Tartalom

Beszédtechnológiák

nemeth@tmit.bme.hu

Ebben a számban egyrészt a gépi beszédkeltés fejlődésének aktuális kérdéseiről olvashatunk, másrészt a szélesebb értelemben vett akusztikai, beszéd- és jelfeldolgozás eredményeiből kaphatunk ízelítőt.

Az első blokk a gépi beszédkeltéshez kapcsolódó négy írásból áll. Elsőként egy áttekintő jellegű cikket olvashatunk arról, hogyan lehet a beszédfelismerésben már hosszabb ideje meghonosodott, rejtett Markov-modell (Hidden Markov Modell, HMM) alapú technológiát a magyar nyelvű beszédszintézis területén is alkalmazni. A gépi beszédkeltés minősége már elért arra a szintre, hogy a szövegek érthetősége ritkán jelent problémát. A hosszabb, géppel előállított felolvasás azonban általában monotonnak, robotosnak tűnik. A monotonitás csökkentésére kidolgozott új eljárást mutat be a második dolgozat. A legjobb hangminőséget ma a jelentős méretű (több óra) hanganyagot alkalmazó és többnyire kötött témakörökre kidolgozott, úgynevezett korpusz-alapú beszédszintetizátorok adják. A harmadik cikk azt vizsgálja, hogy a magyar nyelvre, kötött témakörökre kidolgozott korpusz-alapú technológiát hogyan lehetne a kötetlen szókészlet irányába kiterjeszteni. Ezt a blokkot a beszédadatbázisok pontosabb címkézésének megoldásait elemző írás zárja. Ennek az ad jelentőséget, hogy az adatbázisokra épülő alkalmazások teljesítménye jelentős mértékben függ az adatbázis-címkézés minőségétől.

A második blokk a beszéd- és más akusztikai jeleket változatos megközelítésben elemző öt dolgozatot tartalmaz. Először egy érdekes zajcsökkentési algoritmusról olvashatunk. Ezután a PPKE kutatóinak a Híradástechnika korábbi számaiban már részletesen ismertetett, akusztikus jelből szájmozgást modellező eljárásának egy újabb alkalmazását ismerhetjük meg. A megoldás segítségével IPTV-s jelfolyamba valós időben illeszthető a siket embereket segítő, géppel keltett szájmodell. Majd egy, a számítógépes modellezésnek a teremakusztikában történő alkalmazását konkrét példákkal illusztráló cikk következik. A természetes beszédértésben jelentős szerepe van a prozódiának, például sok esetben egy mondat kérdő vagy kijelentő jellege csak annak alapján dönthető el. A gépi beszédfelismerés azonban ennek a feldolgozására csak ritkán vállalkozik. Egy ilyen kísérletet mutat be a blokk utolsó előtti írása. A záró dolgozat egy Kempelen Farkas óta sokakat megmozgató problémára, sakkozó automata kidolgozására mutat be egy friss hazai kísérletet.

Németh Géza
vendégszerkesztő

Szabó Csaba Attila
főszerkesztő

Rejtett Markov-modell alapú mesterséges beszédkeltés magyar nyelven

TÓTH BÁLINT, NÉMETH GÉZA
Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék
{toth.b,nemeth}@tmit.bme.hu

Lektorált

Kulcsszavak: beszédszintézis, szöveg-beszéd átalakítás, rejtett Markov-modell

Jelen cikk bemutatja a rejtett Markov-modell alapú szövegfelolvasás technológiáját és annak a magyar nyelvre való adaptációját. Ennek a megoldásnak számos előnye van: kis adatbázisméret mellett jó minőségű beszédet képes előállítani, továbbá elvi lehetőséget ad a beszédhang karakterének, stílusának módosítására és érzelmek kifejezésére is meg lehet tanítani a rendszert.

1. Bevezetés

Napjainkban számos automatikus szövegfelolvasási módszer létezik: a beszédkeltés mechanizmusát modellező formáns- és artikulációs szintézistől kezdve a diádos és triádos hullámforma-összefűzéses szintézisen át az elemkiválasztó (korpusz) szintézisig. A legjobb minőséget nyújtó korpusz alapú szövegfelolvasó rendszerek adatbázisának a mérete igen nagy (gigabyte-os nagyságrendbe esik), és a beszélő hangját az adatbázis meghatározza, azon változtatni új felvételek nélkül nem lehet. Új felvételek esetén (amennyiben például érzelmet kifejező beszédet is meg szeretnénk valósítani), számolnunk kell a további stúdiófelvételekkel járó munkával, a felvételek adatbázisba való feldolgozásával és az amúgy is hatalmas adatbázis további növekedésével.

A rejtett Markov-modell (Hidden Markov Model, HMM) alapú szövegfelolvasók szintén az elemkiválasztós rendszerek közé tartoznak, azonban itt az elemeket nem hullámforma egységek jelentik, hanem a hullámformából kinyert spektrális és prozódiai jellemzők sokasága. A HMM-ek feladata ezek közül kiválasztani a felolvasandó szöveget legjobban reprezentáló elemeket, mely elemekből a régről ismert, beszédkódolókban is használt modellekkel készítenek mesterséges beszédhangot.

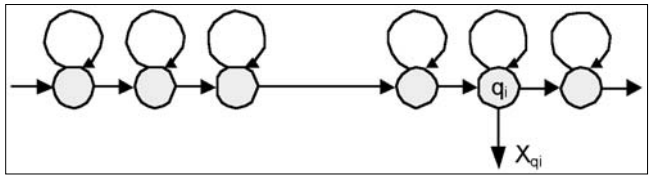
A HMM alapú beszédszintézis számos előnye miatt lett az elmúlt évek egyik legsikeresebb mesterséges beszédkeltési technológiája: kis (1,5-2 Mbyte) adatbázis méret mellett képes jó minőségű, érthető beszédet előállítani, amely hordozza a beszélő hangszínezeti tulajdonságait is. Továbbá lehetőség van több beszélőtől származó felvételek alapján betanított adatbázisból kiindulva viszonylag rövid (5-8 perces) beszédkorpuszokkal a beszélő hangkarakterisztikájának a módosítására [1-4], illetve érzelmek kifejezésére [5].

Jelen cikk először áttekinti a rejtett Markov-modell alapú mesterséges beszédkeltés alapjait, majd ismerteti egy nyílt forráskódú HMM-alapú szövegfelolvasó rendszer magyar nyelvű változata kialakításának lépéseit, bemutatja a rendszerrel végzett meghallgatásos teszt eredményeit, továbbá a jövőbeli terveinkre is kitér.

2. A rejtett Markov-modell alapjai

A rejtett Markov-modellt sikeresen használják a beszédfelismerés [6] és az utóbbi időben a beszédszintézis területén is. Jelen szakasz rövid áttekintést ad a módszer alapjairól, pontos ismertető a [7] cikkben található.

Legyen $\lambda(A,B,\pi)$ egy adott rejtett Markov-modell, melyet paraméterei határoznak meg: A – állapotátmeneti valószínűség, B – kimeneti valószínűség, π – kiinduló állapot valószínűség. Beszédszintézis esetén legyen ez a λ HMM egymást követő kvinfón (öt hangból álló hangsorozat) HMM-ek sorozata *(1. ábra)*. Ezek a kvinfónok határozzák meg azt a szót, amit generálni szeretnénk. Célunk a legvalószínűbb állapotsorozathoz tartozó **X** tulajdonságvektor megtalálása, ami alapján a 3. pontban ismertetésre kerülő módon generálni tudjuk a beszédet.



1. ábra
Összefűzött kvinfón HMM lánc a q_i állapotban, i. időegységben, kimenet X_{qi}

Az X_{qi} kimenet egy M dimenziós tulajdonság-vektor a λ HMM q_i állapotában:

$$X_{q_i} = (x_1^{(q_i)}, x_2^{(q_i)}, x_3^{(q_i)} \ldots x_M^{(q_i)})^T$$

Célunk a λ HMM-ből azt az $\underline{x} = (X_{q_1}, X_{q_2} \ldots X_{q_L}$ kimeneti tulajdonság vektort meghatározni, ami L db állapot mellett maximalizálja a $P(x|\lambda)$ összesített hasonlósági mértéket:

$$\underline{x} = \arg\max_x \{P(x|\lambda)\} = \arg\max_x \left\{ \sum_Q P(x|q,\lambda) P(q|\lambda) \right\},$$

Ahol $Q = (q_1, q_2 \ldots q_L)$ a λ HMM-ben az állapotok sorrendje. A képlet alapján a $P(x|\lambda)$ összesített hasonlósági mértéket a $P(x|q,\lambda)$ kimeneti valószínűség és a

$P(q|\lambda)$ állapotsorrend-valószínűség szorzatának az összes lehetséges Q állapotsorrenden való összegzése adja.

Ennek kiszámolására Viterbi-algoritmust szoktak használni, mert az összes lehetséges állapotsorrend bejárása túl nagy számításigényű. Így

$$\underline{x} \approx \arg\max_{x} \left\{ P(x|q,\lambda,L) P(q|\lambda,L) \right\}$$

A λ HMM $\underline{q}$ állapotsorrendjét $\underline{x}$-től függetlenül lehet maximalizálni:

$$\underline{q} = \arg\max_{q} \left\{ P(q|\lambda,L) \right\}$$

Tegyük fel, hogy minden q_i állapot esetén a kimeneti valószínűségi eloszlás Gauss-i valószínűségsűrűségfüggvény μ_i várható értékkel és Σ_i kovariancia mátrixxal. A λ HMM az összes a várható értékek és kovarianciamátrix halmaza:

$$\lambda = (\mu_1, \Sigma_1, \mu_2, \Sigma_2 \ldots \mu_N, \Sigma_N)$$

Ezt felhasználva a logaritmikus hasonlóságimértékfüggvény a következőképp alakul:

$$\ln\left\{P(x|q,\lambda)\right\} = -\frac{LM}{2}\ln\{2\pi\} - \frac{1}{2}\sum_{t=1}^{L}\ln\left\{|\Sigma_{q_t}|\right\} - \frac{1}{2}\sum_{t=1}^{L}(x_t - \mu_{q_t})^T \Sigma_{q_t}^{-1}(x_t - \mu_{q_t})$$

Ebben az egyenletben ha x-et maximalizáljuk, akkor az

$$\underline{x} = (\mu_{q_1}, \mu_{q_2} \ldots \mu_{q_L})$$

megoldást kapjuk, ahol a kimeneti tulajdonságvektor megegyezik az adott állapotok várható értékeivel. Ez a megoldás a beszédre nem alkalmazható megfelelően, ezért szükségünk van a tulajdonságvektor első és második deriváltjára is:

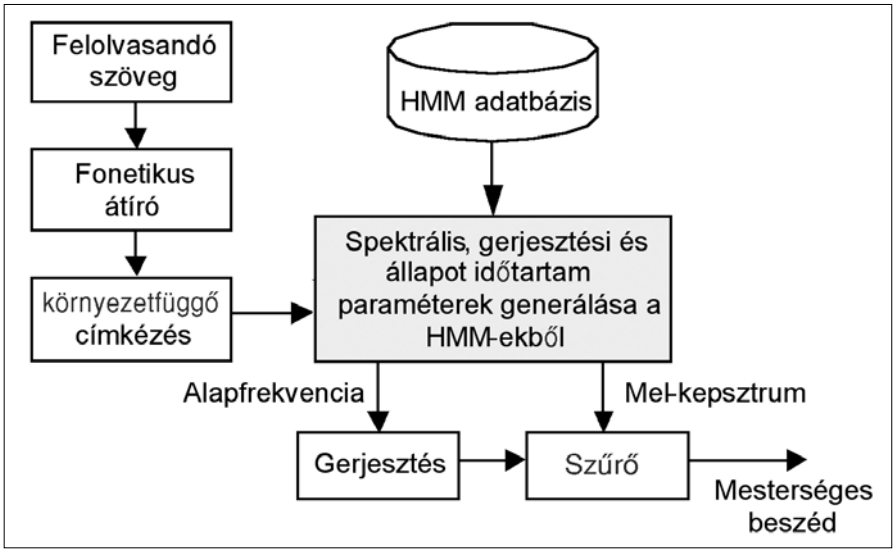
$$\underline{x} = ((x_{q_t})^T, (\Delta x_{q_t})^T, (\Delta^2 x_{q_t})^T)$$

3. A HMM alapú beszédszintézis

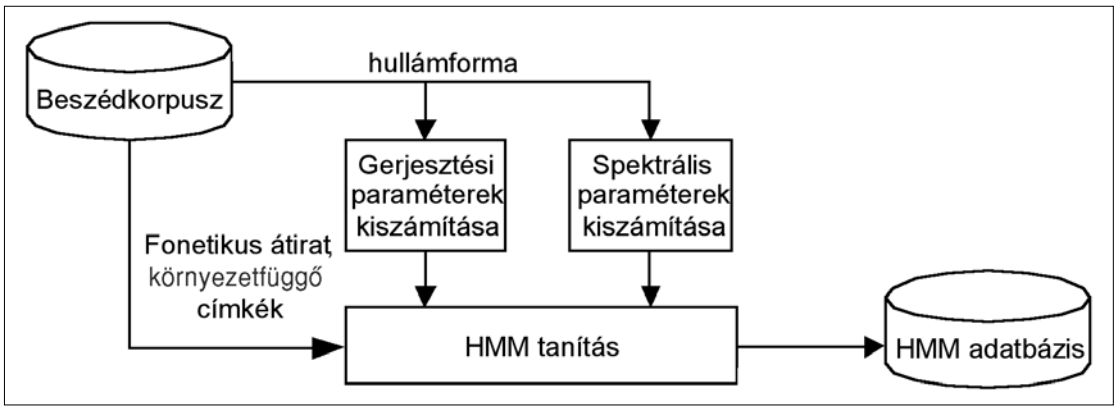
A HMM alapú beszédszintézist két részre oszthatjuk: a tanulás fázisára *(2. ábra)*, melynek során a HMM-eket tanítjuk be a beszédkorpuszunk alapján, illetve a beszédelőállítás fázisára *(3. ábra)*, amikor a betanított HMM-ekből kinyerjük a spektrális paramétereket, az időtartamokat és az alapfrekvenciát.

A tanításhoz szükségünk van egy nagyméretű beszédkorpuszra, annak fonetikus átiratára és a hanghatárok pontos pozíciójára. A hullámformából kinyerjük a mel-kepsztrumot, annak első és második deriváltját, továbbá az alapfrekvenciát, és annak az első és második deriváltját, majd a fonetikus átiratot ki kell bővítenünk környezetfüggő címkékkel (bővebben lásd a 4.2. szakaszt). Ezután elkezdődhet a HMM-ek tanítása. Ennek során a modellt betanítjuk a környezetfüggő címkéknek megfelelően a spektrális és a gerjesztési paraméterekre. Ahhoz, hogy a változó dimenziójú paramétereket (pl. $\log\{F_0\}$ a zöngétlen hangoknál) megfelelőképp tudjuk modellezni, többdimenziós valószínűségi eloszlást kell használni. Minden HMM-nek van egy állapotidőtartam-valószínűségi sűrűségfüggvénye a beszéd ritmusának (hangidőtartamok) modellezése érdekében.

A betanításhoz elsősorban kétfajta módszert lehet használni: betaníthatjuk a HMM-eket egy beszélőtől származó 2-4 órás adatbázissal, illetve betaníthatjuk több beszélőtől gyűjtött adatbázisokkal (beszélőnként 1-1,5



2. ábra
A HMM alapú szövegfelolvasó tanítása



3. ábra
A HMM alapú szövegfelolvasó beszédelőállítási mechanizmusa

óra hanganyag, minimum 3-4 különböző hang), és végül 5-8 perces adatbázissal egy adott hangra adaptálhatjuk [1,2]. Így új hangszínezetű beszéd generálására készíthetjük fel a rendszert adott hangú, igen kis beszédkorpuszok segítségével. A korábbi források [1,2] alapján az adaptív módszerrel előállított hangok jobb minőségűek lesznek, mintha csak egyetlen beszélőtől felvett adatbázissal tanítottuk volna be a rendszert. Ezen túl még számos módszer létezik a beszédhang jellemzőinek a megváltoztatására [3,4].

A beszéd előállítása során első lépésként elkészítjük a szöveg fonetikus átiratát környezetfüggő címkékkel (lásd 4.2.). Következő lépésként a hangidőtartamokat nyerjük ki az állapotidőtartam-valószínűségi sűrűségfüggvényekből, majd a legvalószínűbb spektrális és gerjesztési paramétereket nyerjük ki a HMM-ekből. Ezen paraméterek alapján állítjuk elő a mesterséges beszédet a gerjesztő jel és egy szűrő segítségével (tipikusan mel log spektrum approximációs (MLSA) szűrőt használnak [8]). Korábban egyszerű beszédkódolót használtak a hang előállításához, újabban pedig a jobb minőséget produkáló kevert-gerjesztési modellt is alkalmazzák [9].

4. Magyar nyelvű adaptáció

A kísérleteket a HTS keretrendszer segítségével végeztük el [10]. A magyar nyelvű változat elkészítéséhez szükség volt egy beszédkorpuszra, annak fonetikus átiratára, egy környezetfüggő címkézőre, a magyar nyelvre jellemző döntési fákhoz szükséges kérdések elkészítésére. A következő pontokban áttekintjük a magyar változat létrehozásának fontosabb lépéseit.

4.1. Beszédkorpusz előkészítése

A tanításhoz 600 mondatot használtunk, melyeket professzionális bemondótól vettünk fel, 16000 Hz-en újramintavételeztük, 16 bites felbontással. A mondatok tartalma időjárásjelentés volt, és összesen körülbelül 2 óra a hanganyag hossza. A mondatok fonetikus átiratát elkészítettük és a hanghatárokat bejelöltük automatikus módszerekkel [11].

4.2. Környezetfüggő címkézés

Annak érdekében, hogy a HMM-ek a legmegfelelőbb elemeket válasszák majd ki a beszédelőállítás során, számos fonetikai jellemzőt adunk meg. A jellemzőket minden egyes hangra kiszámoljuk. Az *1. táblázat* foglalja össze a legfontosabb jellemzőket.

1. táblázat
A környezetfüggő címkézéshez használt prozódiai jellemzők

(Megjegyzés: a szótagokat a szótagmagok alapján keressük, számoljuk és jelöljük, tehát nem a nyelvi szótagolási szabályokat vesszük figyelembe.)

Hangok	• Az aktuális hangot megelőző és követő két-két hang (kvinfón). A szüneteket is jelöljük.
Szótagmag	• Szótaghangsúlyok jelölése az aktuális/előző/következő szótagban. • A fonémák száma az aktuális/előző/következő szótagban. • A szótagok száma az előző/következő hangsúlyos szótagtól/szótagig. • A szótag magánhangzója.
Szó	• Szótagok száma az aktuális/előző/következő szóban. • Az aktuális szó pozíciója a mondatrészben (elölről és hátulról is számítva).
Mondatrész (két írásjel közötti szakasz)	• A szótagok és szavak száma az aktuális/előző/következő mondatrészben. • Az aktuális mondatrész pozíciója a mondatban (elölről és hátulról is számítva).
Mondat	• A szótagok száma az adott mondatban. • A szavak száma az adott mondatban. • A mondatrészek száma az adott mondatban.

A címkézést automatikusan végezzük, mely néhány esetben (pl. hangsúlyos szótagok meghatározása) hibás lehet. Ez azonban nem okoz jelentős problémát, hiszen a beszéd előállításakor is ugyanazt az algoritmust használjuk, így hibás címkézés esetén is a HMM következetesen fogja az adott hangoknak megfelelő paramétereket kiválasztani.

4.3. Döntési fák

A 4.2. pontban láthattuk, hogy számos környezetfüggő tulajdonság létezik, melyek összes lehetséges kombinációja óriási szám. Ha csupán a kvinfónok lehetséges változatait számoljuk meg, az is több mint 160 millió, de ezt a számot a többi környezetfüggő tulajdonság még exponenciálisan növeli. Ezért lehetetlen egy olyan, adott nyelvre jellemző beszédkorpuszt előállítani, melyben minden lehetséges kombináció szerepel.

Ezen probléma leküzdése érdekében be kellett vezetni a döntésifa-alapú klaszterezést [12,13]. Mivel a különböző tulajdonságok hatnak mind a spektrális, mind az alapfrekvencia paraméterekre és az állapotidőtartamokra is, ezért ezeket külön-külön kell klaszterezni. A *2. táblázat* mutatja, hogy milyen magyar nyelvre jellemző tulajdonságokat [14] használtunk fel a döntési fák építésekor.

Amennyiben a tanításból például kihagyjuk a mássalhangzók hosszára vonatkozó kérdéseket, akkor a HMM-ek elsősorban rövid mássalhangzókat fognak behelyettesíteni a hosszúak helyére is, hiszen nem klasztereztük ezeket külön és így az adatbázisban lényegesen többször szereplő rövid mássalhangzók kerülnek előtérbe.

4.4. Eredmények

Annak érdekében, hogy objektíven tudjuk értékelni a magyar nyelvű HMM alapú beszédszintézis minőségét, egy MOS (Mean Opinion Score) meghallgatásos tesztet készítettünk el. A tesztben három rendszer vett részt, egy triád-alapú, egy korpuszos és a HMM-alapú szövegfelolvasó.

Fonémák	• Magánhangzó/mássalhangzó • Rövid/hosszú • Zárhang/réshang/zár-rés hang/pergő hang/nazálisok • Képzés helye • Nyelvállás • Ajakállás (kerekített, kerekítetlen)
Szótag	• Hangsúlyos / hangsúlytalan • Az adott szótagra vonatkozó számszerű adatok (lásd 1. táblázat)
Szó	• Az adott szóra vonatkozó számszerű adatok (1. táblázat)
Mondatrész	• Az adott mondatrészre vonatkozó számszerű adatok (1. táblázat)
Mondat	• Az adott mondatra vonatkozó számszerű adatok (1. táblázat)

2. táblázat
A döntési fák építéséhez használt jellemzők

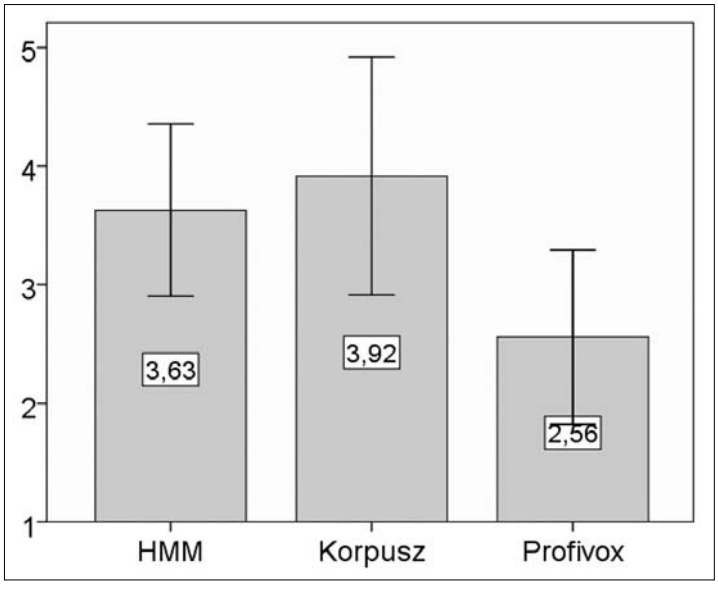
A teszt elején minden rendszertől 3-3 mondatot játszottunk le véletlenszerűen, amelyeket a tesztalanyok még nem értékelhettek. Ez azt a célt szolgálta, hogy az alanyok hozzászokjanak a mesterséges hangokhoz, és hallják előre, hogy nagyjából milyen minőségre számíthatnak.

Ezután minden rendszer mintáiból 29 mondatot játszottunk le, minden tesztalany esetén más-más sorrendben, így zárva ki az esetleges „memória hatásokat" [15]. A tesztmondatok tartalma időjárásjelentés volt. A triád-alapú rendszer kötetlen témakör szintézisére készült. A HMM rendszer időjárásjelentés-tartalmú mondatokkal volt tanítva, illetve a korpuszos rendszer adatbázisa is időjárásjelentéseket tartalmazott. Minden rendszerrel ugyanazt a 29 mondatot generáltuk, de egyik rendszer esetén sem szerepeltek ezek a mondatok az adatbázisban. A tesztalanyok a mondatokat egytől ötig értékelhették (egy volt a legrosszabb, öt a legjobb).

A meghallgatásos tesztet 12 tesztalany végezte el. Az eredményt a *4. ábra* mutatja.

A teszt során a triádos rendszer 2,56 pontot, a HMM alapú szövegfelolvasó 3,63 pontot, a korpuszos rendszer pedig 3,9 pontot kapott átlagban. Ugyanebben a sorrendben a szórásuk 0,73, 1 és 0,73 volt.

4. ábra
A MOS meghallgatásos teszt eredménye (az oszlop magassága az átlagértéket, a függőleges vonal a szórást jelöli)



Fontos kiemelnünk, hogy a korpuszos rendszer ugyan jobb értékeket ért el a HMM alapú szövegfelolvasó rendszernél, de míg az első azonos minőségben csak témakör- (domén-) specifikus mondatokat tud felolvasni, a második általános témájú mondatokat is közel azonos minőségben olvas fel. Továbbá a korpuszos rendszer adatbázisa közel 11 órányi hanganyagot tartalmaz, míg a HMM-ek tanításához elegendő volt 1,5 órányi hanganyag és tanítás után a HMM szövegfelolvasó esetén az adatbázis mérete 2 megabájt alatt marad (szemben a korpuszos rendszer több, mint egy gigabájtos adatbázisával).

A triád-alapú rendszer általános témakörök lefedésére készült, semmilyen témakör-specifikus információ nem került bele. Ez is magyarázhatja az alacsonyabb értékelést. Az eredmények abszolút értéke kevésbé mérvadó, inkább az egymáshoz viszonyított arányok hordoznak érdemi információt.

5. Jövőbeli tervek

Jelen cikk a magyar nyelvű, HMM alapú mesterséges beszédkeltés első változatát ismertette. A jövőben számos továbbfejlesztési irányt tűztünk ki célul, melyek közül első lépésként az adaptív tanításhoz szeretnénk további beszédkorpuszokat rögzíteni, így érve el természetesebb hangzást, továbbá ezáltal lehetőségünk nyílik kis (5-8 perces) adatbázisok segítségével új beszédhangokat és érzelmeket betanítani a rendszerrel.

A kis adatbázisméret előnyei és a jó minőségű beszédhang miatt szeretnénk a rendszert mobil eszközökön is megvalósítani. Ennek érdekében optimalizálni fogjuk a hts_engine-t mobil eszközökre. Lehetséges, hogy a rendszert alapvetően módosítani kell ahhoz, hogy közel valósidejű rendszert kapjunk.

6. Összefoglalás

A cikkben bemutattuk a rejtett Markov-modell alapú szintézis működésének az elvét, a magyar változat létrehozásának a lépéseit és az első magyar HMM-alapú beszédkeltéssel kapcsolatos meghallgatásos teszt eredményeit.

