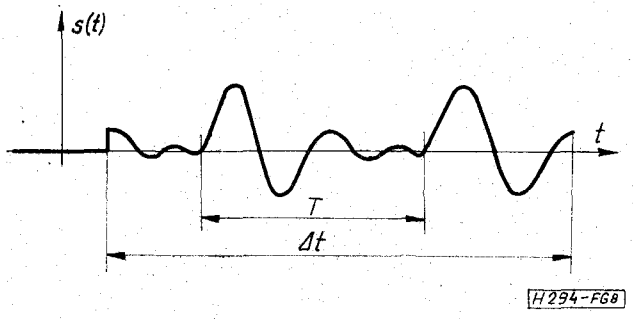


6. ábra

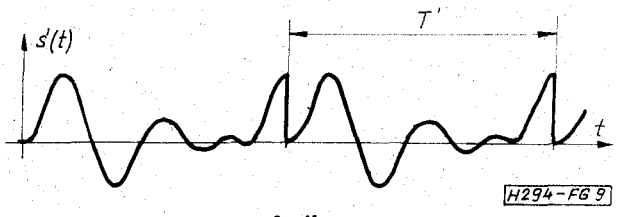
Az idő- és frekvenciatartomány kapcsolatát kifejező bizonytalansági reláció érvényességi területe megkerülhető, ha a frekvenciatartományra vonatkozó adatot (vagy adatokat) is az időtartományban való méréssel állapítjuk meg. Vizsgáljuk meg először a legegyszerűbb esetet, azaz legyen a hangkép egyetlen szinuszos jel (7. ábra). A megfigyelésre rendelkezésre álló Δt idő alatt szinte tetszőleges pontossággal lemérhető a T periódusidő, illetve



7. ábra



8. ábra

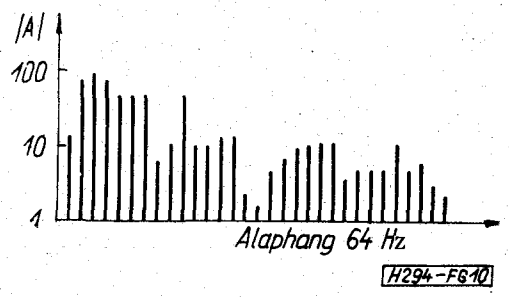


9. ábra

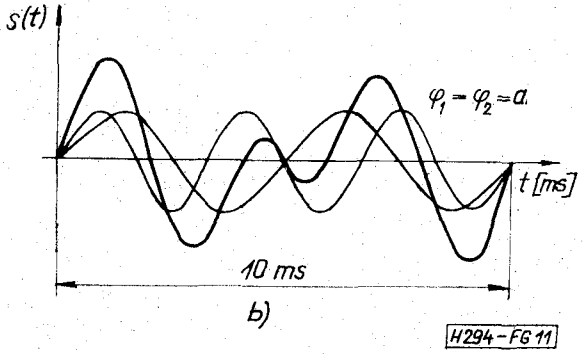
f_0 értéke. A frekvenciamérés hibáját, a Δf-et kizárólag az időmérés pontossága szabja meg. Ha az időmérés abszolút hibája állandó, akkor jó közelítéssel továbbra is érvényes a $\Delta t \Delta f \cong$ konst., ahol a konstans értéke tetszőlegesen kicsi lehet. A periódusidő mérésének ezt az elvét minden digitális műszer kihasználja.

Ha az $s(t)$ időfüggvény periodikus, és nem egyetlen szinuszos jelből, hanem több komponensből áll, akkor az időfüggvény egy periódusánál hosszabb szakasz és a periódusidő ismeretében Fourier-sorfejtéssel meghatározható a komponensek értéke (8. ábra). Az ábrán felrajzolt időfüggvény természetesen nem csak a T szerint fejthető Fourier-sorban, hanem T' idővel is képezhető egy periodikus folytatás (9. ábra). A T' idővel képzett periodikus folytatásból számított komponensek azonban nem az eredeti, hanem a 9. ábrán látható időfüggvényt közelítik. A kétféle módon nyert komponensekből visszatranszformált időfüggvény T időn belül megegyezik, azonban ha a felbontást felismerésre akarjuk felhasználni, akkor a két közelítés között lényeges különbség van. Tételezzük fel, hogy a 8. ábrán látható Δt ideig megfigyelt jel egy gordonkán megszólaltatott hang stacioner része. A végtelen hosszúnak tekinthető hang spektruma a [12] irodalomból átvéve a 10. ábrán látható.