A HMM-alapú szövegfelolvasó rendszerek igazi előnye, hogy kis adatbázisméretek mellett képesek jó minőségű beszédhangot előállítani, illetve könnyebben lehet a hangkaraktert megváltoztatni, érzelmeket kifejezni. Célunk, hogy ipari alkalmazásokban is használható magyar nyelven beszélő szövegfelolvasó rendszerré fejlesszük tovább a jelenlegi változatot.

Köszönetnyilvánítás

Ezúton szeretnénk köszönetet mondani a szubjektív kiértékelésben résztvevő tesztelőknek. Külön köszönet illeti Bartalis Mátyást a web-es tesztfelület elkészítéséért és Mihajlik Pétert a magyar nyelvű beszédfelismerő eszközök használatához nyújtott segítségéért. A kutatást részben támogatta az NKTH a NAP projekt keretében (OMFB-00736/2005).

A szerzőkről

Németh Géza 1983-ban végzett a BME Villamosmérnöki Karán, 1985-ben pedig szakmérnöki diplomát szerzett. 1985-87 között a BEAG Elektroakusztikai Gyárban fejlesztőmérnökként dolgozott, 1987-től a BME Távközlési és Médiainformatikai Tanszékén oktat (Méréstechnika, Kommunikációs rendszerek, Híradástechnika, A jelfeldolgozás elemei, Távközlés, Távközlésmenedzselés, Beszédinformációs rendszerek). Jelenleg a tanszék beszédtechnolólógiai laboratóriumát is vezeti. Irányító szerepet tölt be a beszédkutatási eredmények gyakorlatba való átültetésében, számos gyakorlati alkalmazást az ő vezetésével fejlesztettek ki.

Tóth Bálint Pál 2005-ben kitüntetett diplomával végzett a BME Villamosmérnöki Karán Távközlési és telematikai szakirányon. Ph.D. tanulmányait rögtön a diplomázás után elkezdte beszédszintézis és multimodális felhasználói felületek témakörben. A beszédszintézis területén elsősorban a rejtett Markov-modell alapú szövegfelolvasással foglalkozik, míg a multimodális felhasználói felületek mobil környezetben való alkalmazási lehetőségeit vizsgálja.

Irodalom

[1] T. Masuko, K. Tokuda, T. Kobayashi, S. Imai, "Voice characteristics conversion for HMM-based speech synthesis system," Proc. ICASSP, 1997, pp.1611–1614.

[2] M. Tamura, T. Masuko, K. Tokuda, T. Kobayashi, "Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR," Proc. ICASSP, 2001, pp.805–808.

[3] T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, T. Kitamura, "Speaker interpolation in HMM-based speech synthesis system," Proc. Eurospeech, 1997, pp.2523–2526.

[4] M. Tachibana, J. Yamagishi, T. Masuko, T. Kobayashi, "Speech synthesis with various emotional expressions and speaking styles by style interpolation and morphing," IEICE Trans. Inf. & Syst., Vol. E88-D, 2005, No.11, pp.2484–2491.

[5] S. Krstulovic, A. Hunecke, M. Schroeder, "An HMM-Based Speech Synthesis System applied to German and its Adaptation to a Limited Set of Expressive Football Announcements," Proc. of Interspeech, 2007.

[6] Mihajlik P., Fegyó T., Németh B., Tüske Z., Trón V., "Towards Automatic Transcription of Large Spoken Archives in Agglutinating Languages: Hungarian ASR for the MALACH Project," In: Matousek V, Mautner P (ed.) Text, Speech and Dialogue: 10th Int. Conference, TSD 2007, Pilsen, Czech Republic, Sept. 2007, Proc., Berlin; Heidelberg: Springer, Lectures Notes in Computer Sciences, 2007, pp.342–350. (Lecture Notes in Artificial Intelligence; 4629.)

[7] Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition," Proc. of the IEEE, 77 (2), Febr. 1989, pp.257–286.

[8] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," Proc. ICASSP, 1983, pp.93–96.

[9] R. Maia, T. Toda, H. Zen, Y. Nankaku, K. Tokuda, "A trainable excitation model for HMM-based speech synthesis," Proc. Interspeech, Aug. 2007, pp.1909–1912.

[10] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A.W. Black, K. Tokuda, "The HMM-based speech synthesis system v.2.0", Proc. of ISCA SSW6, Bonn, Germany, Aug. 2007.

[11] Mihajlik, P. Révész, T. Tatai, P., "Phonetic transcription in automatic speech recognition," In: Acta Linguistica Hung., 2003, Vol. 49; No. 3/4, pp.407–425.

[12] J.J. Odell, "The Use of Context in Large Vocabulary Speech Recognition," PhD dissertation, Cambridge University, 1995.

[13] K. Shinoda, T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," J. Acoust. Soc. Jpn.(E), Vol. 21, No.2, 2000. pp.79–86.

[14] Gósy M., Fonetika, a beszéd tudománya. Budapest, Osiris Kiadó, 2004.

[15] Jan P.H. van Santen, Perceptual experiments for diagnostic testing of text-to-speech systems, Computer Speech and Language, 1993, pp.49–100.

Szövegfelolvasó természetességének növelése

CSAPÓ TAMÁS GÁBOR, NÉMETH GÉZA, FÉK MÁRK

BME Távközlési és Médiainformatikai Tanszék
{csapo, nemeth, fek}@tmit.bme.hu

Lektorált

Kulcsszavak: beszédszintézis, prozódiai modell, prozódiai változatosság, F_0 másolás

A cikk röviden bemutatja a jelenlegi beszédszintézis-rendszerekben alkalmazott prozódiai modelleket, illetve egyik gyengéjüket: az emberihez hasonló változatos prozódia modellezésének hiányát. Részletesen ismertetjük az általunk kidolgozott módszert a hosszabb időtartamú szintetizált beszéd monotonitásának csökkentésére. Egy természetes mondatokból álló beszédkorpuszt felhasználva, az alapfrekvencia-menet másolásával valósítottuk ezt meg. Végül bemutatjuk, hogyan történt a módszerünkkel előállított mondatok minőségének értékelése.

1. Bevezetés

A beszédszintézis-rendszerek minőségét annak alapján ítélik meg, hogy az általuk keltett beszéd mennyire hasonlít az emberi beszédre. A jelenlegi rendszerek többsége egy szabályrendszer segítségével a nyelvi elvárásoknak megfelelően adott szöveghez mindig azonos prozódiát rendel. Ugyanakkor ahhoz, hogy a gépi megoldás ne tűnjön monotonnak, az emberhez hasonlóan változatosságot kell létrehozni, azaz ugyanazt a mondatot nem mindig ugyanúgy kell bemondania a rendszernek. Vizsgálataink célja, hogy egy nagyméretű beszédkorpuszt elemezve megtudjuk, a prozódiai változatosság milyen mértékben valósítható meg a BME Távközlési és Médiainformatikai Tanszéken fejlesztett korábbi ProfiVox rendszer kiegészítésével [1].

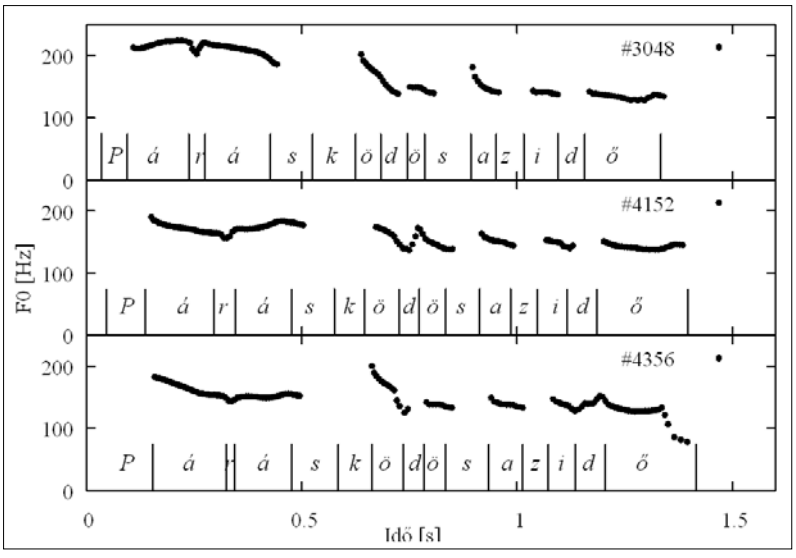
Cikkünk fő témája a szövegfelolvasó rendszerek egyik legfontosabb komponensének, a prozódia előállításának vizsgálata. A prozódia tervezésére sokféle modell ismert, úgymint a leíró jellegű, szabályalapú, gépitanulás-alapú, illetve szuperpozíciós modellek. A Profivox beszédszintetizátor első változata szabályalapú és szuperpozíciós [2], azaz a bemeneti szöveghez tartozó prozódiát ember által definiált szabályok alapján hozza létre több szinten. A szintek modellezése külön-külön történik, először meghatározva a mondatdallamot (emelkedő, egyenletes, eső), utána a szó- vagy szótagszintű hangsúlyokat, végül a mikrointonációs változásokat.

Számos olyan módszer ismert a szakirodalomban, melyek a prozódiát valamilyen természetes beszédből álló korpusz alapján hozzák létre [3,4,5]. Az emberihez hasonló dallammenet létrehozása azzal garantálható, hogy a szintetizálandó mondat alapfrekvencia-menetét az adatbázisból vett kisebb-nagyobb elemek (például szótag, szó) segítségével határozzák meg.

Kutatásunk során jelentős kezdeti eredményeket értünk el a beszédszintetizátorok prozódiájának változatosabbá és természetesebbé tétele területén nagyméretű természetes beszédkorpusz felhasználásával [6]. Munkánkban a Profivox magyar nyelvű diád-triád alapú beszédszintetizátort alkalmaztuk [1]. A jelen cikkben ismertetjük a prozódia változatosabbá tételére kidolgozott módszert, majd bemutatjuk, hogyan történt a módszerünkkel előállított mondatok minőségének értékelése.

2. Prozódiai változatosság

Az emberi beszédben a prozódia rendkívül változékony jellemző. Egy-egy mondatot még akarattal sem tudunk többször ugyanúgy elmondani, a mindennapi beszédben pedig óriási különbségek tapasztalhatóak dallam, hangsúly és ritmus terén is, ahogy ezt az *1. ábra* mutatja. Az ábrán a *„Párás, ködös az idő."* mondat három ku-



1. ábra
Prozódiai változatosság az emberi beszédben. (Mondat: „Párás, ködös az idő.")

lönböző kiejtési módját láthatjuk. A három változat hasonló, de mégis észrevehető különbség van közöttük az alapfrekvencia-menetben (F_0) és a hangok időtartamában (függőleges vonalak).

A legtöbb beszédszintetizátor rendszer ezzel szemben determinisztikusan állítja elő a prozódiát, azaz egy-egy bemeneti szöveghez a beszédszintetizátor futása során mindig ugyanaz a dallam tartozik. Ez sokszor ismétlődő, monoton dallamminták túlzott előfordulásához vezet, ami zavaró a szintetizált beszédben. A prozódiaminták ismétlődése azért fordulhat elő a szövegfelolvasó rendszerekben, mert például egy elemkiválasztásos szintetizátor mindig a legjobb prozódiát próbálja egy-egy mondathoz rendelni. Így az emberi beszéd változatossága (ami az 1. ábrán is látható) lecserélődik a legjobb, leggyakoribb mintára. Ez viszont az emberi fül számára, ami a változékonysághoz szokott, könnyen felismerhető. Beszédünk stílusát sokszor szándékosan is variáljuk, ha különböző dolgokat akarunk kifejezni. Sokszor éppen azért használunk más-más prozódiát, hogy ne tűnjön monotonnak beszédünk. Éppen ezért a beszédszintetizátornak sem szükséges mindig a legjobb prozódiát megtalálnia, inkább egy elfogadható tartományt érdemes definiálni, amin belül megfelelőnek tartjuk a minőséget.

Chu és társai [7] bemutatnak egy szótag- és szó-alapon működő beszédszintetizátor rendszert, ami megkísérli a prozódiai változatosság létrehozását. A módszer célja, hogy ne mindig csak a legjobb lehetőséget keresse meg, hanem a rossz lehetőségek kihagyásával a maradékból véletlenszerűen válasszon. A megközelítés sikeresnek bizonyult és használható az angol, illetve mandarin nyelv szintézisére.

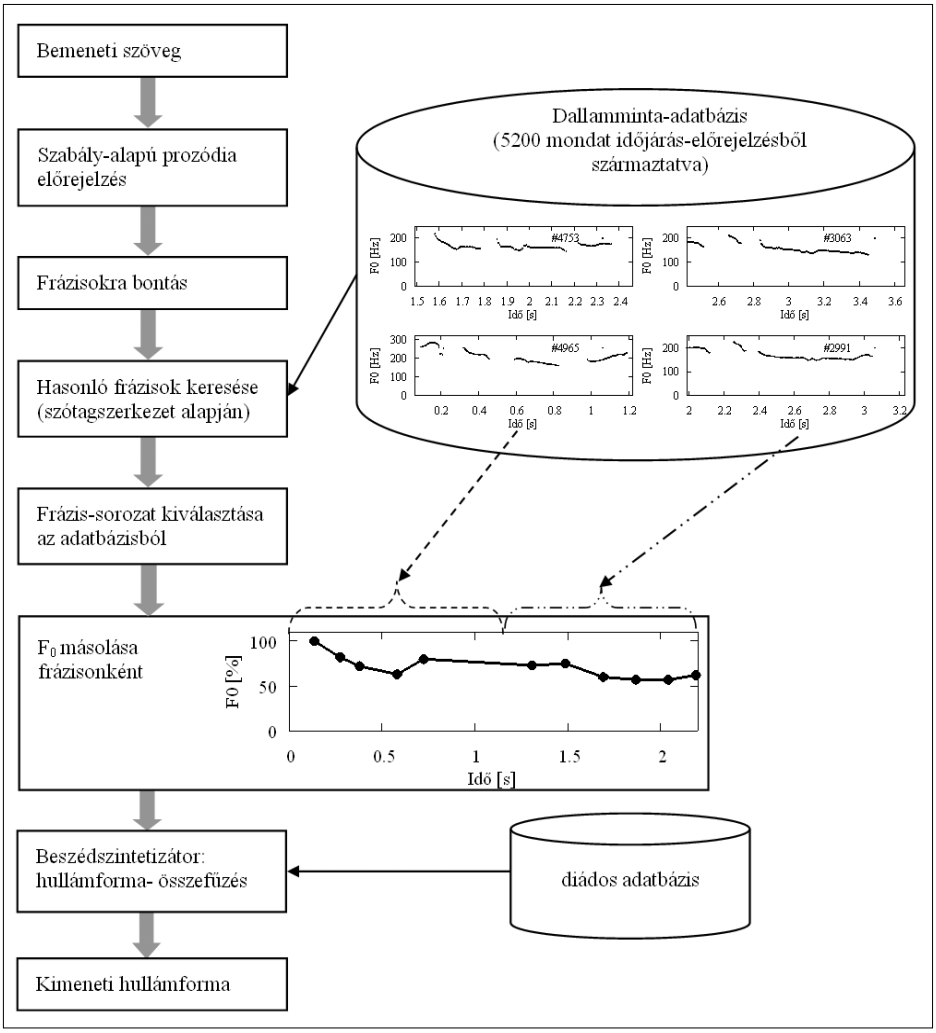
3. Dallammásolás frázisok alapján

Azt, hogy a szövegfelolvasó egy-egy bemeneti mondatához ne mindig ugyanolyan prozódiájú mondatot szintetizáljunk, úgy valósítjuk meg, ha a bemeneti szöveghez többféle dallammenetet tudunk generálni és ezek közül a rendszer szintéziskor egyet véletlenszerűen választ ki. Ekkor ugyanis csökken a monotonitás, hiszen nem-determinisztikussá válik a mondatokhoz történő dallammenet-hozzárendelés.

A prozódiai változatosság eléréséhez az szükséges, hogy egy-egy mondathoz legalább 3-4 lehetséges dallammenetet tudjunk rendelni. Kutatásunk során egy 5200 mondatból álló, magyar nyelvű beszédkorpuszon [8] végeztünk kísérleteket. A prozódia tervezését korpusz alapon oldottuk meg, a természetes mondatok dallamát lemásolva. A dallam szöveghez rendelése során szótagszerkezet (az egyes szavak szótagszáma a mondatban) alapján keresünk F_0-mintákat a korpuszban. Az, hogy egy mondathoz hány teljes dallammintát tudunk előállítani, függ attól, hogy mekkora F_0 másolási egységekkel dolgozunk és mekkora a beszéddallam-adatbázis mérete. Beszédkorpuszunk 5200 időjárás-előrejelzés témájú, az átlagos beszédhez képest hosszú mondatból áll. Az F_0 egységek méretét első kísérleteinkben teljes mondatra, majd a rövidebb frázisra (beszéd során egy levegővétellel kimondott egység) választottuk.

Ahhoz, hogy a hosszabb, több frázisból álló mondatokhoz is találhassunk prozódia-mintát, a mondatok felbontására volt szükség. Egy frázishoz nagyobb valószínűséggel lehet találni egyező szótagszerkezetű mintát, mint a teljes mondathoz. Ha például egy szintetizálandó mondat három frázisból áll *(„Csütörtökön rendkívül melegre, magas hőmérsékleti értékekre számíthatunk, főleg a déli térségekben.”)*, egyben kezelve nehezen találhatunk hozzá szerkezetileg hasonlót, míg frázisokra bontva a keresés egyszerűbbé válik.

2. ábra Módszerünk működési folyamata



A beszédkorpusz mondatait tehát automatikus módszerrel bontottuk fel frázisokra a szöveges átírásuk alapján. Ezen frázisokat sorszámukkal és néhány paraméterükkel (szótagszerkezet, hangsúlyszerkezet, pozíció a mondaton belül, F_0-menet, átlagos F_0 érték) jellemeztük. Összesen 13415 frázisra bontottuk így a beszédkorpuszt, létrehozva ezzel egy dallamminta-adatbázist. Átlagosan egy mondat 2,57 frázisból, egy frázis pedig 13,78 szótagból áll az egész korpuszt figyelembe véve.

A hangsúlyszerkezetet, a pozíciót és az átlagos F_0 értéket azért tároltuk el, hogy a prozódia létrehozásakor a frázisok kiválasztásában ezeket is figyelembe lehessen venni. A prozódiaminták kiválasztásakor és egymás után fűzésekor tehát különböző „kényszerek" segítségével biztosíthatjuk a természeteshez hasonló dallammenetet (például hangsúlyok figyelembe vétele a szótagszerkezet mellett). Ezek segítségével a dallammásolás hatékonysága és természetessége tovább növelhető.

A bemeneti szöveghez a módszer segítségével teljesen automatikusan történik meg a teljes mondatra vonatkozó dallammenet meghatározása.

Módszerünk működésének folyamata a *2. ábrán* látható. A bemeneti szöveg alapján a hangidőtartamok és az intenzitás meghatározása a Profivox korábbi modellje alapján, szabályalapon történik, ezt tehát változatlanul hagytuk. A dallam meghatározása során először frázisokra bontjuk a teljes bemeneti mondatot, majd mindegyikhez keresünk prozódiamintát az adatbázisból. A keresés szótagszerkezet alapján történik. A lehetséges mintasorozatok közül egyet véletlenszerűen kiválaszt a rendszer (bizonyos kényszerek figyelembe vételével), és megtörténik az F_0-szakaszok másolása frázisonként. A véletlen választás miatt ritkábbá válnak az ismétlődő dallamminták, ami javítja a szintetizált beszéd minőségét. A módszer utolsó lépéseként egy diádos adatbázis segítségével történik meg a hullámforma összefűzés, vagyis a szintetizált beszéd létrehozása.

4. Teszt és eredmények

Kiválasztottunk a beszédkorpuszból 10 időjárás-előrejelzés témájú mondatot, és ezeket szöveges átírásuk alapján újraszintetizáltuk az itt bemutatott módszer segítségével, különféle dallammenetekkel.

A változatok között szerepelt mondatonként egy-egy olyan változat, ami a Profivox korábbi, szabályalapú dallammodelljével készült, illetve két-három olyan variáns is, amelynek dallama frázis alapján történő másolással jött létre.

A létrehozott mondatok tesztelését a BME Távközlési és Médiainformatikai Tanszéken kifejlesztett webes tesztelő rendszerben végeztük. A mondatokból mondatpárokat hoztunk létre, melyek egy-egy mondat két változatát tartalmazták. Összesen 37 ilyen mondatpár készült el. A tesztet elvégzők feladata az volt, hogy eldöntsék, a mondatpár első vagy második tagját tartják természetesebbnek, vagy nem tudnak különbséget tenni a két változat között. Egy-egy mondatot többször is meghallgathattak, hogy döntésüket könnyebben meg tudják hozni. A mondatok lejátszása véletlen sorrendben történt.

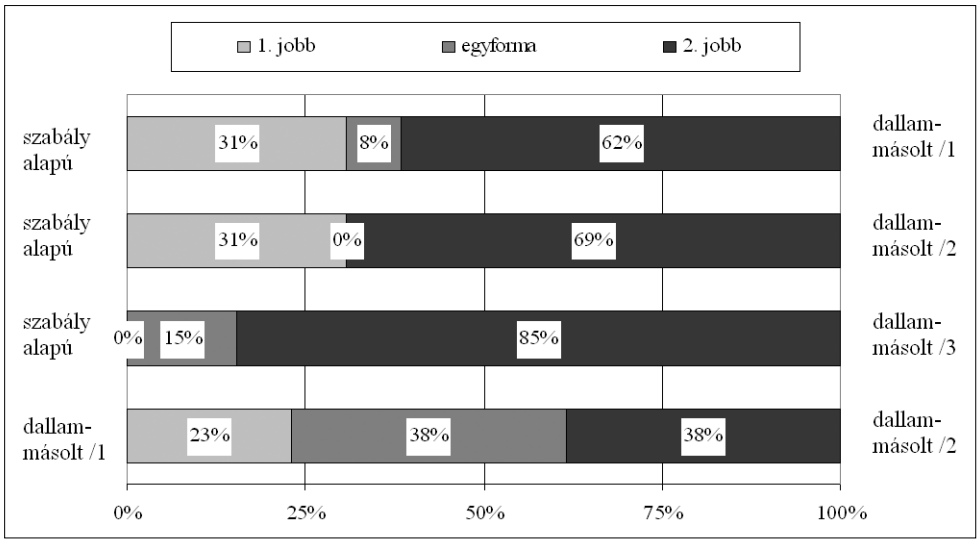
A tesztelőknek a *http://speechlab.tmit.bme.hu/csapo* oldalt meglátogatva egy rövid ismertetőt kellett elolvasniuk a teszt menetéről, majd néhány információt kértünk be róluk (becenév, életkor, nem). Ezután megkezdődhetett a mondatpárok meghallgatása. A szintetizált hangok meghallgatása után a tesztelők megjegyzést is írhattak észrevételeikről.

A mondatpárok meghallgatását 13 tesztelő végezte el. A tesztelők mindannyian ép hallású, magyar anyanyelvű emberek voltak, a 20-64 év közötti korosztályból. Egy részük a témához értő tanszéki munkatárs volt, míg a többiek az egyetemi hallgatók köréből kerültek ki. A rendszer rögzítette a teszt elkezdésének és befejezésének időpontját, így azt a tesztelőt kizártuk az eredmények kiértékeléséből, aki 10 percnél rövidebb idő alatt végezte el a tesztet (hiszen ennyi idő minimálisan szükséges lett volna az összes mondat meghallgatásához). A teszt átlagos meghallgatási ideje 19 perc volt.

A teszt kiértékeléséből az derült ki, hogy a tesztelők az esetek többségében a adatbázisbeli frázisok másolásával létrehozott dallamot preferálták a szabályalapú változathoz képest.

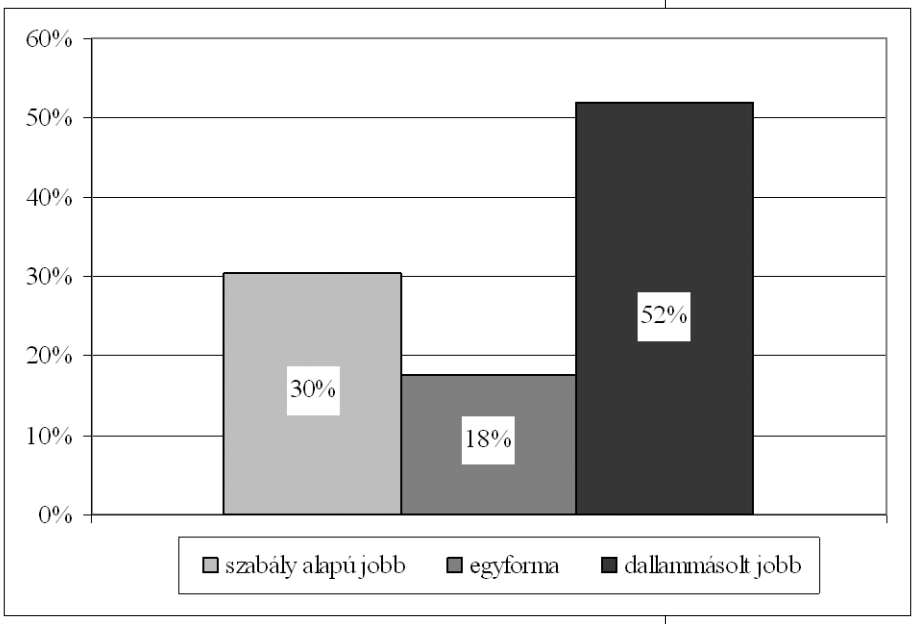
A *3. ábrán* egy mondat négy különböző változatának (egy szabályalapú és három dallammásolt) összehasonlítását láthatjuk, soronként egy mondatpár eredményeit ábrázolva. Észrevehető, hogy a dallammásolt változatokat a tesztelők természetesebbnek érezték,

3. ábra
Egy tesztbeli mondat változatainak összehasonlítása



mint a szabályalapú változatot (első három sor). A két különböző dallamú, új módszerrel létrehozott mondat összehasonlítása (negyedik sor) pedig azt mutatja, hogy a tesztelők mindkét változatot elfogadják, vagyis azok nagyjából egyforma minőségűek.

Összességében elmondhatjuk, hogy a 10 mondatból 5 esetben egyértelműen az új, frázisok alapján működő F_0-másolási módszer volt jobb, 3 esetben nem lehetett dönteni a tesztelők véleménye alapján és 2 mondat esetében a szabályalapú megoldás minőségét értékelték jobbnak.



4. ábra Összesített eredmény

Az összesített eredmény, ami a *4. ábrán* látható, azt mutatja, hogy a tesztelők a dallammásolás módszerét részesítették előnyben. A tesztelők megjegyzései közül fontos kiemelni, hogy egyesek nagyon zavarónak tartották a mondat végi dallamemelést, mert ott mindenképpen a legmélyebb hangot várja a hallgató. Mások szerint a mondatok meglehetősen hosszúak voltak, így nagyon kellett koncentrálni, hogy el lehessen dönteni, melyik a természetesebb közülük. A későbbiekben tehát figyelnünk kell arra, hogy összehasonlítási kísérleteinkben rövidebb mondatokat vizsgáljunk.

5. Összefoglalás

A cikkben ismertettük a mai beszédszintetizátor rendszerek egyik hiányosságát: azt, hogy nem modellezik az emberi beszéd változatosságát. Áttekintettük munkánkat és ennek eredményét. Automatikussá tettük a prozódia másolását, nagyméretű beszédkorpuszban vizsgáltuk módszerünk eredményességét.

A módszerünkkel létrehozott mondatok minőségét egy webes tesztben ellenőriztük. Mondatpáronként kellett a tesztelőknek értékelniük a különböző dallamváltozatú mondatokat. Az eredmények kiértékeléséből kiderült, hogy a dallammásolással létrehozott szintetizált mondatok az esetek többségében jobbak a szabályalapú változatoknál.

Az általunk kidolgozott módszer segítségével természetesebbé tehető a szövegfelolvasók által létrehozott prozódia. Ez az előny számos gyakorlati alkalmazásban használható, mint például SMS-, e-mail-, könyvfelolvasó, vagy telefonos tudakozó. A változatosabb prozódia főleg hosszú szövegek felolvasása esetén előnyös, hiszen ekkor zavaró a beszédszintetizátor monotonitása. A fő cél tehát az, hogy a módszert a Profivox beszédszintetizátorba beépítve szélesebb körben használni lehessen azt.

Érdekes lenne megvizsgálni, hogy más beszédadatbázissal milyen eredményeket tudunk elérni. Olyan korpuszt célszerű választani, amiben rövidebb mondatok vannak, amelyek jobban közelítik az általános beszéd mondathosszát. Azt az irányt is érdemes megvizsgálni, hogyan lehetne a prozódia többi komponensét (elsősorban az időtartamokat) is korpusz alapján létrehozni.

Jelen dolgozat az Interspeech 2007 konferencián bemutatott cikk [6] kibővített változata, amely az azóta elért eredményeket is tartalmazza.

Köszönetnyilvánítás

Ezúton szeretnénk köszönetet mondani a BME Távközlési és Médiainformatikai Tanszék Beszédtechnológiai Laboratóriuma munkatársainak a tanácsokért, a meghallgatásos kísérletben résztvevőknek a teszt kitöltéséért, valamint Bartalis István Mátyásnak a webes tesztelő rendszer beállításáért.
A kutatást az NKTH részben támogatta a NAP (OMFB-00736/2005) és az NKFP (NKFP 2/034/2004) programok keretében.

A szerzőkről

Csapó Tamás Gábor 2008-ban fogja megszerezni informatikai diplomáját a Budapesti Műszaki és Gazdaságtudományi Egyetem Távközlési és Médiainformatikai Tanszékén. Kutatási témája a beszédszintézis, ezen belül a szövegfelolvasók által létrehozott mesterséges beszéd természetesebbé tétele. Ennek során több publikációja született, többek között OTDK 1. helyezést ért el. Az utolsó tanévben köztársasági ösztöndíjban részesült kiemelkedő eredményeiért. Tanulmányait a BME Informatikai Tudományok Doktori Iskolájában tervezi folytatni.

Németh Géza 1983-ban végzett a BME Villamosmérnöki Karán, 1985-ben pedig szakmérnöki diplomát szerzett. 1985-87 között a BEAG Elektroakusztikai Gyárban fejlesztőmérnökként dolgozott, 1987-től a BME Távközlési és Médiainformatikai Tanszékén oktat. Jelenleg a tanszék beszédtechnolólógiai laboratóriumát is vezeti. Irányító szerepet tölt be a beszédkutatási eredmények gyakorlatba való átültetésében, számos gyakorlati alkalmazást az ő vezetésével fejlesztettek ki.

Fék Márk 1997-ben végzett a Budapesti Műszaki és Gazdaságtudományi Egyetem Villamosmérnöki és Informatikai karán, Műszaki Informatika Szakon. 1997-2001 között francia-magyar közös doktori képzésen vett részt a BME-n és a francia ENST-Bretagne-on. Doktori disszertációját a beszéd- és audio-jelek tömörítése témakörében 2006-ban védte meg. 2001-től a BME Távközlési és Médiainformatikai Tanszékén magyar nyelvű beszédszintézissel foglalkozik. Főbb kutatási területei a korpusz alapú beszédszintézis és az érzelemszintézis.

Irodalom

[1] Olaszy, G., Németh, G., Olaszi, P., Kiss, G., Gordos, G., "PROFIVOX – A Hungarian Professional TTS System for Telecommunications Applications," Int. Journal of Speech Tech., Vol. 3, Numbers 3/4, Dec. 2000, pp.201–216.

[2] Olaszy, G., Németh, G., Olaszi, P., "Automatic Prosody Generation – a Model for Hungarian," Proc. Eurospeech 2001, Vol. 1, pp.525–528.

[3] Dong, M., Lua, K.T., "An Example-based Approach for Prosody Generation in Chinese Speech Synthesis," Proc. ISCSLP 2000, Beijing, pp.303–307.

[4] Raux, A., Black, A., "A Unit Selection Approach to F_0 Modeling and its Application to Emphasis," Proc. ASRU 2003, pp.700–705.

[5] Van Santen, J., Kain, A., Klabbers, E., Mishra, T., "Synthesis of prosody using Multilevel Unit Sequences," Speech Communication, Vol. 46, Issues 3-4, pp. 365–375, 2005.

[6] Németh, G., Fék, M., Csapó, T.G., "Increasing Prosodic Variability of Text-To-Speech Synthesizers," Proc. Interspeech 2007, pp.474–477.

[7] Chu, M., Zhao, Y., Chang, E., "Modeling stylized invariance and local variability of prosody in text-to-speech synhtesis," Speech Communication, Vol. 48, 2006, pp.716–726.

[8] Fék Márk, Pesti Péter, Németh Géza, Zainkó Csaba, Generációváltás a beszédszintézisben. In: Híradástechnika, LXI. évf. 2006/3, pp.21–30.

Pollák-Virág díjasok

A Pollák-Virág díjbizottság javaslata alapján a Híradástechnika folyóirat 2007. évi cikkei közül az alábbiak kaptak Pollák-Virág díjat:

Kutatási cikkek kategória

Mitcsenkov Attila, Meskó Diána, Cinkler Tibor:
Forgalomhoz alkalmazkodó védelmi módszerek (2007/2. szám)

Nagy Lajos:
Determinisztikus beltéri hullámterjedési modellek (2007/3. szám)

Kőrösi Attila, Székely Balázs, Lukovszki Csaba, Dang Dihn Trang:
DSL hozzáférési hálózatokban alkalmazott csomagütemezők sorbanállási modellezése és analízise teljes és részleges visszautasítás esetére (2007/4. szám)

Perényi Marcell, Soproni Péter, Cinkler Tibor:
Multicast fák rendszeres újrakonfigurálása többrétegű optikai hálózatokban (2007/8. szám)

Szentpáli Béla:
Mikrohullámú térmérő szondák (2007/11. szám)

Áttekintő cikkek kategória

Babics Emil, Horváth A. Róbert, Meskó Örs:
Flexibilis leágazó és kapcsoló eszközök a DWDM hálózatokba (2007/6. szám)

A díjazottaknak gratulálunk!

Magyar nyelvű, kötött témájú korpusz-alapú beszédszintézis és a kötetlenség felé vezető út vizsgálata

ZAINKÓ CSABA

BME Távközlési és Médiainformatikai Tanszék
zainko@tmit.bme.hu

Lektorált

Kulcsszavak: korpusz-alapú beszédszintézis, beszédadatbázisok, prozódiai modul

A beszédszintetizátorok között a korpusz-alapú szintetizátorral lehet jelenleg a legjobb minőségű beszédet előállítani. Ennek ára, hogy csak adott témájú szövegek szintetizálását tudja ilyen minőségben garantálni. A cikk azt tárgyalja, hogy ha egy ilyen kötött témájú korpuszos szintetizátort kívánunk kötetlen szövegekre kibővíteni, akkor annak milyen lehetőségei és korlátai vannak. A vizsgálat során a szintetizátor beszédadatbázisát elemeztük és megvizsgáltuk, hogy elegendően változatos-e tetszőleges szöveghez, illetve megfelelő számú elemet tartalmaz-e a jó minőséghez. Végül a szintetizált mondatokat egy meghallgatásos teszt keretében értékeltettük tesztelőkkel.

1. Bevezetés

A korpusz-szintetizátorokat általában meghatározott témájú szövegek szintetizálására fejlesztik (például időjárásjelentés, menetrendi tájékoztató, árlista felolvasó) [1]. A szintetizátor egy válogató algoritmusból és a hozzá tartozó beszédadatbázisból áll. Egy új témakörre való fejlesztés során általában csak a beszédadatbázist kell elkészíteni, mivel a szintetizátor válogató algoritmusa már megfelelően tesztelt, jól válogat. A munka nagy részét ebben az esetben az adatbázis elkészítése jelenti, azaz a megadott témájú szövegek felolvasása és előkészítése a szintézishez (tisztítás, címkézés, zöngés-zöngétlen határok bejelölése stb.). Ezek után a szintetizátor az adott témában tetszőleges mondatokat képes beszéddé alakítani, amely a technológiából adódóan közel emberi minőségű.

A kísérletben megvizsgáljuk, hogy a BME-TMIT-en készített, kötött tematikára készült beszédadatbázisok összeépítésével (egyazon bemondó hangjára) milyen minőségben lehet tetszőleges tartalmú mondatokat szintetizálni. A kutatás irányt ad arra is, hogy a korpuszos technológiánál milyen problémákkal kell számolni, ha kötetlen, általános beszédszintézist kívánunk megcélozni.

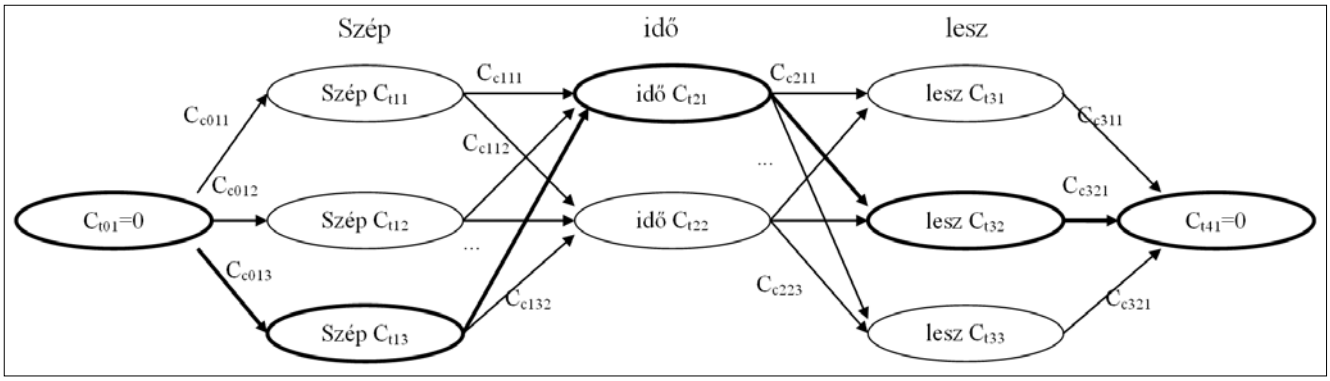
Az első részben bemutatjuk a szintetizátor működését, majd a beszédadatbázist vizsgáljuk meg, hogy milyen a mennyiségi és a minőségi összetétele. Alapvetően ez határozza meg, hogy milyen mondatok szintetizálására alkalmas a rendszer. A beszédadatbázis részletes elemzése után bemutatjuk, hogy milyen kísérleti mondatokat állítottunk elő és azokat a tesztelők hogyan értékelték. Az utolsó részben végül megvizsgáljuk a hangminőség javításának lehetőségeit.

2. Kötött témakörre fejlesztett korpusz-alapú beszédszintetizátor működése

A korpusz-alapú, elemkiválasztásos szintetizátor – továbbiakban korpuszos szintetizátor – egy olyan beszédgenerátor, amely nagy mennyiségű előre rögzített beszédből (beszédkorpuszból) válogatja ki a megfelelő elemeket és állítja elő ezek felhasználásával a szintetizált beszédet. A működés menetét az *1. ábrán* látjuk.

A példamondat a következő: *„Szép idő lesz”*. A szintetizátor a beszédkorpuszból válogat, meghatározza, hogy melyek azok a beszédrészletek (főleg szavak), amelyek felhasználhatók a mondat előállításához. Ezeket az ábrán a szavak alatt található ellipszisek jelölik,

1. ábra Példa az elemek kiválasztására és költségeire



amelyek most szó-méretű elemek, de lehetnek kisebbek is. A talált jelölteket egy mérőszámmal (célegyezési – target – költséggel, az ábrán C_{txy}) látja el, amely meghatározza, hogy mennyire alkalmas az adott elem a keresett pozícióra. A költség egyfajta büntetés, minél nagyobb, annál kevésbé alkalmas az adott helyre. Az egymás melletti pozícióra kiválasztott jelöltek között is kiszámol egy költséget a rendszer (összefűzési – concatenation – költséget, az ábrán C_{cxyz}), amely megadja, hogy mennyire illeszkedik a két elem egymáshoz. Itt is annál nagyobb a költség, minél rosszabbul illeszkedik a két elem. A végső elemsor kiválasztásához az összköltség minimalizásának segítségével jutunk el, amely a felhasznált elemek célegyezési és összefűzési költségeinek összegéből áll. A mondat a legkisebb összköltségű elemsorból fog előállni. Ezt a válogatást a Viterbi algoritmus határozza meg [2]. A számításhoz egy kezdő- és végelemet is felhasználunk, amely egy szünet- vagy csendjellegű elem, ezeket az ábrán C_{t01} és C_{t41}-el jelöltük.

Az ábrán a példamondat előállításához kiválasztott elemeket megvastagítva láthatjuk. Ennek az elemsornak a költsége: C_{t01}+ C_{c013}+ C_{t13}+ C_{c131}+ C_{t21}+ C_{c212}+ C_{t32}+ C_{c321}+ C_{t41}.

Abban az esetben ha nem található meg a keresett szó, akkor a szóhoz tartozó beszédhangokat keresi a rendszer. Ha a példában szereplő *„idő”* szó nem szerepelne az adatbázisban, akkor az *„i” „d” „ő”* hangokat keresi a rendszer a megfelelő környezetben és a szavakhoz hasonlóan számolja a cél- és az összefűzési költségeket.

Mint korábban említettük, az előállított beszéd minősége nagy részben függ attól, hogy a szintetizálni kívánt mondat mennyire illeszkedik a beszédkorpusz témájához. Ha hasonló szavakból álló mondatot szeretnénk előállítani, mint amilyenek a korpuszban szerepelnek, akkor hosszabb beszédelemekből (szavak, szófüzérek) tudja a szintetizátor előállítani a mondatot, a jelöltek is többen lesznek egy-egy pozícióra, így nagyobb eséllyel tud jobban illeszkedőt találni. Az elemek összeillesztésének száma is kevesebb lesz, így az esetleges illesztetlenségi hibák is kisebb számban és mértékben jelennek meg az előállított beszédben.

Összefoglalva azt mondhatjuk, hogy ebben az esetben kevésbé sértjük meg azt a *tételt*, ami azt mondja, hogy *az optimális beszédjel egyedi és egyszeri produktum*. (Ezt például a dadogó megsérti, mivel szaggatottá teszi a jelet, ezért beszéde távol lesz a köznapi normától). A tétel vonatkoztatása az adatbázisra azt jelenti, hogy minél hosszabb beszédegységeket sikerül kiválasztani, annál optimálisabb lesz a hangzás. A legoptimálisabb az a helyzet, amikor a teljes keresett mondat benne van az adatbázisban. Ilyenkor az előbbi tétel teljes mértékben teljesül [3]. Ha eltérő tematikájú mondatot szintetizálunk, akkor kisebb elemeket kell használni, azok az adatbázis különböző helyeiről származhatnak, az ottani elemeket egymástól eltérő időpontokban ejtette a bemondó, tehát az előbbi tétel sérül. Ennek eredménye a több illesztési pont szükségszerű megjelenése is, amely a percepció számára is jól hallható hangzásingadozást okozhat.

A szintetizált beszéd előállításakor a hangsorozat kialakítása mellett a prozódiát is meg kell valósítani. A prozódia alatt a hangsúlyok helyét, a dallammenetet, a szüneteket és a tempóváltást értjük, amely fizikailag az egyes hangok hangmagasságában, energiájában és időtartamában jelenik meg.

A korlátozott tematikára fejlesztett szintetizátor nem tartalmaz külön prozódia generáló és megvalósító egységet, hanem az az elemkiválasztó algoritmusba van beépítve [2]. Mivel az adatbázis elemei természetes emberi bemondásokból származnak, tartalmazzák annak a mondatnak a prozódiáját is, amelyben szerepelnek. A prozódiai információk figyelembevétele a célegyezési költségben (C_t) történik. A költségben büntetve van, ha a mondat más részéből venné az elemet a válogató algoritmus. A példamondatunkban szereplő *„lesz”* szó dallammenete csak akkor megfelelő, ha szintén a mondat végéről származik. Ha mondat közepéről vagy elejéről származó *„lesz”* szót használna a szintetizátor ebben a pozícióban, akkor természetellenes hangzást kapnánk.

3. Mi kell egy általános korpusz-alapú szintetizátorhoz?

Az általános tematikájú szintetizáláshoz két ponton kell vizsgálnunk a korpuszos, kötött témakörű szintetizátor adatbázisát. Az egyik az, hogy a szükséges hangsorépítő elemek rendelkezésre állnak-e a beszédadatbázisban a tetszőleges mondatok előállításához. A második pedig az, hogy a korlátozott tematikájú szintetizátor algoritmusai mennyire alkalmasak arra, hogy tetszőleges mondatot állítsanak elő prozódiai szempontból.

3.1. Beszédadatbázis

A vizsgálathoz három különböző tematikájú, ugyanazon bemondótól rögzített beszédkorpuszt egyesítettünk. Az első rész időjárásjelentés-típusú mondatokból

	Időtartama	Mondatok száma	Szavak száma	Hangok száma
Időjárás	10,7 óra	5821	102940	488093
Menetrend	1,1 óra	515	8656	39027
Szám	14 perc	205	1006	7042
Összesen	**12 óra**	**6541**	**112602**	**534162**

1. táblázat
A vizsgálathoz felhasznált adatbázisok méretei

állt, amely különböző időjárási internetes oldalak tartalma alapján készült. A második rész egy állomás menetrendi információit felolvasó rendszer adatbázisa, amely a járatok érkezésével és indulásával kapcsolatos bemondásokat tartalmazza. A harmadik rész egy olyan adatbázis, amely 1200 többjegyű szám felolvasását tartalmazza [4].

Az adatbázisok néhány jellemző adatát az *1. táblázat* mutatja. Mindhárom adatbázis felvételei azonos körülmények között, azonos stúdióban, azonos mikrofonnal készültek.

Látható, hogy az első – időjárásos – adatbázis a legnagyobb. Az adatbázisból a szintetizátor az aktuális prognózisokat olyan minőségben tudja felolvasni, hogy a hangzás minősége az emberi bemondásokkal közel azonos [2]. Az adatbázis tematikája a napi prognózisoknál bővebb, orvos- és közlekedésmeteorológiai témájú mondatokat is tartalmaz. A második adatbázis kisebb és a mondatok változatossága sem túl nagy, sok azonos szerkezetű és jellegű mondat is található benne. A harmadik adatbázis csak számokat tartalmaz, a három közül ez a legszűkebb tematikájú. Ez az adatbázis a többihez képest kis mérete ellenére alkalmas a számok 1 milliárdig történő emberi minőségű szintetizálására. Ez azért lehetséges, mert a felolvasott többjegyű számok a fonetikai kapcsolódások figyelembevételével, alapos tervezés után lettek meghatározva [4].

3.1.1. Szó-méretű elemek

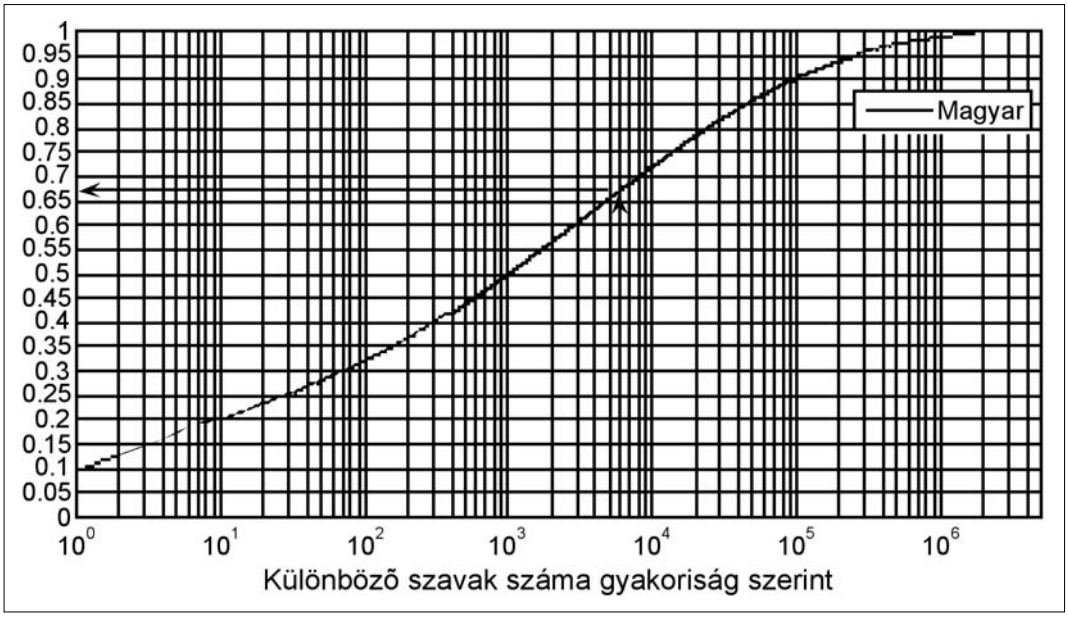
A korpuszos szintetizátor általában akkor adja a legjobb hangminőséget, amikor a leghosszabb, egybefüggő beszédrészleteket tudja felhasználni az adatbázisból. Ebben a vizsgálatban a szó az alapelem, amelyből az összesített adatbázis 112602 db-ot tartalmaz. A különböző szóalakok száma 6281. A nyelvben előforduló gyakoriságukat figyelembe véve meghatározhatjuk, hogy ezek a szavak a szintetizálandó mondatok szavainak hány százalékát teszik ki. A szavak statisztikai elemzéséhez egy saját gyűjtésű, korábbi szövegadatbázis adatait használtuk fel [5]: Digitális Irodalmi Akadémia, internetes újságok cikkei és a Magyar Nemzeti Szövegtár (1999), összesen 80 millió szó. A *2. ábrán* látható, hogy a nyelv leggyakoribb szavai a teljes nyelv szavainak hány százalékát fedik le. Amennyiben tehát a leggyakoribb 6000 szó állna rendelkezésünkre, akkor csak a 67%-ot tudnánk lefedni (nyíllal jelezve az ábrán).

Méréseink szerint a rendelkezésünkre álló – nem a leggyakoribb – 6281 szóval a mért szövegadatbázis 45%-a fedhető le, ami a szintetizálás szempontjából azt jelenti, hogy hozzávetőlegesen minden második szó esetén tudunk szó-méretű elemet felhasználni, a közbenső szavakat kisebb egységekből kell előállítani. Ez összehasonlítva a korlátozott tematikájú rendszerekkel, lényegesen rosszabb minőséget prognosztizál. A szűk tematika esetén átlagosan csak minden 15. szót állítunk elő kisebb elemekből, ami biztosítja a jó minőséget.

3.1.2. Szónál kisebb méretű elemek

A 45%-os szófedési adatból következik, hogy a kisebb elemekre gyakran lenne szükség a szintézis során. A kisebb elemek közül az egyedi hangok, a hangkapcsolatok, és a hanghármasok előfordulását vizsgáltuk meg. A magyar nyelvű szintézishez minimálisan 33 különböző hang szükséges a szünetet – mint a hang induló és befejező szakaszát – is beleértve. Vizsgálatainknál a „dz", „dzs" hangokat ritka előfordulásuk miatt, valamint a rövid-hosszú oppozíciót nem vettük figyelembe. Az így mért adatbázisban 534162 hang szerepel. Egyedi hangok összefűzéséből azonban nem lehet jó minőségű beszédet előállítani, figyelembe kell venni a hangkörnyezetet is.

Az egymásra hatások miatt az adatbázis fonetikai gazdagságáról jobb képet ad, ha a hangkapcsolatokat vizsgáljuk meg. Ilyen hangkapcsolatoknak nevezzük a kettős hangkapcsolatokat (diádok), amelyek más szintézistechnikákban rendszeresen használt elemek. Egy diád egy hangkapcsolatban szereplő két egymás melletti félhangból áll. Az összesített adatbázisban több,



2. ábra Leggyakoribb szavak fedése

mint félmillió diád szerepel. A matematikailag lehetséges 1089 (33*33) darab különböző diádból csupán 855 dbot találtunk meg az összesített adatbázisban. Ha csak azokat az diádokat számoljuk, amelyek legalább 15-ször előfordultak, akkor csak 703 különböző diád áll a rendelkezésünkre. A nagyon ritkán előforduló diádokkal az lehet a probléma a szintetizáláskor, hogy a kevés jelölt miatt, nagyon korlátozott azoknak az utaknak a száma, amelyből a szintetizátor kiválaszthatja a legjobbat, így a minőség várhatólag rosszabb lesz.

Az 1089 különböző diád élő nyelvben nem létezik, mert a nyelvtani és fonológiai szabályok miatt bizonyos kapcsolatok nem valósulhatnak meg. Például kizárólag a mássalhangzókat vizsgálva, a gyakorlatban csak 423 ilyen kettős kapcsolat van jelen a beszédben [6], amennyiben az abszolút hangsorkezdő-záró állapotot is ide számoljuk. Annak megállapítására, hogy melyek azok a diádok, amelyek tetszőleges szöveg szintetizálásakor szükségesek lehetnek, a szavaknál mutatott statisztikai módszerhez hasonlót használtunk. A szószablya [7,8] magyar webkorpusz (mint független adatbázis) mondatait a szintetizátor betű-hang átalakító rendszerével átírattuk fonetikus formába, majd előállítottuk ezekből ugyanazokat az adatokat, amelyeket az összesített adatbázisból is.

A szószablya korpusz adatai a *2. táblázatban* láthatók. Egy mondat átlagosan 83 diádból épül fel.

A különböző diádok száma itt már nagyobb, mint az összesített adatbázisban. A gyakorisági adatok szerint azok a diádok amelyek a szószablya webkorpuszban szerepelnek, de az összesített adatbázisban nem, az összes diád 1%-át teszik ki, ami azt jelenti, hogy átlagosan minden századik felhasználandó diád hiányozni fog. Ha csak diádokból építenénk fel a mondatot, akkor átlagosan 1,2 mondatonként lenne hiányzó diádunk, ami – ha csak ebből a szempontból vizsgáljuk – jó minőséget eredményezhetne.