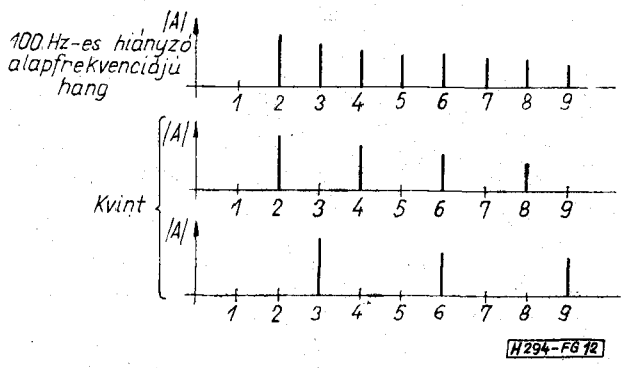
Tételezzük fel továbbá, hogy a jel sávkorlátozott, azaz a frekvenciatartománya véges. Ha egy ilyen időfüggvény Δt ideig figyelhető meg, és az időfüggvény T periódusidejű szakaszát Fourier sorba fejtjük, akkor megkapjuk a 10. ábrán látható komponenseket. Ezek a komponensek nemcsak közelítik, hanem pontosan megadják az $s(t)$ időfüggvényt. A célszerűtlenül felvett T' idővel való sorfejtés eredménye egy végtelen sok komponensből álló spektrum lesz, mely az eredeti időfüggvényt T' időn belül is csak négyzetes értelemben közelíti. Ezenkívül a megszólaltatott hang leglényegesebb információját, a hang magasságát a spektrum nem tartalmazza. A következőkben vizsgáljuk egy alapfrekvenciát nem tartalmazó hangot. (11*a* és 11*b* ábrák) Az időfüggvény periodikus, a periódusidő fázishelyzettől függetlenül 100 Hz periódusidejével, azaz 10 ms-al egyenlő. Ezt a jelet azonban nem 100 Hz-nek, hanem két hangnak halljuk (kvint). Ha az időfüggvényben további komponensek is jelen vannak, (400, 500, 600 Hz stb.), akkor egyre pregnásabb lesz a periodicitása, és nem különálló frekvenciákat, hanem egy 100 Hz-es hangot hallunk. Ezzel szemben ha két harmonikusokban gazdag 200 Hz-es és 300 Hz-es hangot hallgatunk, akkor az időfüggvény 100 Hz-es periodicitása elenére is kvintet hallunk. Önmagában az időfüggvény periodicitásával tehát az alaphang hallása nem magyarázható. Feltételezve, hogy a T időt, továbbá a spektrumot is ismerjük, már különbséget tehetünk az alaphang nélküli 100 Hz és a harmonikusokban gazdag, együttesen megszólaló 200 Hz és 300 Hz között. A 12. ábrán felrajzolt spektrumokból jól látható, hogy a kvint spektrumából a 100 Hz-en kívül még további komponensek is hiányoznak (500 Hz és 700 Hz). Természetesen a felsorolt példákban is megtehetjük, hogy a Fourier-komponenseket nem a T periódusidővel, hanem egy T' idővel periodikussá tett időfüggvényből származtatjuk, azonban így semmiféle összefüggést nem kapunk az időfüggvény eredete, és a Fourier komponensek között. Az elmondottakra talán a legjellemzőbb példa a 13. ábrán látható, ugyanis a T' idővel képzett periódikus folytatás az f_0 frekvenciát nem tar-