A hanghármasok vizsgálatára azért van szükség, mert a korpuszos szintetizátor hang-alapú működése során akkor lehet a legjobb a kiválasztott hang minősége, ha a szintetizálandó mondat minden hangját (a környezetével együtt) megtaláljuk a beszédadatbázisban is. Ezt úgy biztosíthatjuk a keresésnél, hogy egy hang bal és jobb oldali szomszédját is figyelembe vesszük a célegyezési költség számításakor. Akkor optimális a helyzet, ha a szomszédos hangok ugyanazok, mint a szintetizált mondatban. Az adatbázis vizsgálatakor tehát most azt nézzük, hogy az ott megtalálható hanghármasok mennyire fedik le a magyar nyelvben használtakat. Az összesített adatbázisban 8727 db különböző hanghármas található, amiből 5748 db fordult elő legalább ötször.

A hanghármasok statisztikai vizsgálatához a – diádoknál is használt – szószablya webkorpuszt használtuk. Az elkészített fonetikus átiratban megvizsgáltuk, hogy milyen hanghármasok fordultak elő a webkorpuszban. Összesen 27982 különböző hanghármast találtunk, melyek közül 16643 fordult elő gyakran (legalább ezerszer).

Abban az esetben, ha az összesített adatbázisban előforduló összes hanghármas fedését vizsgáljuk, akkor az ott találtak a webkorpusz 96%-át fedik le. Ha a 15 vagy többször előforduló hanghármasokat vesszük csak figyelembe, akkor a fedés csak 82%-os. Ezt az adatot annak függvényében kell vizsgálni, hogy jó minőségű beszédet abban az esetben is elő lehet állítani, ha az adott hanghármas nincs meg pontosan, csak a hang artikulációs pozíciója egyezik. Az azonos képzési helyű mássalhangzók (consonant-C) hatása a hozzájuk kapcsolódó magánhangzókra (vowel-V) hasonló [9]. Tehát ha egy VCV kapcsolatban a C-re csak azonos képzési helyű C1 helyettesítőt találunk, akkor a C1-hez kapcsolódó magánhangzó ugyanolyan akusztikai szerkezettel fog rendelkezni, mint a VCV kapcsolatban, a helyettesítés tehát nem rontja az akusztikai eredményt.

Az összesített adatbázisról általánosságban elmondhatjuk tehát, hogy hang-szinten alkalmas tetszőleges beszéd előállítására, hosszabb elemek szintjén azonban túl hiányos.

3.2. Prozódia

Az emberi minőséghez közelítő szintetizált beszéd előállításához nem elég az, ha az adatbázisban megtalálhatók az előállítandó hangsornak megfelelő hangsorépítő elemek, hanem szükség van arra is, hogy a szintetizált mondat megfelelő prozódiával is rendelkezzen. Ha a prozódia nem megfelelő, a hallgató nem fogadja el természetes hangzású beszédnek a mondatot. A prozódia helyes előállítása legalább olyan nehéz feladat, mint a hangsorépítő elemek biztosítása.

A vizsgált, korlátozott tematikájú szintetizátorok adatbázisa csak kijelentő mondatokat tartalmaz. Mivel a kérdő mondat prozódiája jelentősen eltér ezektől, ezért a kérdő mondatokat az aktuális algoritmusok ezzel az adatbázissal nem képesek előállítani. A kérdő mondatok előállításához vagy olyan adatbázis kell, amely nagy számban tartalmaz kérdő mondatokat is, vagy olyan prozódia kiválasztó és megvalósító algoritmusok szükségesek, amelyek ezeket meg tudják valósítani. A továbbiakban már csak azt vizsgáljuk, hogy kijelentő mondatok esetében milyen esély van a helyes prozódia megvalósítására.

A vizsgált kötött témájú korpusz-alapú szintetizátorban a prozódia modellezése úgy történik, hogy figyeljük a szavak mondatbeli pozícióját [3]. A mondatokat első lé-

2. táblázat A szószablya korpusz főbb adatai

	weblapok száma	mondatok száma	szavak száma	hangok száma
Szószablya webkorpusz	1,2 millió	42 milló	589 milló	3,5 milliárd

pésben tagmondatokra bontjuk, majd ezen belül is meghatározzuk a szó helyzetét. A talált szóalakok vizsgálata során láthattuk, hogy azok átlagosan 45%-ban fedik le a magyar nyelvet, tehát a helyes prozódia is ilyen arányban állhat elő a szavakból a legjobb esetben. Az adatbázisban kis számban előforduló szavak esetén az is előfordulhat, hogy a szó ugyan egészben szerepel az adatbázisban, de nem a megfelelő mondatbeli pozícióban, ezért nem a megfelelő prozódiai információt hordozza.

Abban az esetben, ha kisebb elemekből, építi fel a mondatot a szintetizátor, akkor már nem veszi figyelembe ezeket a mondatbeli pozíció információkat. Előfordulhat tehát, hogy egy hangsúlyos szót olyan szavak elemeiből állít elő, amelyek hangsúlytalanok, ezért a kimenet is hangsúlytalan lesz.

A prozódia megvalósításáról tehát összegezve azt mondhatjuk el, hogy csak akkor várható el viszonylag elfogadható hangzás, ha a szintézis szó szinten tudja biztosítani a hangsorépítő elemeket és ezekből is elég számú van a beszédadatbázisban, amelyek a prozódiai változatosságot biztosítják.

4. Meghallgatásos tesztek

A beszédszintézis rendszerek minőségét meghallgatásos tesztek során végzett szubjektív minősítéssel lehet összehasonlítani. Ennek egyik módja a MOS (Mean Opinion Score – átlagos szubjektív osztályzat) teszt alkalmazása. A tesztekhez mondatokat válogattunk két témakörből. Az elsőben hírolvasásból, a másodikban egy meséből származtak a mondatok. Az előállított tesztanyag 5-5 szintetizált mondatot tartalmazott, amelyek eltérő hosszúságúak voltak. A mondatokat meghallgató és értékelő személyek számára az volt az utasítás, hogy egy 5-ös skálán értékeljék a minőséget (5-ös a legjobb érték). A tesztben továbbá szerepeltek a korpuszos szintetizátor eredeti mondatai is, amelyek a tematikának megfelelő időjárás jelentések voltak. A teszt internetes elérhetőségű volt, a tesztelők a mondatokat véletlen sorrendben hallgatták meg. A teszt tartalmazott egy bevezető részt is, amely azt a célt szolgálta, hogy a tényleges értékelés előtt már képet kapjanak arról, hogy milyen minőségű mondatokat fognak hallani. A teszt során a tesztelők nyilatkoztak arról is, hogy milyen eszközön, milyen környezetben hallgatják a mondatokat.

A tesztet 10 személy értékelte ki; 3 nő és 7 férfi. Az átlagéletkor 32 év volt. A tesztelők mindegyike csendes környezetben hallgatta meg a mondatokat, a legtöbben átlagos minőségű eszközökön. A tesztelők fele-fele arányban használtak hangszórót és fejhallgatót.

A *3. ábrán* az első oszlop mutatja a korpuszos szintetizátorral előállított, a témakörbe vágó mondatok értékelését. A második, vonalazott oszlop a hír és mese témakörökből válogatott mondatok átlaga. Az utolsó két oszlopon a két témakör külön-külön számított átlaga látható. A tematikán kívüli mondatok érthetősége roszszabb és kevésbé természetesek, mint az adatbázisnak megfelelő tematikájú korábban szintetizált mondatok. A különbség a két átlag között több mint 2, ami azt jelenti, hogy a minőségromlás jelentős. Az eredményekből az is megállapítható, hogy az eredeti tematikához közelebb álló hírjellegű mondatok jobbak, mint a tematikától messze álló meserészlet, bár ezek eltérése kicsi, ha a témakörbe vágó mondatokhoz viszonyítjuk.

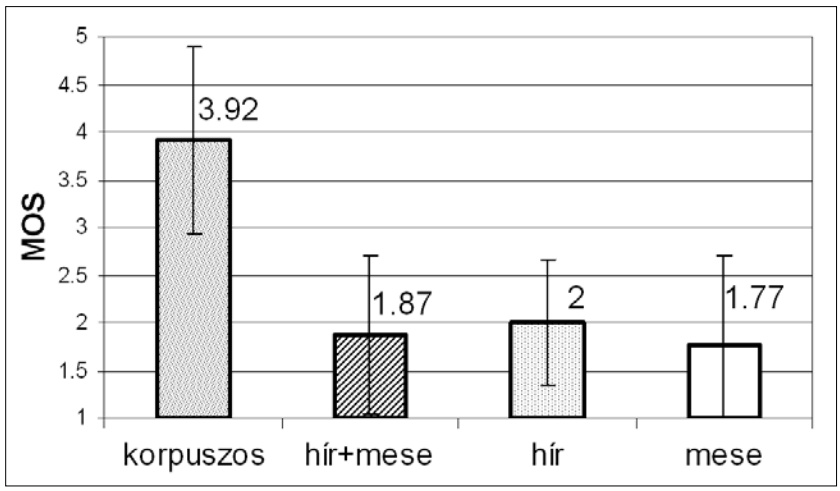
A meghallgatás utáni szabad véleményalkotás során kiderült, hogy a tesztelők szerint a mondatok egyes részei mind prozódiában, mind akusztikai szerkezetben lényegesen különböztek egymástól. Voltak részek, amelyek sokkal jobb osztályzatot kaptak volna, de a mondat többi része lehúzta az értékelést. A legtöbbet említett jelenség az egyenetlen minőség volt.

5. Fejlesztési lehetőségek

Az adatbázisok elemzéséből látható, hogy méretük növelése egyértelműen javíthatja a generálandó szintetizált beszéd minőségét. Ezt a kötött témakörű rendszerek fejlesztése során már többször alkalmaztuk. Ha újabb mondatok szintetizálásának igénye jelent meg és a szintézis hangminősége nem volt megfelelő, akkor egy jól megtervezett hangfelvétellel az adatbázist úgy bővítettük, hogy ezután ezeket az újabb mondatokat is jó minőségben tudta előállítani a rendszer. Amennyiben viszont azt szeretnénk, hogy tetszőleges tematikájú mondatot is szintetizálni tudjunk megfelelő minőségben, akkor az adatbázist olyan mértékben kellene bővíteni ezzel a módszerrel, amely nehezen vagy gyakorlati szempontból egyáltalán nem megoldható.

A jelenleg használt adatbázis 6281 különböző szót tartalmaz. Ha azt szeretnénk elérni, hogy az adatbázisban a magyar szavak 95%-a szerepeljen, akkor a 2. ábrából leolvashatjuk, hogy ehhez hozzávetőlegesen 150

3. ábra
Szubjektív minősítés átlagai az egyes tematikákra



ezer szót kellene felvenni legalább ötfajta mondatbeli pozícióban. Ez azt jelenti, hogy a meglévő adatbázishoz képest körülbelül 700 ezer szót tartalmazó mondatkorpuszt kellene a bemondóval bemondatni és feldolgozni. Ez a meglévő adatbázis 10 órájához képest, nagyságrendileg újabb 100 óra felvételt jelentene, ha sikerülne egyáltalán ezeket a mondatokat megalkotni. Ennek teljesítése irreális követelmény.

A másik megközelítés lehet a minőség javítására, hogy a korábbi szintetizátortechnikáknál használt prozódiai modulok kimeneti információit használjuk fel az általános korpuszos szintetizátorban. Tapasztalatból tudható azonban, hogy az emberi hangminőség – amelyet a szintetizátor akkor nyújt, amikor a saját tematikájának megfelelő mondatokat állít elő – nem érhető el ezzel a technikával. Ezzel a módszerrel azonban ki lehet egyenlíteni azokat a minőségbeli durva egyenetlenségeket, amelyek a meghallgatásos teszt során az észlelők kifogásoltak. Egy korábbi, elemösszefűzéses technikájú szintetizátor 2,5-es szubjektív minősítést ért el egy hasonló meghallgatásos teszt során[2]. Tehát ha ennek a szintetizátornak a prozódiai információt és a korpuszos szintetizátor bővebb hangadatbázisát egyesíteni tudjuk, akkor várhatóan a mostani 2 körüli minősítést a régebbi technikájú szintetizátor 2,5-es minősége fölé tudjuk vinni.

6. Összefoglalás

A korlátozott tematikára tervezett beszédadatbázis és a hozzá kapcsolódó korpuszos beszédszintetizátor változtatás nélkül nem alkalmas tetszőleges tematikájú mondatok előállítására. Amennyiben mégis ilyen irányú fejlesztést kívánunk elindítani, akkor a szintetizátor minőségének egyik javítási megoldása lehet az adatbázis növelése. Ez a jelentős mennyiségű adatbővülés miatt nehezen megvalósítható.

A másik megoldás a prozódiai modul fejlesztése, amellyel az érthetőség jól javítható. Ennek a hátránya, hogy további jelfeldolgozást kíván meg, amely a természetes hangzást ronthatja, de elkerülhető vele az egyenetlen minőség a hangzásban.

Köszönetnyilvánítás

Köszönöm a BME TMIT Beszédtechnológiai laboratórium munkatársainak segítségét, bátorítását.
A kutatást részben az NKFP 2. programja támogatta (szerződésszám: 2/034/2004).

A szerzőről

Zainkó Csaba 1999-ben végzett a BME Villamosmérnöki és Informatikai Kar Médiainformatika szakirányon és azóta a Távközlési és Médiainformatikai Tanszék Beszédtechnológiai laboratóriumában dialógusrendszerek és az ahhoz kapcsolódó komponensek kutatásával és fejlesztésével foglalkozik. Részt vett az első magyar nyelvű elektronikus levél felolvasó és a számszerinti tudakozó fejlesztésében. Jelenleg a korpusz-alapú beszédszintézis technológiájának vizsgálata áll kutatási témájának középpontjában.

Irodalom

[1] Németh Géza, Olaszy Gábor, Fék Márk:
Új rendszerű, korpusz alapú gépi szövegfelolvasó fejlesztése és kísérleti eredményei.
Beszédkutatás 2006. Szerk.: Gósy Mária.
MTA Nyelvtudományi Intézet, 2006, pp.183–196.

[2] Fék M., Pesti P., Németh G., Zainkó Cs.:
Generációváltás a beszédszintézisben.
Híradástechnika, 2006/3. pp.21–30.

[3] Olaszy Gábor:
A korpusz alapú beszédszintézis nyelvi, fonetikai kérdései.
Híradástechnika 2006/3. pp.43–50.

[4] Olaszy G., Németh G.:
IVR for Banking and Residential Telephone Subscribers Using Stored Messages Combined with a New Number-to-Speech Synthesis Method.
In: Human Factors and Voice Interactive Systems, Ed.: Daryle Gardner-Bonneau.
Kluwer Academic Publishers, 1999, pp.237–256.

[5] G. Németh, Cs. Zainkó:
Multilingual Statistical Text Analysis, Zipf's Law and Hungarian Speech Generation,
Acta Linguistica Hungarica, Vol. 49. (3-4), 2002, Akadémiai Kiadó, pp.385–405.

[6] Olaszy Gábor:
Mássalhangzó-kapcsolódások a magyar beszédben.
Tinta Kiadó, Budapest, 2007.

[7] Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor:
Creating open language resources for Hungarian,
In: Proc. of the 4th International Conference on Language Resources and Evaluation (LREC) 2004.

[8] Kornai, A., Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D.:
Web-based frequency dictionaries for medium density languages,
In: Proc. of the 2nd Int. Workshop on Web as Corpus, Ed.: Adam Kilgarriff, Marco Baroni, ACL-06, 2006, pp.1–9.

[9] Olaszy Gábor:
Az artikuláció akusztikai vetülete – a hangsebészet elmélete és gyakorlata.
KIFLAF 2003, Szerk.: Hunyadi László.
Debreceni Egyetem, pp.241–254.

Beszédadatbázisok előkészítése
kutatási és fejlesztési célok hatékonyabb támogatására

NÉMETH GÉZA, OLASZY GÁBOR,
BARTALIS MÁTYÁS, ZAINKÓ CSABA, FÉK MÁRK, MIHAJLIK PÉTER

BME Távközlési és MédiainformatikaiTanszék
{nemeth, olaszy, bartalis, zainko, fek, mihajlik}@tmit.bme.hu

Lektorált

Kulcsszavak: beszédadatbázisok, címkézés, korpusz alapú beszédszintézis, hanghatárkorrekció

A nagyméretű beszédadatbázisok készítése az utóbbi évtizedekben vált szükségessé, hogy támogassák egyrészről a beszédkutatást, másrészről a működő beszédinformációs rendszerek fejlesztését. Az ilyen adatbázisok akkor szolgálhatják jól a tudományt, ha részletes belső címkézéssel is rendelkeznek. Jelen cikkben olyan adatbázisokkal foglalkozunk, amelyek egyetlen bemondótól felvett, sok mondatból álló, több órányi anyagot tartalmaznak. Az ilyen beszédadatbázisok címkézésénél alapvető gond, hogy a címkézést teljes mértékben emberi erővel nem lehet elvégezni a munka nagysága miatt. A cél viszont az, hogy a címkék a lehető legpontosabban legyenek bejelölve a hullámformában. A cikkben ismertetett új hibrid eljárást eredményesen lehet alkalmazni az ilyen munkákhoz, szinte hibamentes címkézés érhető el, időigénye is elviselhető (2-3 nap egy több órás adatbázisra). Az így készített beszédadatbázisokban a keresés megbízható eredményeket ad, melyet fel lehet használni a beszédkutatásban, a beszédszintézisben és a beszédfelismerésben is.

1. Bevezetés

A részletes címkézés érintheti a szegmentális szerkezetet (hanghatárok, szavak határai), valamint a szupraszegmentális szintet (hangsúlyok, dallammenetek, szünetek, prozódiai egységek). Az adatbázisok címkézési munkáit nagy méretük miatt csak jelentős szoftver-támogatással lehet költséghatékonyan elvégezni. Léteznek már évek óta magyar beszédadatbázisok, amelyeket főleg beszédfelismerő algoritmusok tanítására fejlesztettek [7,8]. Ezekben általában sok beszélőtől vettek beszédmintát és a címkézési munkákat még többnyire jelentős mértékben kézi erővel végezték.

Jelen cikkben olyan adatbázisokkal foglalkozunk, amelyek egyetlen bemondótól felvett, sok mondatból álló anyagot tartalmaznak. Ezek címkézéséről van szó. Egyelőre a hanghatárok bejelölésével kapcsolatos szoftverrendszer fejlesztéséről és annak működési tapasztalatairól számolunk be. A rendszert a BME Távközlési és Médiainformatikai tanszékén fejlesztették és az ottani beszédadatbázisokhoz használják. Az eljárás fontos tulajdonsága, hogy szoftveres elemek és emberi erőforrás váltogatják egymást a feldolgozás során. A beszédfeldolgozás egyes pontjain még ma sem lehet kihagyni az emberi döntéshozatali tényezőt.

Az eredmények igazolják, hogy ilyen hibrid eljárással elérhető a szinte hibamentes címkézés, ennek ára viszont a bonyolult, kissé időigényesebb feldolgozás. Az ilyen adatbázisokból pontos és megbízható adatok nyerhetők. A vizsgált adatbázisokról kapott információk azt is megmutatják, hogy az egyes beszélők közötti hangszintű beszédképzési eltérések számszerű adatokkal is jellemezhetők, ami a személyre szabott szoftveres beszédjellemzés egyik kísérleti megvalósításának is tekinthető.

2. A munka célja, módszere és a feldolgozott anyag

A nagyméretű beszédadatbázisok hullámformáját ma már el lehet látni hanghatár-címkékkel szoftver segítségével, azonban az eredmény sohasem teljesen pontos. Ez annak a következménye, hogy a beszédjel biológiai mechanizmus terméke. A beszéd előállítása során a pillanatnyi motoros és artikulációs történések határozzák meg a kisugárzott hullámformát (ugyanazon mondat többszöri kiejtése során minden esetben más akusztikai eredményt kapunk, a hangzás csak globálisan, nyelvi szempontból lesz ugyanaz). A gépi címkézés pontossága sok tényezőtől függ. A jelen kísérlet célkitűzése az, hogy tegyük teljessé a címkézést, a szoftveres alapcímkézés eredményét javítsuk tovább célzott szubrutinokkal, a géppel nem javítható hibákat pedig emberi erővel javítsuk.

Eredményként olyan beszédadatbázist kapunk, amelyikben egyrészről úgymond minden hanghatár helyesen van bejelölve, másrészről a beszédhangok minőségi osztályozására is vannak jelzések. Ez utóbbi megjegyzésen a következőket értjük. Az emberi beszédben a hangkapcsolódások artikulációjából adódóan előfordulhatnak olyan hangok, amelyek belső szerkezetüket tekintve torzultak és nem felelnek meg a fonetikai hangleírásoknak [4]. Ezek a torzult hangok ugyan szerves részét képezik a hangsornak, de csak a saját szélesebb hangkörnyezetüket tekintve (a szó, amiben szerepelnek) adják meg az emberi percepciós rendszernek a megértéshez szükséges akusztikai információkat. Az ilyen hangok megjelölése azért fontos, mert az adatbázis felhasználása során ezek a hangok szétválaszthatók a többitől (például hangzásvariációk keresésekor, vagy beszédszintézisnél, amikor el kell dönteni, hogy mikor melyik hangot használjuk

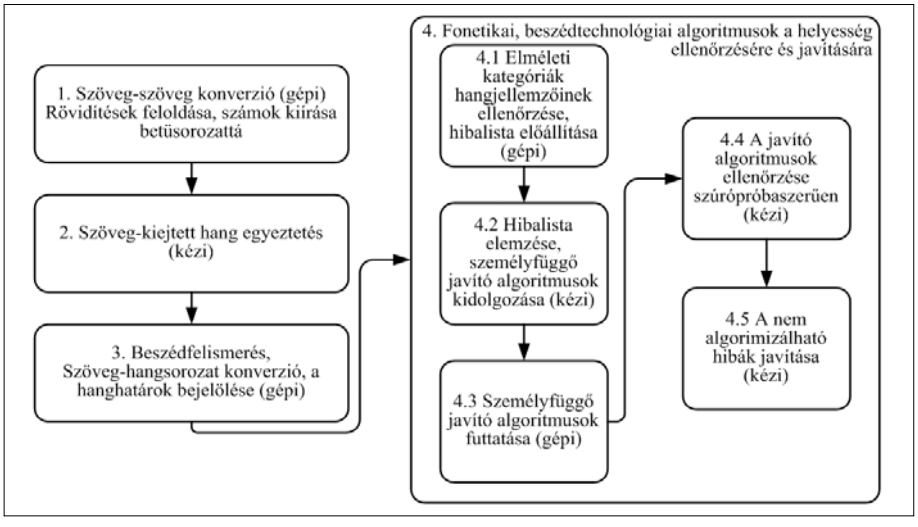
stb.). Az ilyen, jól címkézett beszédadatbázisok a későbbiekben sokféle célra felhasználhatók az oktatásban, a kutatásban és az alkalmazás-fejlesztésekben is.

A cél megvalósítására a hagyományos egylépcsős gépi megoldással szemben fokozatos kialakítású, többlépcsős javító rendszert fejlesztettünk ki. A kísérletek azt mutatták, hogy a nagy pontosságú címkézéshez olyan hibrid megoldást kell keresni, amelyben az egyes lépcsők futtatása között emberi döntéseket is kell hozni, hangolni kell bizonyos szoftverelemeket (nem teljesen automatikus címkézésről van szó). A célkitűzéshez tartozott az is, hogy minimumra csökkentsük a manuálisan javítandó hibák számát, vagyis ésszerűen kézben tartható hibamennyiség maradjon a szoftveres feldolgozás után (max. 500 hiba/adatbázis, amely manuálisan 2-3 óra munkával javítható). A feldolgozáshoz és az eredmények megjelenítéséhez a Praat 4.0 szoftvert is használtuk [1].

A gyakorlati tapasztalatok azt is megmutatták, hogy a címkézési folyamat beszélőfüggő megoldást kíván egészen addig, amíg elegendő mennyiségű, különböző beszélőtől felvett adatbázis feldolgozása meg nem történik (az adott nyelvre). Az ilyen adatbázisok készítésénél maga a hangfelvétel létrehozása is komoly munkát igényel, ezért jelenleg kevés beszélőtől áll rendelkezésre nagyméretű beszédadatbázis. A jelen munkában csak a kezdetekről tudunk beszámolni, mindössze négy beszédadatbázis hangfelvételét készítettük el és dolgoztuk fel (három képzett női beszélő: N1, N2, N3 és egy férfihang: F1). Megvizsgáltuk a beszélők közötti kiejtési (hangképzési) jellegzetességeket is, amelyek befolyásolják a helyes szoftveres címkézést. Mindegyik adatbázis a BME Távközlési és Médiainformatikai Tanszék beszédkutató laboratóriumában található. Az adatbázisok néhány jellemző adatát az *1. táblázatban* láthatjuk.

3. A címkézésre kidolgozott módszer

A új, hibrid gépi címkézési eljárás moduljai az *1. ábrán* láthatók.



1. ábra
A beszédadatbázis hanghatárainak címkézési folyamata

4. Az alapcímkézés

Az alapcímkézést géppel végezzük (4.3. pont), a hibajavításokat gépi és emberi feldolgozással. A hangfelvétel elvégzése után két előkészítő lépést kell elvégezni, hogy a gépi alap címkézést elkezdhessük.

4.1. Első lépés

A felolvasáshoz alkalmazott karaktersorozatot betűsorozattá alakítjuk, ahol a szövegben rövidítés vagy szám van, azt feloldjuk és betűkkel kiírjuk. Az így átalakított eredeti szöveg csak betűkaraktereket tartalmaz majd. A konvertálást célprogram végzi, amely a Profivox beszédszintetizátor szöveg-szöveg átalakító moduljának felhasználásával készült [6].

Példa:
Eredeti szöveg: A hőmérséklet június 3-án Bp-en -3 C° körül várható.
Átirat: A hőmérséklet június harmadikán Budapesten mínusz három celziusz fok körül várható.

4.2. Második lépés

Ebben a lépésben végezzük el a szöveg-beszédhullám szinkron ellenőrzését. Egy szakértő meghallgatja a mondatokat, közben vizuálisan ellenőrzi a hozzájuk tartozó szöveget (a 4.1. szerint). E vizsgálattól függ a további munka sikeressége. Ezt a folyamatot csak ember tudja elvégezni. Ez a munkafázis meglehetősen sok időt vesz igénybe és komoly koncentrációt kíván. A talált hibák kijavítása után elméletileg a két médium szinkronban van egymással (a gyakorlat azt mutatja, hogy néhány hiba azért benne marad valamelyikben, ez a későbbi szoftverellenőrzéskor kiderül és akkor javítjuk).

1. táblázat
A feldolgozott beszédadatbázisok adatai

Adatbázis	Mondat-szám	Szavak száma	Hangok száma	Hangzási idő (perc)	Beszédseb. hang/s	Artik. seb. hang/s	Megjegyzés
N1	5821	102940	488093	641,7	12,89	13,28	időjárásjelentés
N2	3643	43345	259353	331,6	13,30	13,44	telefon árlista
N3	792	9086	39353	50,3	12,68	12,71	prompt
F1	430	3071	11822	15,3	12,66	12,66	általános szöveg

4.3. Harmadik lépés

A bevezető két lépés után következik az alapvető gépi hangfelismerés és címkézés (a beszédhangoknak és határaiknak a meghatározása). Szükség van az adott hangfelvételre, annak ortografikus szöveges átiratára és az alkalmazandó (opcionálisan vagy kötelezően) végbemenő hasonulási, összeolvadási stb. jelenségek megadására. Az úgynevezett kényszerített illesztési (forced alignment) üzemmódú beszédfelismerő első lépésként a csak betűkből álló szöveget (a 4.2.-ből) hangszimbólumok sorozatává alakítja. Ez a szimbólumsorozat fogja segíteni a hanghatárok felismerését (a gép tudja, hogy milyen hang lehet az adott ponton). A hangszimbólumokat itt most két ferde zárójel közé írva adjuk meg.

Példa:
A hőmérséklet június harmadikán Budapesten mínusz három celziusz fok körül várható.
/a/ /h/ /o3/ /m/ /e1/ /r/ /s/ /e1/ /k/ /l/ /e/ /t/ /j/ /u1/ /n/ /i/ /u/ /s/ /h/ /a/ /r/ /m/ /a/ /d/ /i/ /k/ /a1/ /m/ /b/ /u/ /d/ /a/ /p/ /e/ /s/ /t/ /e/ /m:/ /i1/ /n/ /u/ /sz/ /h/ /a1/ /r/ /o/ /m/ /c/ /e/ /l/ /z/ /i/ /u/ /sz/ /f/ /o/ /k:/ /o2/ /r/ /u2/ /l/ /v/ /a1/ /r/ /h/ /a/ /t/ /o1/.

A hangszimbólumok a hozzájuk tartozó betűkkel megegyezők, kivéve az ékezetes betűket, amelyek alapját a főkarakter és az ékezetnek megfelelő szám kombinációja adja. Például: o=o, o1=ó, o2=ö, o3=ő. A hosszú hang jelölésére a kettőspontot használjuk.

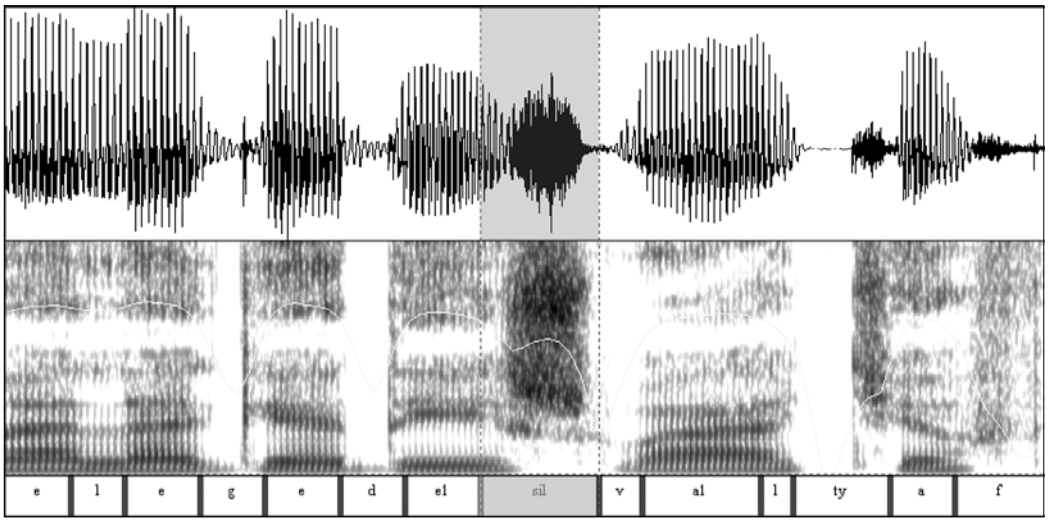
A hangszimbólum sorozat alapján a kényszerített illesztést alkalmazva – amikor is a felismerési hálózat a szavak lineáris szekvenciájából adódik – a felismerő egyszeri futtatásával előállnak mind a hanghatárok, mind a hang-identitások. Az eljárás kezeli a szóhatárokon átívelő hasonulási jelenségeket és az ezekből adódó hangkieséseket is. További részletek a gépi címkézési algoritmusról a [2,3]-ban találhatók.

A megközelítés feltételezi, hogy a felismerés során használt, az akusztikus modellek betanításához használt sokbeszélős adatbázis címkézése pontos volt. Amennyiben ez nem áll fenn, a szisztematikus címkézési hibák továbbadódnak. Ezek kiküszöbölése speciális célú algoritmusokkal lehetséges. A statisztikai elvből is következik, hogy néhány hanghatár pontatlan lehet. Végül meg kell említeni, hogy mivel a beszédfelismerő telefonsávi beszéddel lett betanítva, a magas frekvenciakomponensekben gazdag hangok, mint a „c”, „sz” stb. határainak pontos felismerése elvileg is lehetetlen. Robusztussága miatt viszont bevált ez a megközelítés.

Fontos továbbá azt is megjegyezni, hogy ennek a hangfelismerési eljárásnak az a tulajdonsága, hogy nem vizsgálja sem a zöngés/zöngétlen állapotot, sem a hangidőtartamot, csak spektrális jellemzők alapján dönt. A program felismerési pontossága a fent felsorolt problémák ellenére nagyon jó, 95-96%-os. A hangfelismerés eredményeként megkapjuk a hanghatárok (címkék) sorozatát, amit a hullámformával párhuzamosan megjeleníthetünk *(2. ábra)*.

A megjelenítés alapján vizuálisan is ellenőrizhetjük a címkézés pontosságát (a 4.3. munkafázist többször is el kell végezni, ha a 3.1.2. pontban hibákat vét az ellenőrző személy). A vizuális ellenőrzést a későbbi hibadetektálásoknál és javításoknál is használjuk. Például a 2. ábrán a szürke sáv hibát jelez. A melegedés váltja fel hangsorban az s hang helyén szünet /sil/ jel szerepel a hangszimbólum reprezentációban. Ez egy szinkronizálási hibából fakad. Az eredeti szövegből kimaradt az s betű, így a felismerőnek eggyel kevesebb hangot kellett a hullámformába beilleszteni, ezért ide szünetet jelölt. További hiba, hogy az /e1/ magánhangzó végéből 4 periódus a zöngétlen /s/ hanghoz van jelölve.

A gépi felismerő tévesztései nagyban függnek attól is, hogy a beszélő személy beszédképzési mozzanatai mennyiben felelnek meg az elméleti fonetikai modell paramétereinek (például zöngésség/zöngétlenség, zárképzés, rövid-hosszú oppozíció stb.). Ha a beszélő gyorsan és laza artikulációval beszél, akkor több címkézési hiba lesz, ellenkező esetben kevesebb. Hibaforrás lehet továbbá a koartikulációból adódó olyan hangelem, amelyik fonemikus szinten nem köthető beszédhanghoz [5], illetve annak részéhez (svá, koartikulációs néma fázis, stb.) Ilyenkor a felismerő hibásan állapíthatja meg a hanghatárt. A harmadik hibacsoport, amikor a beszélőre jellemző hangképzési állapotok eltérnek az elméleti osztályozásoktól (például zöngétlen laterális mássalhang-



2. ábra
A hanghatárok és a hangszimbólumok (lent), valamint a hullámforma (fent) együttes megjelenítése a spektrális tartalommal (középen) együtt.

zó is kialakul, ilyen hangot nem definiál a magyar hangrendszer). Végül címkézési hibát okozhat a szinkrontól való eltérés is. Ilyenkor durva elcsúszások lehetnek (lásd a 2. ábrán). Szinkron hiba esetén a 4.2. pontra kell visszamenni a feldolgozásban, a szinkront helyre kell állítani és a felismertetést újból el kell végezni.

A további feldolgozás során tehát ezt a címkézett adatbázist (a hibáival együtt) tartjuk a következő, javító munkafázisok bemenetének.

5. Fonetikai alapú algoritmusok az ellenőrzésre és javításra

A célkitűzés az, hogy a gépileg helyenként hibásan bejelölt hanghatárokat (a teljes állomány 4-5%-a) a beszédadatbázisban feltárjuk és kijavítsuk. Javítási követelmény, hogy a hibák többségét algoritmussal korrigáljuk és csak kis részét manuálisan. A hibadetektáláshoz egyrészről célzott módon kialakított jelfeldolgozási eljárásokat használunk fel, másrészről a fonetikai hangleírások ismérveit, tehát olyan adatokat, amelyeket a beszédfelismerő nem vett figyelembe. A hibafeltárás során támpontul szolgált a zöngés periódusok utólagos automatikus megjelölése a hullámformában, a hangokra jellemző specifikus intenzitásértékek kiszámítása, valamint a jellemző hanghosszúság figyelembevétele a hangkörnyezet függvényében [6].

A hibakeresés során kialakított algoritmusok a hangra jellemző paramétereket hasonlítják össze a tényleges hullámformával a korábban bejelölt két hanghatár közötti szakaszon. A kategorizálás kétszintű.

5.1. Első szint

Elsőként a durva hibákat szűrjük ki, a rosszul jelölt hangokat keressük. Minden hangot és hanghatár jelölést egymással összevet az erre a célra fejlesztett algoritmus és azokat a hangokat tekintjük hibás jelölésűnek, amelyeknek az akusztikai tartalma nagyrészt (75%) egyáltalán nem felel meg a hang fonetikai leírásának (például egy zöngétlen zárhangnak titulált hangsorrész 80%-ban zöngés periódusokat tartalmaz; egy zörejes, nagy energiájú hangsorrész szünetnek van jelölve). Ez a kereső szoftver a fonetikai jellemzés alapján kategorizál.

Néhány esetben ellentétbe kerülhet az elméleti fonetikai kategorizálás és a hang valóságos fizikai tartalma (ez beszélőtől függ), amit a tények ellenére nem szabad hibának tekinteni. Például az intervokális helyzetű zöngés zárhangokat (*...fejében...*) néhány helyen hibásnak találta a program, azokban a hangkapcsolatokban, ahol a zárhangban nem lehetett periodikus elemeket felfedezni (így ejtette a beszélő). Ez valójában nem hiba. Az algoritmus fejlesztése során az ilyen hibás döntések elkerülésére előírtuk például, hogy VCV hangkörnyezeti esetekben a zöngés zárhangot minden esetben zöngésnek kell tekinteni, függetlenül a hanghullámban mérhető állapotoktól. Egy további ilyen példa, hogy a /h/ hangra engedélyeztük, hogy zöngés és zöngétlen formában is szerepelhet a hangsorban.

A durva hibákat megjelölő szoftver hibalistát generál, amelyben megadja a hiba hangsorbeli helyét, hangkörnyezetét és fajtáját. A listában sokféle hiba szerepel, számuk általában nagy. A gyakorlat azt mutatta, hogy az ilyen durva hibák megszüntetését nem lehet egyetlen javító algoritmussal elvégezni a hibák sokrétűsége miatt, hibatípusonként külön szubrutinokat kellett fejleszteni. A szubrutinok megírásához kategorizáltuk a listából a tipikus eseteket. Tipikus hibának tekintettük, ha ugyanaz a hiba legalább 12-szer megjelent a listában. A továbbiakban manuális, vizuális és auditív vizsgálatot végeztünk a többször előforduló egyforma hibákra (5-7 ugyanolyan hibát vizsgáltunk) annak megállapítására, hogy algoritmizálható-e a hiba kijavítása. Ha egységes képet mutattak a vizsgált esetek, akkor egy-egy egyszerű algoritmust dolgoztunk ki a hiba javításához. Ezt lefuttatva a beszédadatbázison az adott hibák megszűntek és az összesített hibaszám csökkent.

A négy adatbázis vizsgálata során összesen 18 fajta javító szubrutint dolgoztunk ki a durva hibák csökkentésére. Néhány példát mutatunk a *2. táblázatban*. Feltüntettük azt is, hogy a hibásnak talált hangkapcsolatból összesen hány darab van az adatbázisban, ebből látható, hogy a beszédfelismerő csak néhány esetben adott téves felismerést és címkézést. A szürke mezők jelzik azokat a hibákat, amelyek javítására szubrutint dolgoztunk ki. A táblázatból az is kiolvasható, hogy a legtöbb durva hibát az N2 jelű beszélő mondataiban találtuk, aki jellemzően gyorsbeszédű volt.

2. táblázat A feldolgozott beszédadatbázisokban talált hibás hangkapcsolatok előfordulási adatai

Adatbázis	N1		N2		N3		F1	
hibás hangkapcs.	hibás	összes	hibás	összes	hibás	összes	hibás	összes
/s,sz,c,cs/ + /m,n,ny/	8	1340	1	158	1	72	1	18
/amf/		232	22	923		4		2
/cez/			73	1528		61		
/cij/ + V		104	1	116	1	55		4
/l/ + /end_sil/		292	8	642		9	4	11
/gyez/		1	4	707		101		4
/kil/		152	271	6304		72		9

Az adatok azt mutatják, hogy a hibák előfordulása egyrészről szövegfüggő (más témájú szövegben esetleg elő sem fordul az adott hangkapcsolat), másrészről a beszélőtől is függ.

PÉLDA: /a/ + /m/ + /f/ kapcsolat

HIBA: az /m/ hang teljes egészében az /f/ első felére van címkézve, így a nazális hang zöngétlen szakaszra esik *(3. ábra felső része).*

JAVÍTÁSI szubrutin: az /m/ jobb oldali határát az /f/ kezdetére (zöngés-zöngétlen váltási pont) kell elcsúsztatni, a jobb oldali határt pedig az /a/ /m/ hangkapcsolat belsejében kell kijelölni a következők szerint:

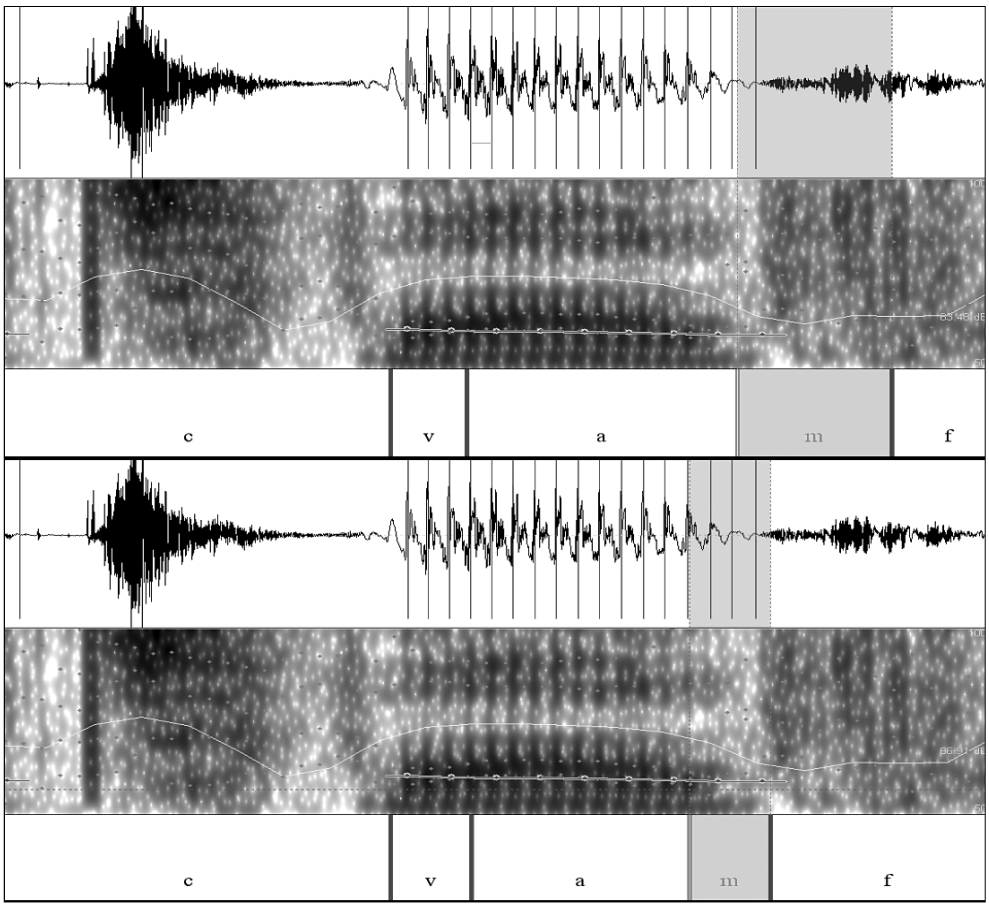
– amennyiben a zöngés szakaszra a /v/ /a/ /m/ hangok vonatkoznak, akkor az időtengely mentén 20-60-20%-ban kell felosztani *(3. ábra lent)* a három zöngés hang teljes időtartamát.
– ha csak /a/ /m/ hangok vannak a zöngés szakaszban, akkor 50-50%-ban kell felosztani.

A durva hibákat kijavító szubrutinokban döntően a zöngés-zöngétlen szakaszok váltakozására támaszkodunk, valamint a hangidőtartamokkal kapcsolatos kutatási eredményekre. Ez utóbbi alapján írjuk elő a több zöngés hangot tartalmazó zöngés szakaszokban az egyes hangokra vonatkozó időarányokat. A durva hibák nagy részét tehát célzott szubrutinokkal megszüntethetjük, a maradékot kézzel kell javítani (ezek a hibák általában egyediek, tehát gépi algoritmust nem éri meg rájuk készíteni). Az ilyen esetekben tehát célszerű felhasználni az emberi közreműködést.

Példák a kézzel javítandó hibákra:

- /s/, /sz/, /c/, /cs/ + /m/, /n/, /ny/ kapcsolatoknál a kapcsolatban résztvevő nazális hang baloldali hanghatára az előző zöngétlen hang végére van címkézve, nagy részben zöngétlen szakaszra. A javítás során a nazális mássalhangzó bal hanghatárát a zöngétlen-zöngés váltópontig kell eltolni, majd a zöngés szakaszra a hanghatárt úgy állítjuk be, hogy a nazális hang és a következő magánhangzó 20-80%-ban osztozzon a kapcsolat időtartamán.
- Felesleges szünetek (nem rövid /sil/, de például zöngés, vagy nagy intenzitású rész /sil/ jelöléssel), ezek általában glottalizáció környékén jelentkeznek például az /o3/ + /sil/ + /o2/ jelölésből /o3/ /o2/ lett a kézi javítás után.
- Rövidre módosult zárszakasz helyes jelölése például a nazális után az /m/ + /b/, /n/ + /c/ esetén
- Átírási hiba (szinkron hiba) a szövegben *kétezer,* a hangban /k/ /e/ /t:/ /o3/ /e/ /z/ /e/ /r/. Ilyenkor több hibásan címkézett hang lesz egymás mellett.
- Számok átírásánál előforduló hiba, a felismerő nincs alternatív kiejtésre tanítva, például: /k/ /i/ /l/ /e/ /n/ /c/ /sz/ /a1/ /sz/, illetve /k/ /i/ /l/ /e/ /n/ /c:/ /a1/ /sz/.

A hibajavítás első szakaszának végére a négy adatbázison összesen 7385 esetben javított a gépi algoritmus durva hanghibát, 4383 esetben pedig a hibás szü-



3. ábra
Példa a nazális hang rossz címkézésére az /amf/ kapcsolatban az N2 jelű adatbázisból (fent). Az automatikus javítás utáni állapotot a lenti részen láthatjuk.

net-jelölések korrekciójára került sor. A kézi javítások száma összesítve 1172. A hibajavítás végére tehát olyan adatbázisaink lettek, amelyekben minden hang a neki megfelelő akusztikai tartalomhoz van jelölve a hullámformában, csak kisebb mértékű hanghatár-elcsúszások lehetnek még jelen, mint hibák. Ezek száma a négy adatbázisban 30658 volt. Ezek javításáról lesz szó a következő pontban.

5.2. A kismértékű hanghatár-elcsúszások feltárása és javítása

A kismértékű hanghatár-elcsúszások feltárásánál és javításánál ugyanazon módszert alkalmaztuk, mint amit a durva hibáknál (lista, tipizálás, szubrutin). Az ilyen hibák nagy része a zöngés-zöngétlen átmenetek határán jelentkezik olyan formában, hogy a tényleges átmeneti ponttól (ami 10-15 ms-os sávra tehető) távolabb van megjelölve a hanghatár, mint ahogy kellene.

Példa:
a *...forintos lebeszélhetőség...* szövegrész szóhatárán az /o/ /s/ hangkapcsolatban az /s/ hang kezdete az /o/ hang 70%-os pontjára van jelölve, tehát zöngés része is van a zöngétlen réshangnak *(4. ábra)*.

A tervezett algoritmussal az összes hasonló hibát sikerült megszüntetni, kézi javításra egyáltalán nem volt szükség.

6. Eredmények

Az új eljárás fejlesztése során világossá vált, hogy a beszédjel gépi címkézéséhez minden olyan információt fel kell használni, ami jellemzi a beszédjelet, tehát kombinálni kell a jelfeldolgozási és a fonetikai eljárásokat, modelleket a pontosabb címkézéshez. Mindezekhez azonban hozzá kell tenni, hogy az itt ismertetett módszer nem alkalmazható automatikus feldolgozásra, mivel emberi tényező is szerepel benne és a feldolgozási idő is hosszúnak tekinthető.

Az új eljárással két-három napi munkával el lehet végezni egy új, több órányi hanganyagot tartalmazó beszédadatbázis címkézését. Az ilyen feldolgozás közel 100%-os pontosságú címkézést eredményez, ami a későbbi használat során fontos tényező lehet.

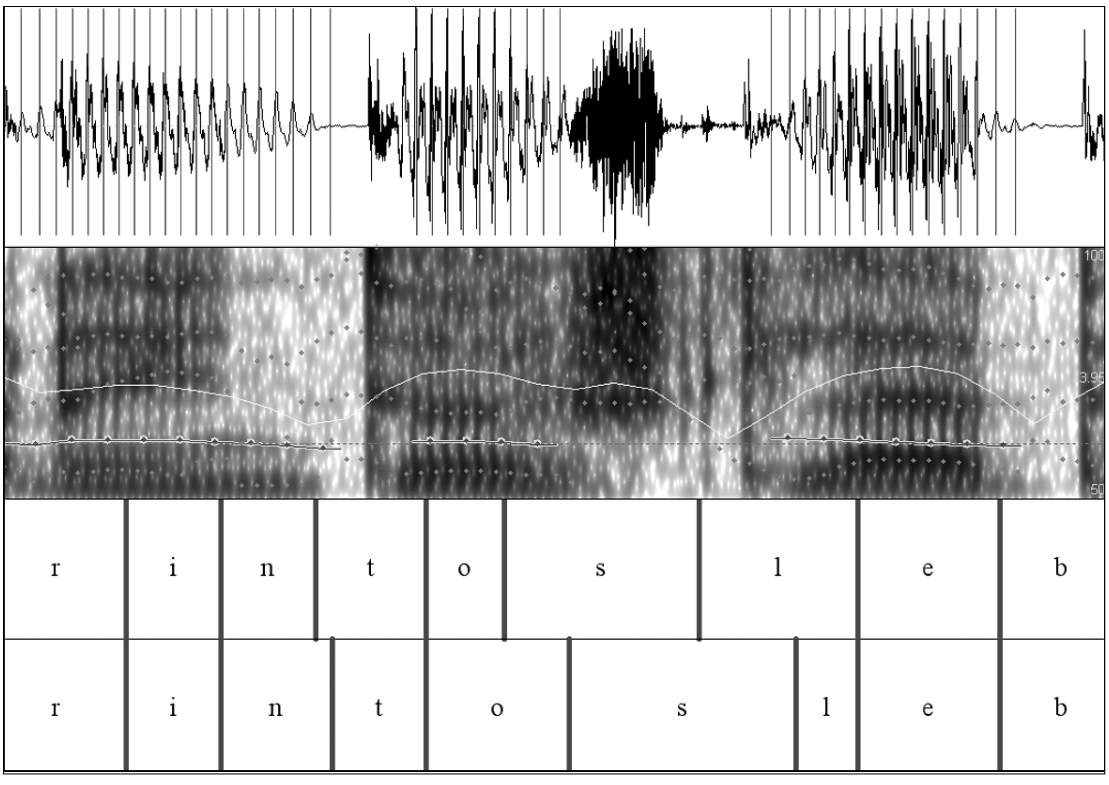
7. Összefoglalás

A több órányi beszédet tartalmazó beszédadatbázisok címkézésénél alapvető gond, hogy a címkézést teljes mértékben emberi erővel nem lehet elvégezni. A cél viszont az, hogy a címkék a lehető legpontosabban legyenek bejelölve a hullámformában.

A fent ismertetett új hibrid eljárást eredményesen lehet alkalmazni az ilyen munkákhoz, szinte hibamentes címkézés érhető el, időigénye is elviselhető (2-3 nap). Az így előkészített beszédadatbázisokban való keresés megbízható eredményeket ad. Ezt fel lehet használni a beszédkutatásban, a beszédszintézisben és a beszédfelismerésben is.

Köszönetnyilvánítás

A kutatást részben az NKTH támogatta a Jedlik Ányos program keretében (TELEAUTO projekt).



4. ábra
Példa a hanghatár elcsúszására (felső hangsor) és annak gépi javítására (alsó hangsor)

A szerzőkről

Olaszy Gábor 1967-ben végzett a BME Villamosmérnöki Kar Híradástechnikai szakán. 1975 óta foglalkozik beszédkutatással, fonetikával (MTA Nyelvtudományi Intézet 2006-ig). Kutatási területei: a beszéd akusztikai szerkezete, szegmentális és szupraszegmentális elemek kutatása, fonetikai modellezés, beszédtervezés, címkézési hibajavító algoritmusok tervezése, többnyelvű szöveg-beszéd átalakító beszédszintetizáló rendszerek tervezése, hullámforma szintézis fonetikai alapjainak kutatása, professzionális beszédkeltők tervezése, készítése, tesztelése. 1983 óta dolgozik a BME Távközlési és Médiainformatikai Tanszék beszédkutató csoportjában is.

Zainkó Csaba 1999-ben végzett a BME Villamosmérnöki és Informatikai Kar Médiainformatika szakirányon és azóta a Távközlési és Médiainformatikai Tanszék Beszédtechnológiai laboratóriumában dialógusrendszerek és az ahhoz kapcsolódó komponensek kutatásával és fejlesztésével foglalkozik. Részt vett az első magyar nyelvű elektronikus levél felolvasó és a számszerinti tudakozó fejlesztésében. Jelenleg a korpusz alapú beszédszintézis technológiájának vizsgálata áll kutatási témájának középpontjában.