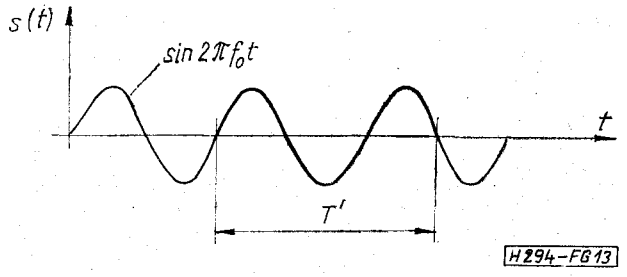
10. ábra



11. ábra



12. ábra



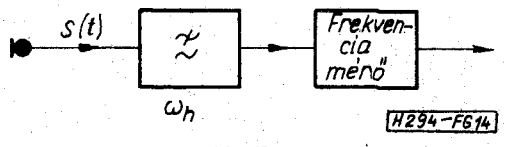
13. ábra

talmazza. Megállapíthatjuk tehát, hogy akár zenei hangok, akár beszédhangok kvázistacioner részleteinek feldolgozásához (lényegkiemeléshez) először meg kell állapítani a T periódusidőt, majd ennek ismeretében kell elvégezni a Fourier-transzformációt. Ezt az eljárást az irodalomban szinkron Fourier-transzformációnak nevezik.

Zöngés beszédhangok szinkron Fourier-transzformációval történő analizálásának legnagyobb nehézsége a periódusidő megállapítása. Folyamatos beszédben a zöngés részleteken belül a zönge frekvenciája, azaz a T periódusidőnek megfelelő frekvencia változik, továbbá változnak a jelet előállító komponensek amplitúdói is. A változások különösen a különböző fonémák kapcsolódási helyén nagymértékűek, de a fonéma közepe környékén sem teljesen egyformák a T idejű szakaszok. Az $s(t)$ időfüggvény regisztrátumának ismeretében a T idejű szakaszok „ránézéssel" történő felismerése semmiféle nehézséget nem jelent, azonban a folyamat gépi megvalósítása nem könnyű feladat.

A periódusidő megállapítására több eljárás ismeretes. A következőkben röviden ismertetünk néhány módszert, melyek különböző elven alapulnak:

a) Az egyik legegyszerűbb megoldás a periódusidő meghatározására, ha az $s(t)$ jelből egy aluláteresztő szűrővel az alapfrekvenciát kiszűrjük, és frekvenciáját digitális elven lemérjük (14. ábra). Ez az egyszerű megoldás több hátránnyal rendelkezik. Legnagyobb hibája, hogy minden beszélőhöz illeszteni kell, mert a határfrekvenciának az első

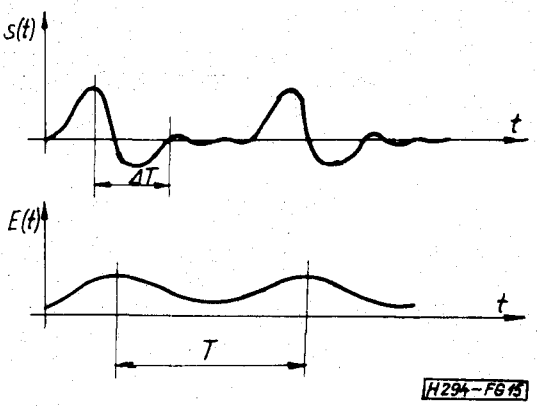


14. ábra

és a második komponens közé kell esnie. (Az egy beszélő jelében előforduló változások maximálisan 10—20% nagyságúak.) Ezenkívül olyan jelek periódusidejének meghatározására, melyek nem rendelkeznek alapfrekvenciával, nem alkalmas. A telefoncsatornán átvitt jel az esetek döntő többségében nem tartalmaz alapfrekvenciát, ugyanis a zönge frekvenciája 75 Hz és 400 Hz közé esik. (Férfi beszélők átlaga kb. 125 Hz.)

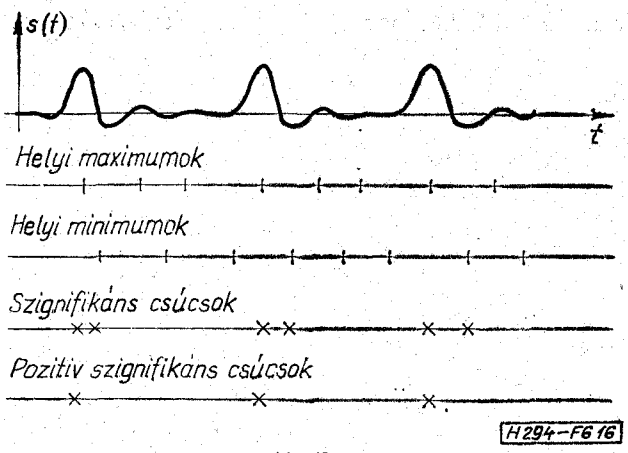
b) Egy másik megoldás az $s(t)$ kváziperiodikus jelnek azt a tulajdonságát használja ki, hogy a jel rövid idejű energiája a periódusidővel együtt változik (15. ábra). Az $s^2(t)$-et ΔT ablakidőre integrálva, és az ablakot a t időtengely mentén folytonosan eltolva az $E(t)$ jelet kapjuk. A ΔT időre számított energia maximumai megadják a T periódusidőt. Az eljárás nehézkes a ΔT idő helyes megválasztása miatt, ugyanis rövid ΔT idő esetén az $E(t)$ függvény több helyi maximummal rendelkezik, túl hosszú ΔT idő esetén pedig nem kapunk határozott maximumokat.