Bartalis Mátyás 2005-ben végzett a BME Villamosmérnöki és Informatikai karán, Médiainformatika szakirányon. Oklevele megszerzése óta a BME Távközlési és Médiainformatikai tanszékének Beszédtechnológiai laborjában dolgozik. Fő tevékenysége a beszédszintetizátorokon alapuló alkalmazások fejlesztésében való részvétel, valamint a beszédszintetizátorok adatbázisainak fejlesztése, javítása.

Fék Márk 1997-ben végzett a BME Villamosmérnöki és Informatikai karán, Műszaki Informatika Szakon. 1997-2001 között francia-magyar közös doktori képzésen vett részt a BME-n és a francia ENST-Bretagne-on. Doktori disszertációját beszéd és audio jelek tömörítése témakörében 2006-ban védte meg. 2001-től a BME Távközlési és Médiainformatikai Tanszékén magyar nyelvű beszédszintézissel foglalkozik. Főbb kutatási területei a korpusz alapú beszédszintézis és az érzelemszintézis.

Németh Géza a BME Villamosmérnöki Karán 1983-ban végzett, 1985-ben szakmérnöki diplomát szerzett. 1985-87 között a BEAG Elektroakusztikai Gyárban fejlesztőmérnökként dolgozott, 1987-től a BME Távközlési és Médiainformatikai Tanszékén oktat (Méréstechnika, Kommunikációs rendszerek, Híradástechnika, A jelfeldolgozás elemei, Távközlés, Távközlésmenedzselés, Beszédinformációs rendszerek). Jelenleg a tanszék beszédtechnológiai laboratóriumát is vezeti. Irányító szerepet tölt be a beszédkutatási eredmények gyakorlatba való átültetésében, számos gyakorlati alkalmazást az ő vezetésével fejlesztettek ki.

Mihajlik Péter 1999-ben végzett a BME Villamosmérnöki és Informatikai Karán, Villamosmérnöki Szakon, távközlési főszakirányon és orvosbiológiai technika mellékszakirányon. A gépi beszédfelismeréssel PhD hallgatóként 1999-ben kezdett foglalkozni. 2002 óta a BME Távközlési és Médiainformatika Tanszékén dolgozik – jelenleg tudományos segédmunkatársi minőségben – ahol elsősorban a magyar nyelvű gépi beszédfelismerés témakörében végez kutatásokat.

Irodalom

[1] Paul Boersma P., D. Weenink:
Doing Phonetics by Computer [Comp. software], 2005.
www.praat.org

[2] Mihajlik Péter, Tatai Péter:
Automatikus fonetikus átírás magyar nyelvű beszédfelismeréshez.
In: Gósy Mária (szerk.), Beszédkutatás 2001,
MTA Nyelvtudományi Intézet, Budapest.
pp.172–185.

[3] Mihajlik Péter, Tatai Péter, Gordos Géza:
Automatic Phonetic Transcription and Its Application in Speech Recogniser Training:
A case study for Hungarian
In: Divenyi P., Greenberg S., Meyer G. (ed.),
Dynamics of Speech Production and Perception,
Amsterdam: IOS Press, 2006.
pp.245–262. (NATO Science Series, I.) 374,
Life and Behavioural Sciences.

[4] Olaszy Gábor:
A nazálisok okozta szerkezetváltozás a zár-, rés- és zár-rés hangoknál mássalhangzó kapcsolatokban.
In: Gósy Mária (szerk.), Beszédkutatás 2006,
MTA Nyelvtudományi Intézet, Budapest.
pp.32–43.

[5] Olaszy Gábor:
Mássalhangzó-kapcsolódások a magyar beszédben.
Tinta Kiadó, Budapest, 2007.

[6] Olaszy G., Németh G., Olaszi P., Kiss G., Zainkó Cs., Gordos G.:
Profivox – a Hungarian TTS System for Telecommunications Applications.
International Journal of Speech Technology. Vol. 3-4.
Kluwer Academic Publishers, 2000.
pp.201–215.

[7] Vicsi Klára, Vígh Attila:
Az első magyar nyelvű beszédadatbázis.
In: Gósy Mária (szerk.), Beszédkutatás 1998,
MTA Nyelvtudományi Intézet, Budapest.
pp.163–177.

[8] Vicsi, Klára:
Beszédadatbázisok a gépi beszédfelismerés segítésére,
Híradástechnika 2001/1, Budapest.
pp.5–13.

Beszéd kiemelése zajból a rekonstruált fázistérben

PINTÉR ISTVÁN

Kecskeméti Főiskola GAMF Kar, Kalmár Sándor Informatikai Intézet,
Automatizálási és Alkalmazott Informatikai Szakcsoport
pinter.istvan@gamf.kefo.hu

Lektorált

Kulcsszavak: beszéd kiemelése zajból, jel-altér, rekonstruált fázistér, dimenzió-beágyazás

A dolgozat beszédjel zajból való kiemelésére szolgáló módszert ismertet, amely a rekonstruált fázistér és a dimenzió-beágyazás fogalmaira épül. Az algoritmus a beszédet a zajtól a transzformált térben elvégzett nemlineáris művelettel választja szét. A dolgozat a beszéddel nem korrelált additív zaj esetében elért jelenlegi eredményeinket mutatja be.

1. Bevezetés

A gépi beszédfeldolgozásban régóta meglévő feladat a beszédjel kiemelése a zajból [1]. Az elmúlt mintegy három évtizedben több módszert is kidolgoztak a zaj csökkentésére. A legtöbb eljárásban közös az a feltételezés, hogy a lineáris beszédjel-modellnek megfelelő zajjellemzők időben lassan változnak. Példaként a hallásmodell-alapú szűrősorral végzett feldolgozás említhető, ahol a rész-sávokban Wiener-szűrést alkalmaznak [2]. A tapasztalatok szerint mintegy 6...9 dB-nél nagyobb jel/zaj viszony (SNR) esetén érhető el e módszerekkel jó eredmény [3]. Ennél kisebb SNR illetve nemstacionárius zaj esetén a zajcsökkentő módszerek alapja többnyire nemlineáris modell. Nemlineáris rendszermodellre példa az emberi hallórendszer egy modellje, amit zajcsökkentő eljárásban is alkalmaznak [4], míg nemlineáris jelmodell a jelen dolgozatban is vizsgált beszédmodell, nevezetesen a rekonstruált fázistérben történő beszédábrázolás [5].

A cikk felépítése a következő. Először ismertetjük a zajmentes beszédjel ábrázolását, amit a továbbiakban felhasználunk. Ezt követi a beszéd-altér fogalmára alapozott zajcsökkentési eljárás leírása a szakirodalom alapján, ahol bemutatjuk a rekonstruált fázistérben működő változatot is. A negyedik szakaszban ismertetjük az eljárás megvalósításával elért zajcsökkentési eredményeinket, majd a cikket a következtetések zárják.

2. A zajmentes beszédjel ábrázolása a transzformált térben és a rekonstruált fázistérben

A cikkben ismertetett zajcsökkentő eljárás két feltételezésre épül: az egyik, hogy létezik a beszédminta-sorozat optimális ábrázolása, a másik pedig, hogy a beszédfeldolgozási feladatokhoz is használható a rekonstruált fázistér fogalma.

Ami az első feltételezést illeti, ebben az esetben a feldolgozás alatt álló szegmens N számú α_n beszédmintájából vektort alkotunk, így a szegmens az N dimenziós tér egy pontjának felel meg. Ez az $\underline{s}$ vektor az úgynevezett $\{\underline{t}_n\}$ természetes, ortonormált bázisban a bázisvektorok lineáris kombinációjaként írható fel, ahol az együtthatók a beszédminták: $\alpha_n=(s,\underline{t}_n)$, és az N dimenziós $\underline{t}_n$ oszlopvektor n-edik komponense 1, a többi 0.

A gépi beszédfeldolgozás gyakorlati problémáinak megoldása során szerzett tapasztalatok szerint a beszédjel esetén létezik olyan bázis, amelybéli reprezentációban N-nél kevesebb számú összetevővel is leírható az $\underline{s}$ vektor [6]. Emiatt fel lehet tenni azt a kérdést, hogy melyik az az ortonormált bázis, amelyben $\underline{s}$ kevesebb összetevővel adható meg

$$\hat{\underline{s}} = \sum_{n=0}^{L-1} a_n \cdot \underline{v}_n \tag{1}$$

alakban, ahol $\{\underline{v}_n\}$ a keresett ortonormált bázis és L<N, továbbá ez az előállítás optimális abban az értelemben, hogy a

$$J(\underline{e}) = E\left\{\|\underline{e}\|^2\right\} = E\left\{\|\underline{s}-\hat{\underline{s}}\|^2\right\} = \\ = \sum_{n=L}^{N-1} \underline{v}_n^T \cdot E\left\{\underline{s}\cdot\underline{s}^T\right\} \cdot \underline{v}_n = \sum_{n=L}^{N-1} \underline{v}_n^T \cdot \underline{\underline{R}} \cdot \underline{v}_n \tag{2}$$

kritériumfüggvény értéke, vagyis az eltérésnégyzet várható értéke minimális (ideális esetben L<N és $||\underline{e}||=0$). A továbbiakban feltesszük, hogy $E\{\underline{s}\}=0$, így $\underline{\underline{R}}=\underline{\underline{K}}$, ami a kovarianciamátrix. Ismeretes, hogy a $\{\underline{v}_n\}$ optimális ortonormált bázist a kovarianciamátrix sajátvektor-rendszere adja, az eltérésnégyzet várható értéke pedig

$$J(\underline{e}) = \sum_{n=L}^{N-1} \lambda_n$$

ahol λ_n a kovarianciamátrix n-edik sajátértéke. Az $\underline{s}$ beszédminta-vektor új (optimális) bázisbeli reprezentációját a

$$\underline{\underline{T}} = \left(\underline{v}_0^T, \underline{v}_1^T, \ldots, \underline{v}_i^T, \ldots, \underline{v}_{N-1}^T\right)^T \tag{3}$$

mátrix segítségével lehet kiszámítani, ami praktikusan rendre a nagyság szerint csökkenő sorrendben felírt sajátértékeknek megfelelő sajátvektorokból, mint sorvektorokból áll.

A második feltételezés a beszédjelnek a rekonstruált fázistérben történő ábrázolására vonatkozik. A rekonstruált fázistér fogalma a diszkrét dinamikai rendszer $\underline{x}_{n+1}=\underline{F}(\underline{x}_n)$ alakú mozgásegyenletéhez kapcsolható, ahol $\underline{x}_n$ és $\underline{x}_{n+1}$ a D dimenziós fázistérbeli pontok, $\underline{F}$ megfelelő leképezés. A fázistérbeli pontok $\{\underline{x}_n\}$ halmaza a trajektória, vagy pálya. Ez a pálya közvetlenül nem figyelhető meg, csak az $\underline{x}_n \rightarrow g(\underline{x}_n)$ nemlineáris leképezésen át. Így adódik az $\alpha_n = g(\underline{x}_n)$ megfigyelhető (mérhető) valós szám, vagy beszédminta. Ezeket rendre egymás után T_{MV} időnként véve adódik az $\{\alpha_n\}$ beszédminta-sorozat.

Igazolható, hogy ha $M>2\cdot D+1$, akkor az α_n számsorozatból az eredeti $\{\underline{x}_n\}$ vektorsorozattal ekvivalens $\{\underline{y}_n\}$ vektorsorozat állítható elő az úgynevezett dimenzióbeágyazás módszerével. A dimenzióbeágyazás az

$$\underline{y}_n(M,\tau)=\left(\alpha_n,\alpha_{n+\tau},\ldots,\alpha_{n+(M-1)\cdot\tau}\right) \tag{4}$$

vektort eredményezi, ahol $\tau>0$ az időeltolás (itt mintaszámmal adott), $M>0$ a beágyazási dimenzió. A fentebb említett ekvivalencia azt jelenti, hogy létezik olyan egyértelműen invertálható, sima $\underline{h}:\underline{y}_n\ (M,\tau)\rightarrow\underline{x}_n$ leképezés, amivel a két vektorsorozat egymásba átvihető [7]. A dimenzióbeágyazás műveletéhez szükséges M dimenzió és a τ időeltolás értékét numerikus kísérletekkel lehet meghatározni az adott beszédtechnológiai alkalmazási feladathoz. A szakirodalmi adatok szerint az $M\cdot\tau\cdot T_{MV}$ beágyazási időablak 1...5 ms [8].

3. Zajcsökkentés a rekonstruált fázistérben altér módszerrel

A rekonstruált fázistér fogalmával leírható zajcsökkentő algoritmus lényegében általánosítása egy, a szakirodalomban régebben közölt eljárásnak [6], ezért először ezt ismertetjük.

A módszer alapja az előző pontban ismertetett tulajdonság, vagyis az, hogy az optimális beszédábrázolás az N számú bázisvektor helyett L számú bázisvektorral is megoldható, ideális esetben zérus eltérésnégyzettel. Az N mintából álló beszédvektor így az N dimenziós tér L dimenziós alterében található, emiatt ezt beszéd-altérnek is nevezik. A zajcsökkentő algoritmus ebben a beszéd-altérben állít elő optimális becsült beszédvektort a kiindulásképpen rendelkezésre álló zajos beszédmintákból. A feladat az, hogy a beszédjellel nem korrelált, additív zajjal terhelt

$$\underline{u}=\underline{s}+\underline{w} \tag{5}$$

beszédmintasorozat ismeretében adjuk meg a tiszta beszédvektor $\underline{\tilde{s}}$ becslését úgy, hogy az $\underline{s}-\underline{\tilde{s}}$ eltérésvektor hosszának várható értéke a legkisebb legyen, azaz

$$E\left\{\left\|\underline{s}-\underline{\tilde{s}}\right\|^2\right\}\rightarrow\min \tag{6}$$

teljesüljön. Hasonlóan az előző szakaszban foglaltakhoz, itt is meg kell találni az optimális ortonormált bázist, ám most csak $\underline{u}$ ismert. Feltesszük, hogy $E\{\underline{w}\}=0$, és mivel előző feltevésünk miatt $E\{\underline{s}\}=0$, adódik $E\{\underline{u}\}=0$. További feltevés, hogy a 0 várható értékű zaj típusa fehér zaj, vagyis kovarianciamátrixa $\underline{\underline{K}}^{ZAJ}=\sigma^2\cdot\underline{\underline{I}}$, ahol $\sigma>0$ és $\underline{\underline{I}}$ NxN méretű egységmátrix. Belátható, hogy ekkor a korrelálatlanság miatt a zajos beszéd kovarianciamátrixa a beszéd és a zaj kovarianciamátrixok összege:

$$\underline{\underline{K}}^{ZAJOS}=E\left\{\underline{u}\cdot\underline{u}^T\right\}=\underline{\underline{K}}^{BESZÉD}+\underline{\underline{K}}^{ZAJ}, \tag{7}$$

továbbá igazolható az is, hogy a zajos beszéd kovarianciamátrixának és a zajmentes beszéd kovarinaciamátrixának sajátvektorai azonosak. Ez utóbbi tulajdonság teszi lehetővé, hogy a zajcsökkentés a meglévő zajos beszédvektorból kiindulva ugyanabban az ortonormált bázisban végezhető el, mint amiben a beszéd reprezentációja optimális és az optimális beszédreprezentációt adó $\{\underline{v}_n\}$ ortonormált bázis kiszámítható a rendelkezésre álló zajos beszéd kovarianciamátrixából is, a zajmentes beszédvektor kovarianciamátrixának ismerete nélkül.

Továbbá, mivel a kovarianciamátrixok összegezhetők, belátható, hogy a zajos beszéd kovarianciamátrixa a transzformált térben a következő diagonálmátrix lesz:

$$\underline{\underline{K}}^{ZAJOS}\Big|_{\{\underline{v}_n\}}= \\ =\operatorname{diag}\left(\lambda_0+\sigma^2\ \ldots\ \lambda_{L-1}+\sigma^2\ \ \sigma^2\ \ldots\ \sigma^2\right). \tag{8}$$

Az első szakaszban ismertetett feltevésünknek megfelelően a $\underline{v}_0,\underline{v}_1,\ldots,\underline{v}_{L-1}$ vektorok által kifeszített altérben a zajmentes beszéd ideálisan reprezentálható. Szemléletesen szólva a zajos beszéd esetében itt „beszéd és zaj is található”, míg ezen altér ortogonális komplemensében, a $\underline{v}_L,\underline{v}_{L+1},\ldots,\underline{v}_{N-1}$ vektorok által kifeszített altérben „csak zaj található”.

Ezek után a zajcsökkentő eljárást a $\underline{\underline{H}}$ lineáris transzformáció alakjában keressük, vagyis

$$\underline{\tilde{s}}=\underline{\underline{H}}\cdot\underline{u}\,. \tag{9}$$

A becslés hibája az $\underline{r}=\underline{s}-\underline{\tilde{s}}$ maradékjel. A [6] szerzői megmutatták, hogy az

$$\underline{r}=\underline{r}^{BESZÉD}-\underline{r}^{ZAJ} \tag{10}$$

maradékjel két összetevőből áll, egy a beszéddel, egy a zajjal korrelált. Emiatt nemcsak a beszéddel korrelált hibaösszetevő minimalizálása a feladat, hanem ezzel egyidejűleg a zajjal korrelált összetevő előírt szint alatt tartása is cél az optimális lineáris transzformáció keresésekor. A feladatot [6]-ban mind az időtartományban, mind a spektrális tartományban előírt feltételek esetében megoldották. Saját, időtartományra vonatkozó eredményeinket [9]-ben tettük közzé. A második esetben is a beszéddel korrelált hibaösszetevő minimalizálása a cél, de most minden egyes spektrális komponensre külön-külön írunk elő zajszintcsökkentési feltételt, azaz

$$J\left(\underline{r}^{BESZÉD}\right)\xrightarrow[\underline{\underline{H}}]{}\min \tag{11}$$

feltéve, hogy:

$$E\left\{\left|\underline{v}_n^T\cdot\underline{r}^{ZAJ}\right|^2\le\beta_n\cdot\sigma^2\right\}\ \ n=0,1,\ldots L-1, \\ E\left\{\left|\underline{v}_n^T\cdot\underline{r}^{ZAJ}\right|^2=0\right\}\ \ n=L,\ldots N-1. \tag{12}$$

Szemléletesen szólva az első feltétel a beszéd-altérben megmaradt zajra vonatkozó komponensenkénti előírás rendre β_n>0-val megadott feltételekkel, a második a zaj-altérbeli komponensek nullázását írja elő.

Az optimális transzformáció mátrixa a Karush-Kühn-Tucker-féle feltételes szélsőérték-keresési módszer alapján [6] szerint a következő:

$$\underline{\underline{H}}^{OPT} = \underline{\underline{V}} \cdot \underline{\underline{G}} \cdot \underline{\underline{V}}^{T},$$
$$\underline{\underline{G}} = \mathrm{diag}\left(g_{0,0},\ldots,g_{N,N}\right),$$
$$g_{n,n} = \begin{cases} \sqrt{\gamma_n} & n = 0,1,\ldots,L-1 \\ 0 & n = L,\ldots,N-1 \end{cases} \quad (13)$$

ahol $\underline{\underline{V}}$ a sajátvektorokból, mint oszlopvektorokból álló mátrix. Az említett mű γ_n megválasztására két javaslatot is ad, ezek közül a jelen munkában a

$$\gamma_n = \exp\left(-\frac{\kappa \cdot \sigma^2}{\lambda_n^{BESZÉD}}\right) \quad (14)$$

összefüggéssel dolgoztunk. A zajcsökkentés mértékét a $\kappa \geq 1$ tapasztalati konstanssal lehet beállítani, egyben a tisztított beszéd torzulását is befolyásolva ezzel.

A fent összefoglalt módszer általánosítható a rekonstruált fázistér esetére is. Ugyanis ez utóbbi, mint modellháttér lehetővé teszi, hogy egyetlen megfigyelt $\underline{u} = \underline{s} + \underline{w}$ N dimenziós zajos beszédvektorból állítsunk elő M dimenziós vektorokból álló adatrendszert a dimenzió beágyazás módszerével. Az így keletkező $\underline{\underline{U}}_{MxN}$ trajektóriamátrixra a konstrukciója miatt igaz, hogy

$$\underline{\underline{U}}_{MxN} = \underline{\underline{S}}_{MxN} + \underline{\underline{W}}_{MxN}. \quad (15)$$

Mivel $u_n = s_n + w_n$ minden összetartozó mintára fennáll, a trajektóriamátrixban lévő vektorokból, mint adatrendszerből számolható kovarianciamátrixra nézve teljesül, hogy

$$\underline{\underline{K}}_{\underline{\underline{U}}_{MxN}} = \underline{\underline{K}}_{\underline{\underline{S}}_{MxN}} + \underline{\underline{K}}_{\underline{\underline{W}}_{MxN}}, \quad (16)$$

ahol $\underline{\underline{K}}_{\underline{\underline{W}}_{MxN}} = \sigma^2 \cdot \underline{\underline{I}}_{MxM}$

Emiatt az előzőekben ismertetett gondolatmenet most is alkalmazható, amivel előállítható a becsült $\underline{\tilde{S}}$ trajektória-mátrix. Ebből kell a becsült $\underline{\tilde{s}}$ beszédminta sorozatot visszaállítani, ami $\underline{U}$ előállítása alapján tehető meg. Az eredeti altér-módszertől ez az eljárás abban különbözik, hogy itt más adatrendszer konstruálható a kovariancia-mátrix becslésére és a becsült minta végleges értéke a becsült trajektória-mátrix egynél több eleméből számítható ki.

Az általunk használt trajektória-mátrix a beszédszegmens periodikus kiterjesztésén alapul, ezáltal minden minta pontosan M-szer szerepel, akárcsak a becsült trajektóriamátrixban is, így egyetlen becsült beszédminta előállítása átlagolással történhet és nem szükséges súlyozómátrix, ami más konstrukcióknál megjelenik [10].

A trajektóriamátrix $u_{i,j}$ eleme eszerint

$$u_{i,j} = u_{(j+i\cdot\tau) \bmod N}, \quad (17)$$

itt tehát N a beszédszegmens mintáinak száma, M a beágyazási dimenzió, τ az időeltolás.

A zajcsökkentési eljárás alapja a fenti trajektóriamátrix, mint M dimenziós vektorokból álló adatrendszer alapján becsülhető kovarianciamátrix. Megjegyezzük, hogy ez a kovarianciamátrix különbözik mind a [6]-ban használt empirikus Toeplitz-kovarianciamátrixtól, mind pedig az [5]-ben, illetve [10]-ben ismertetett változatoktól.

4. Zajcsökkentési eljárás megvalósítása, numerikus kísérleti eredmények

A jelen dolgozatban vizsgált kiinduló beszédállomány úgy jött létre, hogy a leírt mondatot egy magyar anyanyelvű, férfi bemondó valósította meg, a beszédmintákat 8 kHz mintavételi frekvenciával és 16 bites lineáris kvantálással állították elő. Az aktív beszédszakaszokon számolt globális jel/zaj viszony 45,8 dB volt. A zajos beszédállományokat ebből mesterséges zajosítással készítettük, a jelen munka során használt zajok az RSG-10 zaj adatbázisból származnak [11], 8 kHz mintavételi frekvenciájúra átalakítva az eredetileg 19980 Hz-cel mintavett jeleket.

A vizsgálatban használt zajtípusok: fehér zaj, rózsaszín zaj, hírközlő csatorna zaja. A zajszint beállításának alapjául a tiszta beszéd aktív beszédszakaszain számolt energia szolgált. A zajcsökkentés hatékonyságát az

$$SRR = 10 \cdot \lg\left(E^{BESZÉD} / E^{MARADÉK}\right), \quad (18)$$

(Signal to Residual Ratio) számmal jellemeztük, ahol a számlálóban az aktív beszédszakaszok indexhalmazán számolt beszédenergia, a nevezőben ugyanezen indexhalmazon számított, zajcsökkentés utáni maradékjel energiája szerepel.

Jel/zaj viszony (dB)	SRR (dB)		
	Fehér zaj	Nagyfrekvenciás hírközlő csatorna zaja	Rózsaszín zaj
15	9,3	9,3	9,0
12	9,1	9,0	8,3
9	8,7	8,6	7,3
6	8,1	7,9	5,7
3	7,2	6,9	3,7
0	6,0	5,7	1,8
-3	4,5	4,2	0,0

1. ábra
A javulás értékei különböző jel/zaj viszony értékek és zajtípusok esetén (szegmenshossz: 800 minta, beágyazási dimenzió: 20, időeltolás: 1 minta, beszéd-altér dimenziója: 7, a tapasztalati konstans: κ=5)

A zajcsökkentést időben átlapolt beszédszegmensek sorozatán haladva, rendre szegmensről szegmensre végeztük egyetlen szegmens mintái alapján. Az adott szegmens mintáit Hanning-ablakkal súlyoztuk, 50%-os átlapolást alkalmaztunk és a tisztított beszédminta sorozatot az átlapolás és hozzáadás módszerével számítottuk ki. A szegmens hosszát, a beágyazási dimenzió és az időeltolás értékét, a beszéd-altér dimenzióját és a κ tapasztalati konstanst a tisztított beszéd meghallgatása alapján határoztuk meg. Az előző szakaszban ismertetett spektrális tartománybeli módszerben szereplő γ_n értékekhez a σ^2 értékét az első zaj-altérbeli sajátértékkel, a $\lambda_n^{\text{BESZÉD}}$ értékeket a beszéd-altérbeli sajátértéknek a σ^2 becslésétől mért eltérésével becsültük. A sajátérték-sajátvektor számítást a Jacobi-módszerrel végeztük, a zajcsökkentő eljárást C nyelvű programmal valósítottuk meg.

Az *1. ábrán* (lásd az előző oldalon) összefoglaltuk a kapott számszerű adatokat, amik megfelelnek a szakirodalomban közölt eredményeknek [5,10]. A táblázatból kiolvasható, hogy az SRR-ben is megjelenő javulás 6 dB-nél kisebb SNR esetén mutatható ki. Ennek oka véleményünk szerint az, hogy a vizsgált eljárásnak nemcsak zajelnyomó, hanem beszédtorzító hatása is van, megfelelően a beszéddel korrelált maradékjel-összetevő létezéséről a 3. szakaszban leírtaknak. A módszer működőképességét szemlélteti az alábbi, időtartománybeli *2. ábra*, mely -3 dB SNR és fehér zaj esetében készült. Amellett, hogy a módszer zajcsökkentő képessége szembeötlő, az ábrán a beszédtorzító hatás is jól kivehető.