c) A periódusidő meghatározását még úgy is megvalósíthatjuk, hogy megpróbáljuk utánozni azokat a funkciókat, melyeket az időregisztrátum ismeretében tulajdonképpen mi is elvégzünk a periódusidő „ránézéssel" történő megállapításánál. Az időfüggvény helyi maximumainak, illetve helyi minimumainak meghatározása után különböző feltételek segítségével meghatározzuk a szignifikáns csúcsokat, és a pozitív szignifikáns csúcsok közötti távolságot tekintjük periódusidőnek (16. ábra). A szignifikáns csúcsok meghatározásához szükséges feltételek megkötéseket tartalmaznak az abszolút értékre (pl. nagyobb az abszolút maximum 90%-ánál), továbbá az időtengelyen mért távolságokra (pl. szignifikáns pozitív illetve negatív csúcsok között legalább 2,5 ms a távolság). A feltételek számának növelésével egyre biztonságosabbá tehetjük a periódus felismerését. E módszer hatásossága zajos beszéd analizálása esetén erősen romlik.



15. ábra

Az *a)*, *b)*, *c)*-vel jelölt és vázlatosan bemutatott módszerek közül az *a)*-val jelölt megoldás már akkor is használható, ha a megfigyelésre rendelkezésre álló idő a periódusidőnél hosszabb, de nem szükséges két teljes periódus. Mint már említettük hiányzó alaphangú jel analizálása esetén nem használható, továbbá nem tudjuk előre, hogy a szűrő



16. ábra

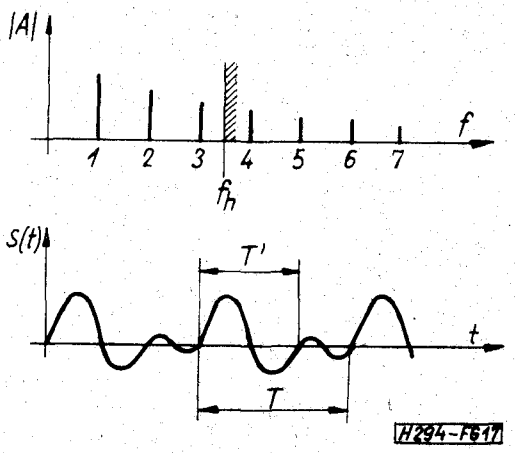
határfrekvenciáját hová kell választani. A *b)* és *c)* eljárások alkalmazásához legalább két teljes periódus szükséges, ugyanis a periódusidő meghatározásának éppen az az alapja, hogy az ismétlődés felismerését oldja meg gépi úton.

A továbbiakban az eddig ismertetett eljárások előnyeit egyesítő új periódusidő meghatározó módszert kívánunk bemutatni.

Tételezzük fel, hogy egy periodikus jel a 17. ábrán látható komponensekkel rendelkezik, és a jelet egy f_h határfrekvenciájú aluláteresztő szűrővel megszűrjük. A szűrő kimenetén nyert jelet jelöljük s(*t*)-vel Az *s(t)* jelről biztosan tudjuk, hogy f_h frekvencia feletti komponenseket nem tartalmaz, azaz ha két tetszőleges pozitív nullátmenet közötti szakasz (pl. az ábrán T'-vel jelölve) periodikus folytatását Fourier-sorba fejtjük, akkor a sorfejtés eredményeként f_h feletti komponenseket is kapunk, tehát a T' nem lehet periódusidő. Távolabb pozitív nullátmeneteket választva, és az így képzett periodikus folytatásra a sorfejtést újból elvégezve, a 17. ábrán felrajzolt esetben már f_h felett nem kapunk zérustól különböző komponenseket, tehát az így felvett idő periódusidő.

Az eljárás nem használja ki az egymásutáni szakaszok hasonlóságát, ezért ha a megfigyelésre rendelkezésre álló idő egy periódusnál hosszabb, akkor már megállapítható a periódusidő.

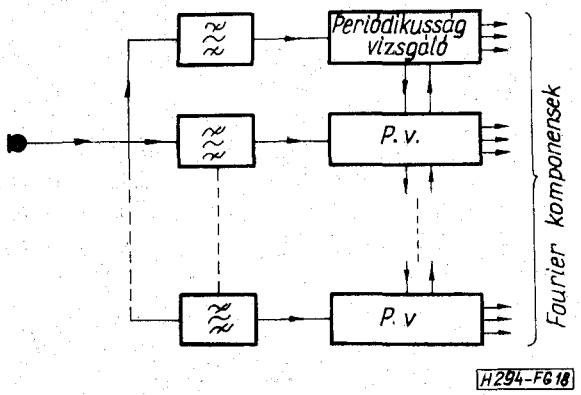
Ha feltételezzük, hogy az emberi hallás hasonló eljárással dolgozza fel a hangingert, természetesen



17. ábra

nem egy aluláteresztő szűrő, hanem egy sávszűrő rendszer felhasználásával, (18. ábra) akkor ezzel a modellel jól magyarázható az alaphanggal nem rendelkező hangok, a kváziperiodikus hangok, valamint az eredőben nem periodikus, de frekvenciában távoli periodikus hangok érzékelése. A modellben célszerű feltételezni, hogy a szomszédos frekvenciasávok periodikusság vizsgálatai között kölcsönös kapcsolat van.

Véleményünk szerint az eljárás zöngétlen hangok esetén is használható. A sávszűrő frekvenciasávján kívül eső komponensek ebben az esetben növekvő T idővel sem lesznek nullák, de tetszőlegesen közelíthetők. Így az eljárás nemcsak periódusidő meghatározására, hanem lényegkiemelésre is használható. Várható, hogy az eljárás viszonylag kevés, és jól szeparálható paramétert eredményez. Sajnos a rendelkezésre álló eszközökkel a modellt nem tudtuk megvalósítani, ezért egyelőre az elérhető információcsökkentés mértékéről, és ezen az úton történő felismerés hatékonyságáról nem tudunk beszámolni.



18. ábra

Végezetül megjegyezzük, hogy az emberi hallásra jellemző, és Püthagorasz óta ismert oktáv kapcsolatot ezzel a modellel sem lehet megnyugtatóan megmagyarázni.

IRODALOM

[1] *G. S. Ohm:* Ann. der Physik *59.* (1843) 497.
[2] *H. v. Helmholtz:* Die Lehre von den Tonenpfindungen. Braunschweig 1913.
[3] *L.M. Grobben:* Appreciation of short tones. Seventh international congress on acoustics, Budapest 1971 Vol. 3. 329—332.
[4] *Türk,W.:* Über physiodogisch-akustischen Kennseiten von Ausgleichsvorgängen. Akust. Z. (1940) 129.
[5] *J. Pfandner:* Der Einfluss niehtlinearer Verzenungen beim gleichzeitigen erklingen zweier oder einer grössen Zahl Harmonicher ohne Grundfrequenz. Seventh international congress on acoustics, Budapest 1971. Vol. 3. 673—676.
[6] *G. v. Békésy:* Experiments in Hearing Mc Graw Hill. 1960.
[7] *Reichardt, W., G. MacGinitie:* Zur Theorie der lateralen Inhibition. Kybernetik *1* (1962).
[8] *G. v. Békési:* J. Aconst. Soc. Amer *31* (1959).
[9] Zwicker, E.: Über ein einfaches Funktionsschema des Gehörs. Acustica *12* (1962).
[10] *Licklider, J. C. R.:* Experientia *7* (1951).
[11] *R. M. Fano:* Short-Time Autocorrelation Functions and Power Spectra. J. Acoust. Soc. Am. *22* (1950). 546—550.
[12] *Meyer, E., G. Buchmann:* Die Klangspektren der Musikinstrumente. Berl. Berichte (1931).