5. Következtetések

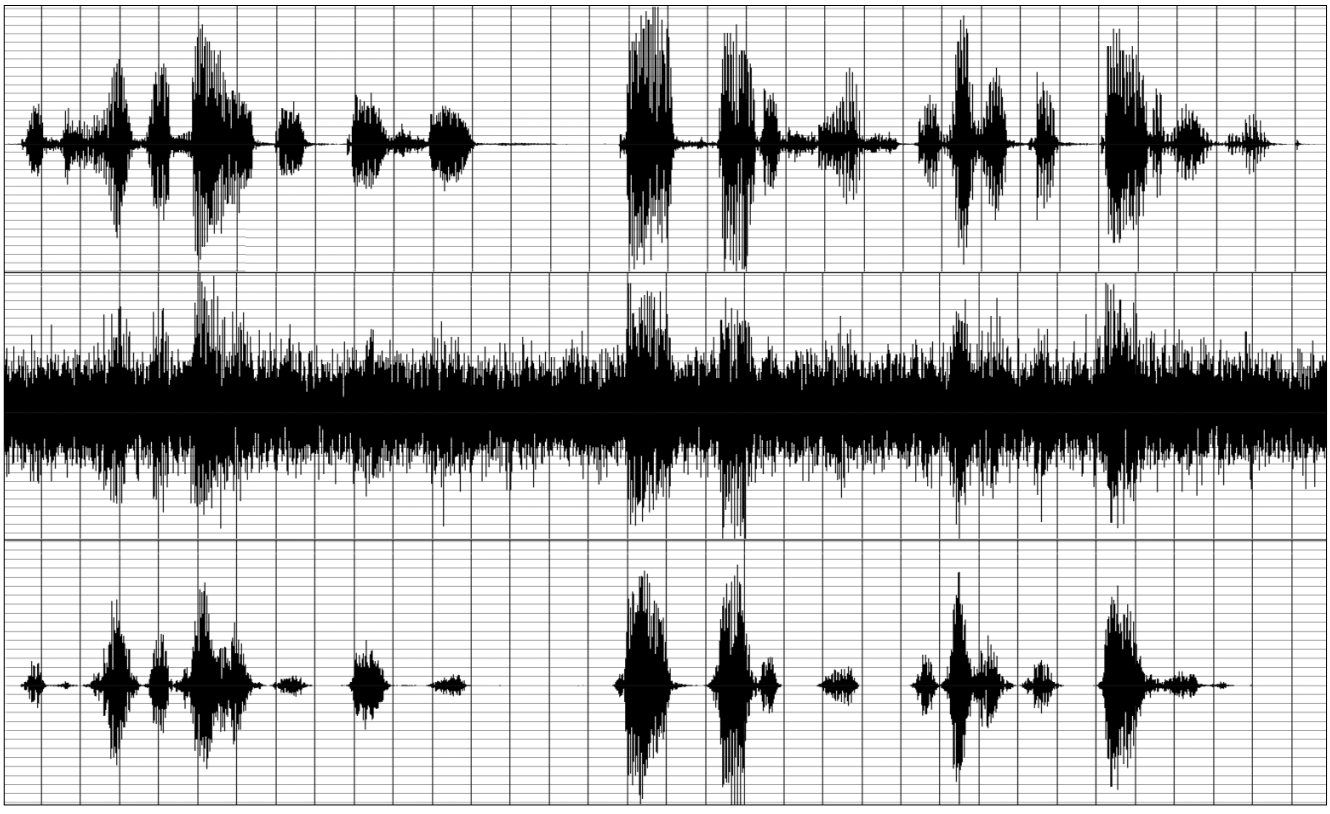
Dolgozatunkban rekonstruált fázistérben működő zajcsökkentő eljárást ismertettünk.

Az eljárás a dimenzióbeágyzásra épül és feltételezi, hogy a beszéd és a zaj altér elfogadhatóan választható szét a dimenzióbeágyzás után kapott adatrendszerből számítható optimális ortonormált bázis által kifeszített euklideszi térben. Az említett ortonormált bázist az adatrendszer kovarianciamátrixának sajátvektorai alkotják, amit a Jacobi-eljárással számítottunk ki. Az adatrendszert a tisztítandó beszédszegmens periodikus kiterjesztésével alkottuk meg, eltérően a szakirodalomban található megoldásoktól. A mi módszerünk nem igényel tapasztalati súlyozómátrixot a beszédminta becslésekor.

A programot egy magyar mondat zajosításával kapott zajos beszéddel teszteltük háromféle zaj és hétféle zajszint esetén. A javulást számszerűen is jellemeztük, a paramétereket a tisztított beszéd meghallgatásával állítottuk be. Megállapítható, hogy a legjobb eredményt körülbelül 100 ms hosszúságú szegmens, 50%-os szegmensléptetés, Hanning-ablak, átlapolás és hozzáadás típusú szegmentált feldolgozás, 20 dimenziós beágyazott tér, 1 minta beágyazási időlépés és 7 dimenziós beszéd altér esetén értük el.

Ezek a vonatkozó szakirodalomban jelenleg megtalálható adatoknak jól megfelelnek, numerikus kísérleteink alapján azt is mondhatjuk, hogy nemcsak fehér zaj, hanem a nagyfrekvenciás csatorna zaja és rózsaszín zaj esetén is. Ugyanakkor a közölt módszer fehér zajra

Zajcsökkentés -3 dB SNR és fehér zaj esetében a táblázatban leírt paraméterek mellett (felül: eredeti bemondás, középen: zajosított beszédmintasorozat, alul: a zajcsökkentés utáni mintasorozat)



kidolgozott, a színes zaj elnyomása nem optimális, ahhoz fehérítő transzformáció beépítése is szükséges.

További feladat a beágyazási dimenzió, az időeltolás és a beszéd-altér dimenzió értékeinek automatikus meghatározása, valamint ezek birtokában a zajcsökkentő eljárás módszeres tesztelése nagy beszéd-adatbázison.

Köszönetnyilvánítás

A szerző ez alkalommal is megköszöni Gordos Gézának, Németh Gézának és Tatai Péternek a segítséget és biztatást, amit beszédfeldolgozási célú algoritmusfejlesztési munkái során kapott.

A szerzőről

Pintér István 1983-ban kapta meg okleveles villamosmérnöki diplomáját a BME Villamosmérnöki Karán, majd ugyanitt PhD fokozatot szerzett a műszaki tudomány informatika tudományszakán 1997-ben. Az egyetem elvégzése után 1 évig a MIKI-ben dolgozott, majd 1984-től kezdve a GAMF-on, illetve 2000-től ennek jogutódján a KF GAMF Karán, jelenleg főiskolai tanárként. Fő érdeklődési területe a gépi beszédfeldolgozás, a digitális jelfeldolgozás és a jelfelismerés. Az első területen új beszédreprezentációk kidolgozásával és a beszédjel zajból való kiemelésével foglalkozik, a második témakörön belül elsősorban diszkrét ortogonális transzformációk jelfeldolgozási alkalmazásaival, a sort a mesterséges neurális hálózatok jelfelismerési célú alkalmazásai zárják. Az említett szakterületeken cikkei jelentek meg hazai és külföldi folyóiratokban, konferencia-kiadványokban. 2007-ben elnyerte a HTE Pollák-Virág díját.

Irodalom

[1] J.S. Lim, A.V. Oppenheim:
Enhancement and bandwidth compression of noisy speech.
Proc. IEEE 67 (12) 1979., pp.1586–1604.

[2] Yang Gui, Kwan, H.K.:
Adaptive subband Wiener filtering for speech enhancement using critical-band gammatone filterbank.
Proc. 48th Midwest Symposium on Circuits and Systems, 2005, Vol. 1., pp.732–735.

[3] Haci Tasmaz, Ergun Ercelebi:
Speech enhancement based on undecimated wavelet packet-perceptual filterbanks and MMSE-STSA estimation in various noise environments.
Digital Signal Processing,
(p.16, in press, available online 12 October 2007)

[4] T.F. Quatieri, R.B. Dunn:
Speech enhancement based on auditory spectral change.
Proc. International Conference on Acoustics, Speech and Signal Processing in Orlando, IEEE,
13-17 May 2002, pp.257–260.

[5] J. Sun, N. Zheng, X. Wang:
Enhancement of Chinese speech based on nonlinear dynamics.
Signal Processing 87 (2007), pp.2431–2445.

[6] Y. Ephraim, H.L. Van Trees:
A signal subspace approach for speech enhancement.
IEEE Trans. on Speech and Audio Processing,
July 1995, Vol. 3., No.4., pp.251–266.

[7] H. Kantz, T. Schreiber:
Nonlinear Time Series Analysis.
Cambridge University Press, 1997.

[8] G. Kubin, C. Lainscsek, E. Rank:
Indentification of Nonlinear Oscillator Models for Speech Analysis and Synthesis.
In: Chollet et al. (Eds.): Nonlinear Speech Modeling.
LN AI 3445, Springer Verlag 2005.,pp.74–113.

[9] I. Pintér:
Noise suppression using non-linear speech model.
Pollack Periodica, Vol. 2. Suppl., 2007.
Akadémiai Kiadó, pp.121–133.

[10] M.T. Johnson, R.T. Povinelli:
Generalized phase space projection for nonlinear noise reduction.
Physica D 201 (2005), pp.306–317.

[11] http://spib.rice.edu/spib/select_noise.html

IPTV hanginformáció siketek számára

TIHANYI ATTILA, FELDHOFFER GERGELY, OROSZI BALÁZS, TAKÁCS GYÖRGY

Pázmány Péter Katolikus Egyetem, Információs Technológiai Kar
tihanyia@itk.ppke.hu

Lektorált

Kulcsszavak: fejmozgatás, beszédjel átalakítás, látható beszéd, DirectShow rendszer

Magyarországon a televíziózás kezdetétől elterjedt, hogy az idegen nyelvű filmek szinkronizált magyar hanggal kerültek adásba, a magyar nézők ezt megszokták. Nő a hallássérültek tábora, a lakosság növekvő hányadának a TV hang a szokásos formában már nem elég. Kidolgoztunk egy megoldást, amely az aktuális beszédhangnak megfelelő szájmozgású fej képét jeleníti meg a képernyő sarkában. A rendszer valós időben működik, jelfolyamra, DVD-re, IPTV-re egyaránt alkalmazható, nyelvspecifikus részleteket nem tartalmaz, ezért bármilyen nyelvhez adaptálható. A megvalósításkor a Windows DirectShow rendszert használtuk keretként.

1. Bevezetés

A magyar TV nézők megszokták, hogy az idegen nyelvű filmek szinkronizált magyar hanggal kerülnek adásba és a többség ezt igényli ma is. Ennek korábban nyilvánvaló politikai indítékai voltak. Bár ez ma már nem áll fenn, de a nézettség adatai meghatározóak egy TV program gazdasági sikerességének szempontjából, így ez a gyakorlat megmaradt. Vannak más igények is és ezek kielégítésére új, működőképes műszaki megoldást kínálunk. A megoldás lényege az aktuális TV műsor beszédhangjának közvetlen átalakítása beszélő fej képévé. Részletesen taglaljuk a nagyothallók és siketek igényeit és az ezzel kapcsolatos elfogadott európai dokumentumokat.

A beszédjel közvetlen szájmozgássá alakításának alapelvét csak vázlatosan ismertetjük, mivel a Híradástechnika hasábjain már több cikkünk jelent meg erről. A megvalósítás újszerű eleme a Windows Direct Show rendszer alkalmazása, mivel ez változatos környezetben is egységes keretbe foglalja a képjelek és hangjelek lejátszását és egy új, helyben generált képrészlet beillesztését. Cikkünk leghosszabb szakasza ezzel foglalkozik.

2. A nagyothallók és siketek igényei, hatások a hallók társadalmára

Nagyothallóknál a hallott beszédhangot kiegészíti a látott szájmozgás. Kifejezi ezt a gyakran emlegetett mondás: „jobban hallom a tévét, ha felteszem a szemüvegem”. Ilyen esetekben (azon túl, hogy ne szinkronizált hang legyen) problémát okozhat az azonos nyelvű látott és hallott beszéd ellentmondása az időbeli eltérés miatt. Az amerikai TV nézőknél egészen sajátos esetet szült ez a probléma. A filmszalagon rögzített, (számos esetben még hangcsíkos) filmeknél műszaki okokból elcsúszhat időben egymástól a kép és hang. Különösen gyakran fordul elő ez a jelenség, ha egy TV adásban egymás után más technológiával rögzített felvételek kerülnek adásba. Annyira érzékenyek erre, hogy egy külön termék került kereskedelmi forgalomba a „lipsync” személyes kiegyenlítésére. Személyes beállító eszközzel állítható a kép-hang időeltérés pozitív vagy negatív irányban. A „lipsync”-re számos példa található, ennek szemléltetésére egy népszerű termékre hivatkozunk [1].

Amíg a magyar nézők jól hallanak, bevéső dik az agyukba, hogy amennyiben érteni is akarják a minden este nézett filmek eseményeit, csak a szinkronhangra érdemes figyelniük, mert ha közben a szájmozgást is nézik, abból csak zavar támad. Egy különös példa erre, hogy amíg a főszereplő figura az angol szöveg szerint azt mondja és tátogja „I am batman.” ezalatt a magyar szinkronszínész hangja azt mondja „Én a denevérember vagyok”. Még a szótagszám is több, mint a duplája. Minden tiszteletünk a jó szinkronszínészeké, akik még ilyet is vállalni kénytelenek és sokszor igen jól megoldják. Érdekes további példa a Frédi-Béni rajzfilmsorozat magyar hangja. A mindkét kultúrában otthonos nézők szerint a magyar hang szellemesebb és élvezetesebb, mint az eredeti, s mivel ez a rajzfilm eleve nem épít a pontos szájmozgás-képre, a magyar változat élvezetében nem zavar az elnagyolt, de más szájmozgás.

Más helyzet áll fenn a siketek és a nagyon töredékes hangot hallók körében. Náluk a szájmozgás képe nem kiegészítő információ, hanem a fő információforrás. A hétköznapi közvetlen kommunikációban nagyon kifinomult módon megtanulják a szájról olvasás művészetét. Munkánk során találkoztunk egyetemi diplomát szerzett kiválóságokkal, akik nem csak a könyvekből, hanem a jó előadók szájmozgásáról leolvasva szerezték meg tudásuk egy részét. Találkoztunk olyan sikettel is, aki cukrászdában eladóként dolgozott és soha nem tévesztette el, hogy a vevő két krémest, vagy három dobos tortát kért, mert pontosan megértette a szájmozgásból. Problémái abból adódtak, hogy amíg lehajolt az áruért, vagy a csomagolásra figyelt, azalatt módosította a vevő a rendelését és ezt nem észlelte.

A siketek számára tehát létkérdés a TV műsorokban a szájmozgás követése az események megértése szempontjából, de több fontos egyéb eset is felsorolható a szinkronizált játékfilmeken túl. A politikai, a magazin és a hírműsorokban gyakran betétrészletek láthatók alámondott hanginformációval. Sokszor olyan ábrát, mozgófilmet dokumentumfilm-részletet láthatunk, amelyet csak a hanginformáció tesz érthetővé. Népszerűek a természetfilmek, városok, tájak ismertetését tartalmazó műsorok is. Ezeknél narrátor mondja az alapvető információt, amelyet a képek, mozgóképek színesítenek, tesznek élvezetessé. Ezek üzenetének lényege nem érheti el a siket vagy erősen nagyothalló nézőket.

Ma Magyarországon 60 ezerre tehető a siketek száma, Európában 6,5 millió ember siket vagy súlyosan halláskárosodott. Ez több nemzeti kisebbség arányát is eléri. Egyes EU tagállamok többségi népessége sem tesz ki ekkora létszámot.

Számos, tudományosan megalapozott elmélet magyarázza, hogy miért nő jelentősen a siket és hallássérült újszülöttek, gyermekek aránya. Ugyanakkor a zajterhelés, a növekvő életkor és az ifjúkorban rendszeres és tartós hangos „zenehallgatás” egyik következménye, hogy a lakosság növekvő hányadának a TV hang a szokásos formában már nem elég. Többféle igény és megoldás megfogalmazódott erre [2,3]. A „Televíziózás Határok Nélkül” című EU direktíva tartalmazza, hogy lehetőleg minden műsort el kell látni felirattal vagy jelnyelvi kiegészítéssel. A jelnyelvi kiegészítésben tételesen szerepel a jelelés kézzel és kiegészítése szájmozgással. Ebben a kérdésben élénk viták zajlanak az érintettek és az őket segítő szakemberek körében. Mi a jobb egy TV műsor esetén: feliratozás vagy jelnyelvi tolmács? Ebbe a szakmai és társadalmi vitába mi nem szállunk bele érvekkel, vagy megfontolásokkal. Kínálunk viszont egy vadonatúj megoldást, amelyben az aktuális TV beszédhang (bármely nyelvű legyen is) kiegészíthető egy azonos idejű szájmozgás-képpel. A fennálló vitát ez nyilván nem dönti el addig, amíg nagyszámú siket néző ezt meg nem tanulja, meg nem szokja, esetleg meg nem szereti.

A feliratkészítés drága, a siket közösség nem is szereti, mert vagy a feliratot olvassa, vagy a filmet nézi. Ha a felirat nagyon szűkszavú, akkor nem érti, ha nagyon pontos, akkor végig sem tudja olvasni, mert előbb vált, mint ahogy a végére érne, ráadásul leköti teljes figyelmét az olvasás. Hallottunk olyan érvelést is, hogy talán ezzel a módszerrel lehetne megtanítani a halló tanulóifjúságot is olvasni, mert a jelen iskolarendszer nem képes kellő hatékonysággal a gyors szövegolvasás és megértés képességét elsajátíttatni.

Mi mindössze egy új eszközt kínálunk a siketek számára. Ennek lényege, hogy valós időben, korlátozott pontossággal a beszédhang jeléből előállítható a szájmozgás képe, bármely nyelvre. A látható szájmozgás ritmusa, tempója, időbeli szerkezete pontosan megfelel az elhangzó beszédhangnak a „lipsync” tűréshatárán belül. Annak eldöntésére, hogy a lehetséges felhasználók hazai 60 ezres vagy európai szinten a 6,5 milliós közössége számára az adott feladatra az ajánlott megoldás jó vagy sem, statisztikailag értékelhető igazolásunk még nincs. Cikkünk a javasolt megoldás műszaki alapjait foglalja össze.

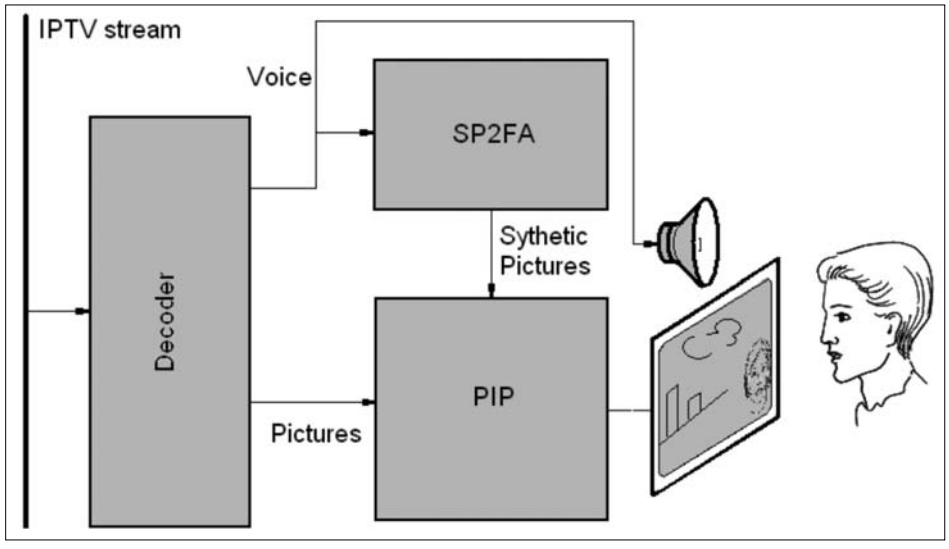
A (beszéd)hanggal vezérelt animált fej vagy száj ígéretes megoldás, mert:

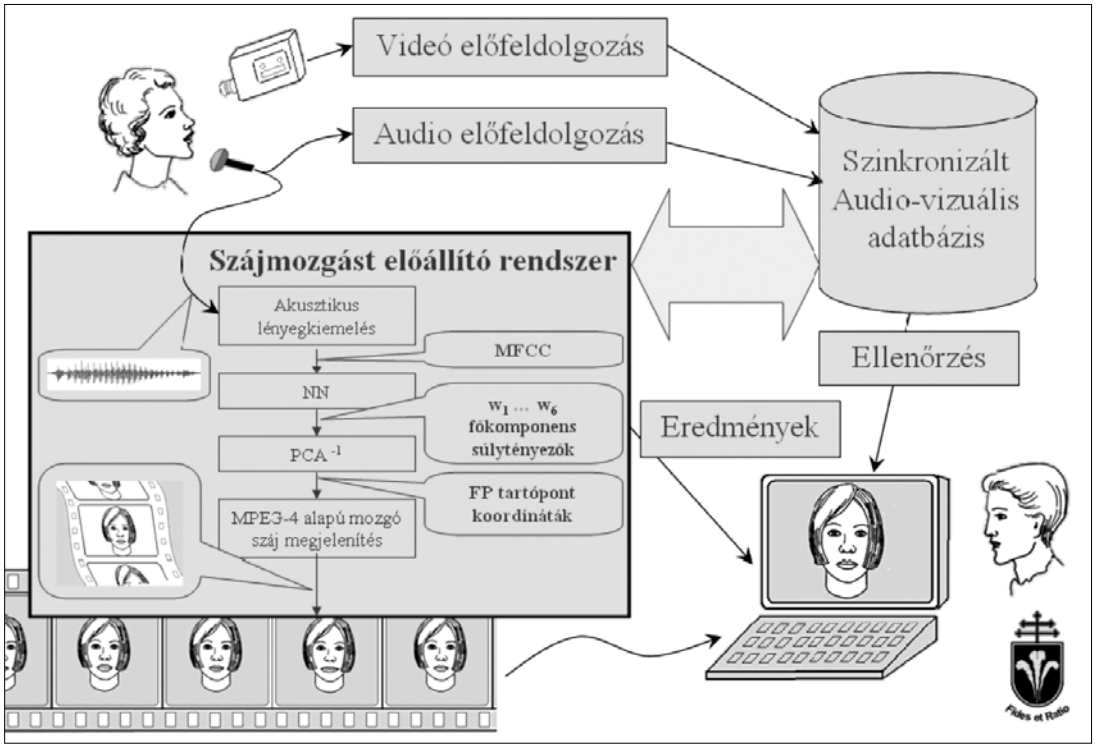
– a siketek deklarált fő igényeihez illeszkedik,
– nyelvfüggetlen megoldást sikerült megvalósítani alapszinten,
– teljesen automatizálható, szemben a feliratot készítő megoldással, tehát hatékony,
– ugyanaz a műszaki megoldás alkalmazható különböző technológiával továbbított vagy tárolt műsorok esetén is (analóg vagy digitális TV adás, IP-TV, DVD, állandó vagy változó sebességű jelátvitelnél is).

3. A TV képhez kapcsolódó beszédjelek átalakítása mozgó fej képévé

Az MPEG-4 kódolást multimédia-alkalmazások, mozgó fejek élethű megjelenítése figyelembe vételével fejlesztették. Egy általános célú, nyílt forráskódú fejmodellt alkalmassá tettünk mozgó száj képének megjelenítésére. Törekedtünk a számítási erőforrások minimális igénybe vételére, hogy az alkalmazás egy egyébként is használt eszközben – például „set top box” – megvalósítható legyen. Fontos eredménynek tartjuk, hogy az MPEG-4 animáció működik akkor is, ha nem képpontok mintavételezése alapján származtattuk a tartópont paramétereket, hanem beszédjelből számoltuk azokat.

1. ábra A teljes rendszer főbb elemei





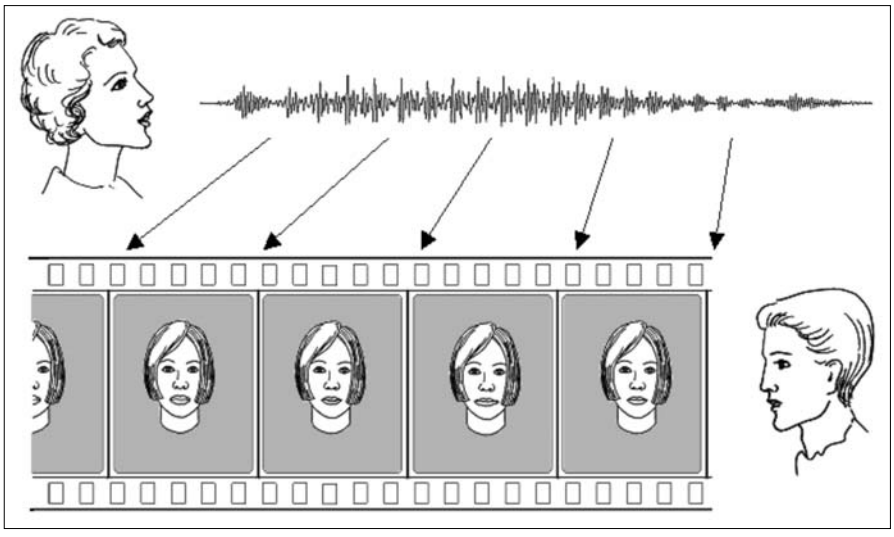
3. ábra
Az SP2FA rendszer felépítése

A teljes általunk kidolgozott rendszer alkalmas arra, hogy egy IPTV médiafolyamban (stream-ben) érkező TV műsor műsorhangjának felhasználásával egy szintetikusan előállított emberi fej-modell száját a beérkező hangnak megfelelően mozgassa. A szintetikusan előállított képet hozzáadja az eredeti műsor képtartalmához és azt együttesen jeleníti meg a felhasználó képernyőjén. A rendszer felépítésének legfőbb elemei az *1. ábrán* (az előző oldalon) láthatók.

Az IPTV stream dekódolása során megtörténik a kívánt csatorna adatainak kiválasztásán túl a csatorna átvitelénél alkalmazott kódolás visszaalakítása, így jutunk a kívánt műsorjel kép és hangtartalmához. Az SP2FA feliratú átalakító tartalmazza a beszédből a fejmodell mozgatási paramétereinek előállítását MPEG-4 kódolás felhasználásával, valamint azt az eljárást, amely a meghatározott jellemzők felhasználásával a beszélő fej mozgóképét állítja elő. A következő szükséges részegység valósítja meg a kép a képben (PIP) rendszer felhasználásával az eredeti TV-képbe a beszélő fej képének beillesztését.

A hangjelből közvetlenül, azaz nyelvi szintek felhasználása nélküli képi átalakítás elvét a *2. ábra* mutatja. A közvetlen átalakítás nehéz és korlátos, de csak ezen az úton érhető el a „lipsync" tűréshatárán belüli időeltérés a hang és a kép között. A mozgó szájról a siketek képesek a beszédet leolvasni, a részlegesen hallókat pedig a hanghoz időben pontosan kapcsolódó képi többletinformációval segíti. A rendszer alapelve és részletes ismertetése a Híradástechnika folyóirat korábbi számaiban megtalálható [4,5]. Az alábbiakban főként azokat a részleteket és megfontolásokat taglaljuk, amelyek kifejezetten az IPTV megvalósításra és a megjelenítő egységre vonatkoznak.

2. ábra
Hangból közvetlen mozgó száj képét előállító rendszer alapelve



Folyamatos beszédjelből mozgókép-folyamot hozunk létre. Ez egy olyan transzformáció, melynek lényegi részét egy neurális hálózat hajtja végre a *3. ábrán* összefoglalt rendszer szerint. A neurális hálózat komplexitását korlátok között kellett tartani, ezért elengedhetetlen volt az emberi beszédfolyamat lényegét jól megragadó, tömör és hatékony leírása a hangzó és a látható beszédnek.

Az SP2FP rendszerben a hanginformáció tömörítésére az MFCC vektorokat használtuk időkeretenként, a képinformáció tömörítésére

az MPEG-4 FP koordináták főkomponenseit használtuk [4,5]. Az első 6 főkomponens jellemző kisebb, mint 2% hibával leírta a szükséges képi koordinátákat. A tömörített beszédjelből és a tömörített vizuális jellemzőkből szinkronizált audio-vizuális adatbázist hoztunk létre. Ez az adatbázis tartalmazza azokat az információkat amit a ténylegesen működő rendszer tanítása során, illetve annak ellenőrzésekor használtunk.

Rendszerünk fő egysége a megfelelően tanított neurális hálózat. A tanítás lényegi újdonsága, hogy nem nagyszámú átlagos beszélő adataival történt, hanem kevés, de hivatásos jeltolmács adatai alapján, akik kifejezetten siketek igényeihez szabják beszédjük tempóját és látható artikulációjuk pontosságát, intenzitását. A neurális hálót ilyen szempontok szerint gyűjtött és előfeldolgozott beszédadatokkal tápláltuk a bemenetén és a siketek igényeihez igazodó képi koordinátákat vártunk a kimeneteken a jeltolmácsok videofelvételeiből származtatva. A rendszer fejlesztésében külön kezelt probléma volt a mozgókép megjelenítés modellje.

4. Megvalósítás a Direct Show keretben

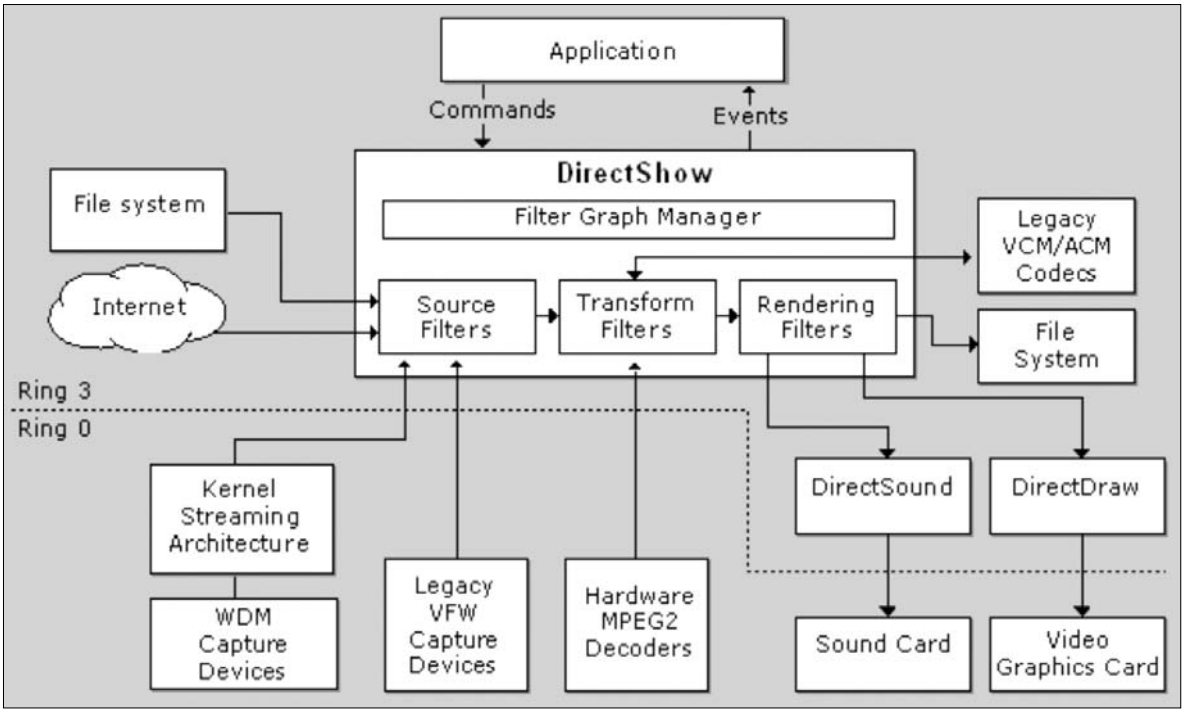
A DirectShow egy olyan környezet, amelyet a Windows operációs rendszer médiakezeléséhez fejlesztett a Microsoft. A rendszer a médiafeldolgozásban már jól ismert, hálózatba szervezhető alapvető funkciókra épül. Az alapvető feldolgozóegységeket jól megfogalmazott input-output rendszerbe helyezték, és lehetővé tették az előre lefordított egységek szabad szervezését is. Egy jól ismert példa egy ilyen alapvető funkcióra a kodek fogalma, ami egy olyan funkciót lát el, ami egy adott reprezentációban elérhető médiaállományt szabványos „kicsomagolt" reprezentációra képes alakítani, amit aztán például a videómegjelenítőre már közvetlenül lehet irányítani.

A DirectShow alapvetően tehát feldolgozóegységek halmaza, amikből hálózatokat lehet építeni. A hálózatépítés nagyrészben automatizálható és automatizált, ez történik például egy avi kiterjesztésű fájl lejátszásakor a Windows rendszerekben. A lejátszás első lépése ugyanis a hálózat generálása a fájl alapján. A fájl információt tartalmaz az azt lejátszani képes kodekről (FOURCC kódok), az audio és a video jelhez ezek akár függetlenek is lehetnek. A megjelenítést pedig a videokártyától is függő meghajtóprogram DirectShowba beépülő feldolgozóegységei vezérelhetik a szabványos, hardveres gyorsítástól mentes megjelenítőkön kívűl.

A DirectShow lehetővé teszi jó minőségű multimédiás tartalom felvételét és lejátszását. Formátumok széles skáláját (például ASF, MPEG, AVI, MP3, WAV stb.) és digitális, illetve analóg felvevőeszközöket is támogat *(4. ábra)*. A multimédiás tartalom feldolgozása sok kihívást jelent:

- Multimédiás folyamok nagy mennyiségű adatot tartalmazhatnak, melyet nagyon gyorsan kell feldolgozni, átalakítani és mozgatni.
- A hangot és a képet szinkronizálni kell, hogy egy időben induljanak és álljanak meg, valamint egyforma sebességgel történjen a lejátszásuk.
- Az adat sok különböző forrásból származhat, mint például helyi fájlokból, hálózatról, televíziós sugárzásból, vagy videokameráról.
- Az adat sokféle különböző formátumban érkezhet, mint például AVI, ASF, MPEG, DV stb.
- Egy multimédiás alkalmazás fejlesztője nem tudhatja előre, hogy milyen hardver áll rendelkezésre a célszámítógépen.

A DirectShow a fenti kihívások mindegyikére kínál megoldást. Hogy a sok különböző forrás, formátum, illetve hardver különbözőségét kezelni lehessen, a Direct Show egy moduláris architektúrát használ, melynek alapeleme a szűrő (filter).



4. ábra DirectShow rendszer felépítése

Egy szűrő bemenetekkel és/vagy kimenetekkel rendelkező komponens, mely egy adott részfeladatot lát el. Alapvetően három típusú szűrő van:

– Forrás-szűrő (Source filter)
– Transzformáló szűrő (Transform filter)
– Megjelenítő szűrő (Rendering filter)

A DirectShow rendszerben minden megoldás valamilyen szűrő felhasználását jelenti. Egy-egy szűrő valósítja meg a bemenetre érkező jelek fogadását (forrásszűrő) és ennek az egységnek a feladata az is, hogy a jelfolyamot a további transzformáló szűrők által feldolgozható formátumra alakítsa. A transzformáló szűrő valósítja meg a megfelelő adatformátumok közötti átalakítást. Ilyen transzformáló szűrő lehet például egy hardware megoldású MPEG dekóder alkalmazása, természetesen csak abban az esetben, ha a megfelelő eszköz rendelkezésre áll. Hasonló feladatot lát el a hang kezelésével kapcsolatos különböző kodek-ek közötti átalakítás is. A megjelenítő szűrő képes az előzetesen más egységek által átalakított jelfolyamok megjelenítésére, hang és video kártyák kezelésére, de ez a részegység a felelős más eszközök – mint például a fájlrendszer, internet csatlakozás – kezelésére is.

Az architektúra alapkoncepciója a gráf-modell, melynek csomópontjai a szűrők. A DirectShow biztosít egy alapkészletet, de a rendszer igazi erőssége abban rejlik, hogy tetszőleges szűrőkkel bővíthető. Az egyes multimédiás megjelenítő eszköz gyártók az általuk készített eszközökhöz biztosítják a megfelelő DirectShow környezetbe illeszkedő forrás szűrőt.

Szemléltetésképpen egy AVI fájl lejátszásának folyamata a Windows médialejátszójában az *5. ábrán* látható, amely folyamat az alábbi lépéseket tartalmazza:

– Nyers adat olvasása a fájlból bájt-sorozatként (Fájl forrás szűrő)
– Az AVI formátum feldolgozása és szétválasztás képkockákra, ill. hangmintákra (AVI Splitter szűrő)
– A képkockák dekódolása (különböző dekódoló szűrők lehetségesek, a tömörítéstől függően)
– A képkockák megjelenítése (Megjelenítő szűrő)
– A hangminták lejátszása a hangkártyán keresztül (Hanglejátszó szűrő)

Hasonló a helyzet akkor is, ha valamilyen eszköz megvalósítja az IP hálózatból érkező TV adás bitfolyamának vételét. Ez esetben a jelforrásból érkező digitális jel MPEG kódolású, tehát a megjelenítés előtt MPEG dekódoló alkalmazása szükséges. A számítógépen történő megjelenítés teljesen hasonló az 5. ábra szerinti megoldáshoz. Egy megjelenítő szűrő segítségével érhető el, hogy a rendelkezésre álló képinformáció a monitoron látható legyen.

A kifejtés kezdetén ismertetett és az 1. ábrán vázolt rendszer leírása a DirectShow rendszer szemléletében válik érthetővé és világossá azáltal, hogy ebben a rendszerben valósítható meg legcélszerűbben a jelenleg rendelkezésre álló műszaki feltételek között. Az SP2FA egység egy DirectShow elemként illeszthető a rendszerbe. A kép a képben (PIP) eljárással az eredeti képanyagba illesztett beszélő fej egy szokásos Direct Show alkalmazási elem.

5. Értékelés

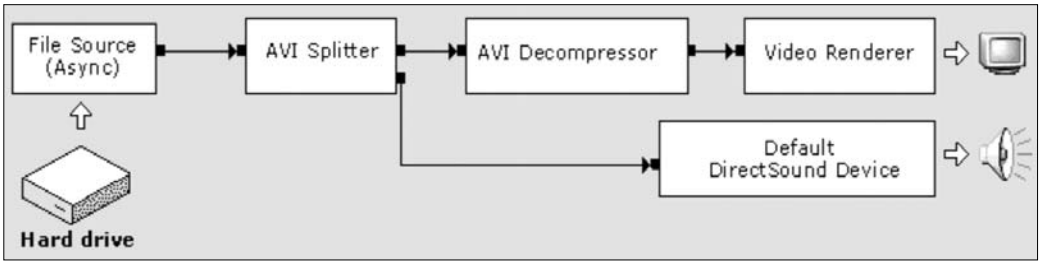
A kísérleti rendszer megvalósult. Szokványos PC erőforrásokon valós időben működik. Szélesebb körű tesztelésére és hosszabb idejű próbájára szükség lenne annak érdekében, hogy a célfelhasználók megtanulják, megszokják.

Köszönetnyilvánítás

A szerzők ezúton is köszönik a Magyar Telekom támogatását a kísérleti rendszer létrehozására.

Irodalom

[1] http://www.felston.com/reviews.htm
[2] RNIB, RNID, EFHOH, EUD, FEPEDA and EBU, Submission in Response to the EC Public Consultation on the Review of Television Without Frontiers Directive, European Voice conference on Television Without Frontiers (Brussels, 21/03/02).
[3] Helga Stevens: Equal rights for deaf people – From being a stranger in one's own country to full citizenship through sign languages, ICED 2005, Maastricht, 17-20 July 2005.
[4] Takács Gy., Tihanyi A., Bárdi T., Feldhoffer G., Srancsik B.: MPEG-4 modell alkalmazása szájmozgás megjelenítésére. Híradástechnika, LXI.évf. 2006/8, pp.22–28.
[5] Takács Gy., Tihanyi A., Bárdi T., Feldhoffer G., Srancsik B.: Beszédjel átalakítása mozgó száj képévé siketek kommunikációjának segítésére. Híradástechnika, LXI.évf. 2006/3, pp.31–38.
[6] MSDN: http://msdn2.microsoft.com/en-us/library/ms783323.aspx



5. ábra Példa a DirectShow működésére, egy avi kiterjesztésű fájl lejátszása kapcsán

Számítógépes teremakusztikai szimuláció hangtér optimalizáláshoz

WERSÉNYI GYÖRGY

Széchenyi István Egyetem, Távközlési Tanszék
wersenyi@sze.hu

Lektorált

Kulcsszavak: teremakusztika, CARA, CAD, utózengés, hangtéroptimalizáció

A teremakusztikai tervezés, utózengési idő számítása és a hangtér optimalizálása régóta a műszaki akusztika egyik nehéz feladata. Tekintettel arra, hogy bizonyos közelítésekkel e számítások egyszerűen és gyorsan gépesíthetők, mára többféle számítógépes tervező program segíthet bennünket. A cikk bemutatja, hogy a CARA (Computer Aided Room Acoustics) program segítségével miként lehet tetszőleges termeket, azok berendezési tárgyait CAD módszerrel megépíteni és az ismert formulák segítségével számításokat végezni az utózengési időre, visszaverődésekre, hangnyomástérképekre. Segítségével optimalizálhatjuk a terem kiépítését, a hangsugárzók és a lehallgatási pozíciók elhelyezését. A vizsgálat aktualitása a Széchenyi István Egyetem új, D1-es jelű felújított nagyelőadójának átadása, a hangosítás vizsgálata. Egy másik példán röviden egy lakószobai házimozi optimalizálását láthatjuk.

1. Bevezetés

Egy terem akusztikai kialakítása, hangosítása vagy éppen hangszigetelése és a „mi szól jól?” kérdések megválaszolása nehéz feladat. Léteznek objektíven vizsgálható, mérhető paraméterek, mint például a hangnyomás(szint) és annak eloszlása, az utózengési idő, a terem módusai és az esetleges állóhullámok kialakulása. Ezek gyakran azonban másodlagosak a szubjektív élvezet szempontjából és csak becslést, közelítő értékeket adnak, illetve iránymutatást tudnak nyújtani a tervezéshez, átalakításhoz [1,2]. A paraméterek kiszámításához azonban segítségünkre lehet a geometriai akusztika, amely tulajdonképpen a geometria optika számításait használja fel. Hasonlóan, végeselem-, peremelem módszerek, nagy számításigényű hangtérleírások egyre pontosabban szimulálják számunkra a „hallanivalót”. Mára a számításigény kielégíthető a számítógépekkel, egyszerű, de nagy mennyiségű számolások rábízhatóak a szoftverekre. Nem várjuk el, hogy pontosan megmondják nekünk, mit, hova, és hogyan kell elhelyezni, de útmutatást adhatnak a helyes kialakításhoz. E szimulációk sikeressége pedig jórészt a felhasznált modellek pontosságán múlik.

A piacon többféle akusztikai tervezőprogram is létezik. Legismertebb közülük a CATT programcsomag [3]. A kevésbé ismertek közé tartozik az itt is bemutatásra kerülő német fejlesztésű CARA (Computer Aided Room Acoustics) [4]. Az interneten elérhető, megrendelhető, ára is gazdaságos. Lehetőséget biztosít a termek létrehozására és berendezési tárgyainak megtervezésére ismerős CAD felületen. Hasonló elveken tetszőleges hangsugárzókat is megépíthetünk, ha nem elégséges a hozzá kapott adatbázis. A kettő együttes ismeretében a program először analizálja nekünk a termet és annak utózengési idejét az ismert formulás segítségével. Majd a hangszórók és a hallgatók elhelyezésével optimalizálási stratégiákat dolgoz ki a jobb hangzás (egyenletesebb eloszlás) érdekében. Utóbbiak gyakran nem egyértelműek, néha több megoldást is kapunk, melyeket aztán saját szubjektív ízlésünk szerint szelektálhatunk.

A cikkben bemutatásra kerül a program néhány alapfunkciója a győri egyetem felújított előadója és egy otthoni nappali szoba példáján keresztül.

2. A geometriai akusztika számításai

A geometriai hullámterjedés az optikából ismeretes. Az a tény, hogy használhatjuk-e az ismert optikai törvényeket (Snellius-Descartes, töréstörvények, elnyelés és visszaverődés, tükörforrások elve stb.), attól függ, mekkora a hullámhossz. Megfelelő frekvenciatartományban jó közelítésekkel számolhatunk, ha a fenti törvényeket alkalmazzuk. Ha a hullámhossz jóval kisebb a fal felületénél, a beesési- és visszaverődési szögekre, a hangutak kiszámításához alkalmazhatók a fénytörési törvények (például beesési szög = visszavert szög). Röviden bemutatjuk, mely paraméterek azok, amelyeket papíron vagy számítógép segítségével meghatározhatunk.

2.1. Utózengési idő

Az utózengési idő definíció szerint az az időtartam, amely alatt a terembe betáplált és állandó szinten tartott hangteljesítmény a hang megszűnése után 60 dB-el esik [5]. Kétféle elterjedt mérési módszere van. A nehézkesebb, amikor a definíció szerint mérünk és egy hangforrás (jellemzően fehér zajszerű, úgynevezett referencia-hangforrás, nagy, állandó teljesítménnyel) kikapcsolása után vizsgáljuk az eredményt. A másik gyakoribb módszer az impulzusválasz vizsgálata, amikor a termet nagy teljesítményű impulzussal gerjesztjük (riasztópisztoly, lufi durrantása). A méréseinket általában valamilyen műszer segíti, a modern kézi zajanalizátorok

nem csupán zajszintet mérnek, hanem többek között az utózengési időt is meghatározzák.

Az utózengési idő frekvenciafüggő. Teli koncerttermek esetén az 1,8-2,5 mp is elfogadható középfrekvenciákon [6]. Jellemzően templomokban 5-8, koncerttermekben 1,5-2,2, színházakban 1,0-1,5, stúdiókban 0,2-0,6, süketszobákban pedig kisebb mint 0,05 másodperc az utózengési vagy lecsengési idő. Az utózengési időből jól lehet következtetni a terem méretére, „zengésére", beszédakusztikai tulajdonságaira.

Az utózengési időt ismert közelítő formulákból számítással is megbecsülhetjük [5,7]. Nem túl kicsi utózengési idő (τ) esetén a *Sabine*-formula az alábbi:

$$\tau = \frac{0{,}161V}{A}, \qquad (1)$$

ahol az utózengési időt sec-ban kapjuk meg, ha *V*-t köbméterben, *A*-t négyzetméterben helyettesítjük, a 0,161-es konstansnak pedig [s/m] a dimenziója. Az *A* itt nem a felületet jelenti közvetlenül, hanem az abszorpciót:

$$A = \sum_i \alpha_i S_i . \qquad (2)$$

Ebben a képletben az *S* változó már ténylegesen egy adott felületet jelent négyzetméterben, a hozzátartozó elnyelési tényezővel (alfa). Az elnyelési tényező általában adott, táblázatból kikereshető [8,9]. Gyakorlatilag arról van szó, hogy a különböző anyagú felületeket súlyozzuk. Így ha van egy betonszoba adott felülettel és alfával, akkor az azon nyitott faajtó felületét is a fa alfájával kell súlyozni. Az alfa mérhető is, és számolható is, ráadásul frekvenciafüggő.

Ez a képlet nagy utózengési időknél használatos, és egyenletes terjedést feltételez minden irányban (izotróp), a terem módusait elhanyagolva. Nagyobb *A* és egyre kisebb τ esetén az eredmény egyre pontatlanabb lesz. Kisebb τ esetén a másik használatos képlet az *Eyring*-formula:

$$\tau = \frac{0{,}161V}{S\ln(1-\bar{\alpha})}, \qquad (3)$$

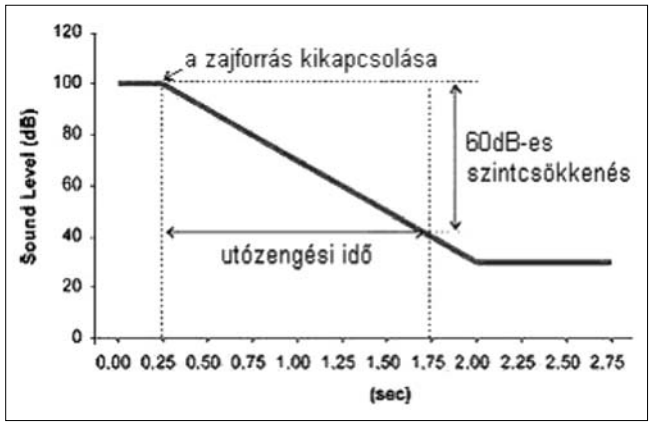
ahol egy átlagos alfával dolgozunk:

$$\bar{\alpha} = \frac{\alpha_1 S_1 + \alpha_2 S_2 + \ldots + \alpha_i S_i}{S_1 + S_2 + .. + S_i}, \qquad (4)$$

és $$S = S_1 + S_2 + .. + S_i . \qquad (5)$$

Akkor a legpontosabb ez a formula, ha az α-k körülbelül egyenlők (hátrány), ugyanakkor matematikailag korrektebb, mert süketszobára, ahol alfa értéke egy, τ-ra zérus jön ki.

Ezekhez a számításokhoz csak a terem geometriai méreteire és anyagára van szükség. Az anyagok felületének és elnyelési tényezőjének ismeretében (utóbbiakat táblázatból kiolvashatjuk), viszonylag egyszerű módon számolhatunk. Ebben a számítógép sokat segíthet.



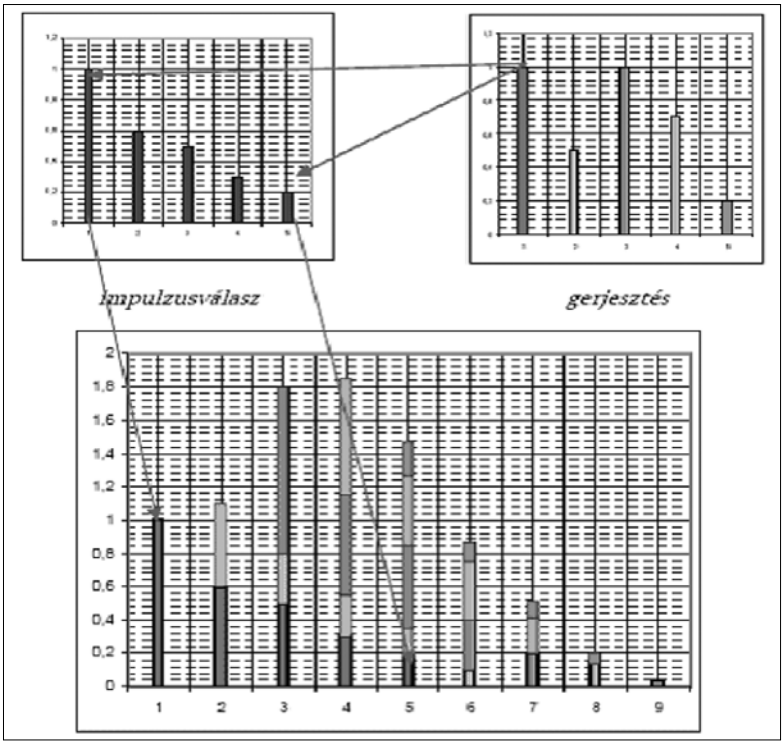
1. ábra Az utózengési idő szemléltetése

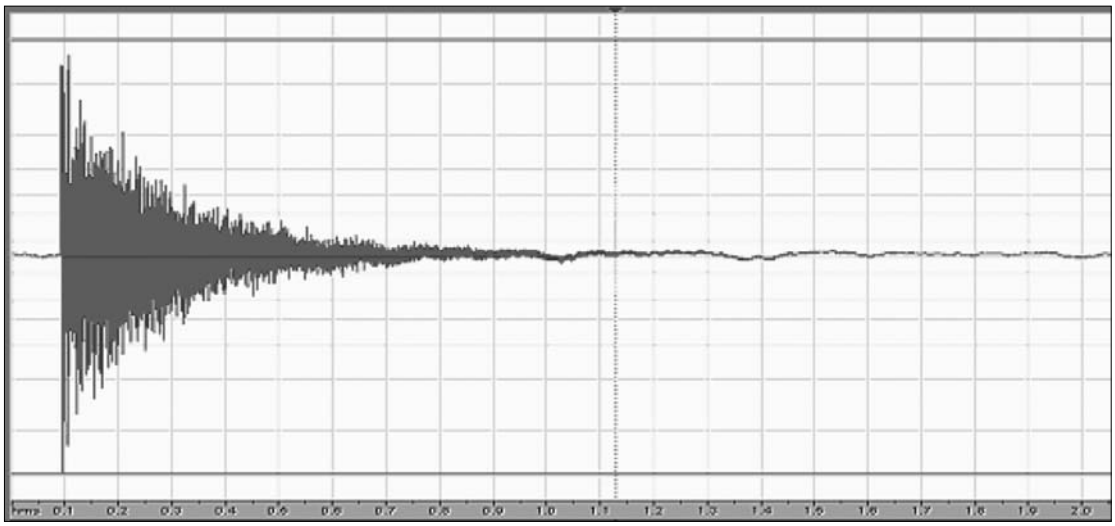
2.2. Echogram és a terem impulzusválasza

A terem válaszát egyszerűen meghatározhatjuk az impulzusválaszával. Az impulzusválasz rendszerleíró függvény és mint ilyen, az időtartományban teljes egészében hordozza az adott rendszer átviteli tulajdonságait. Egyszerűen, gyorsan mérhető, hátránya, hogy általában kis energiájú (rossz jel-zaj viszonyú), különösen mélyfrekvencián nehéz egy termet gerjeszteni. Ismert, hogy az impulzusválasz Fourier-transzformáltja a komplex átviteli függvény, melyet könnyedén meghatározhatunk.

Korábban a számításigény miatt az időtartománybeli manipuláció nem volt lehetséges. Ezért a frekvenciatartománybeli szorzással és FFT, IFFT algoritmusokkal gyorsítottuk a folyamatokat. Manapság az időtartománybeli konvolúciónak nincs különösebb akadálya, így

2. ábra
A konvolúció hatása egy impulzusválasz és gerjesztés esetén



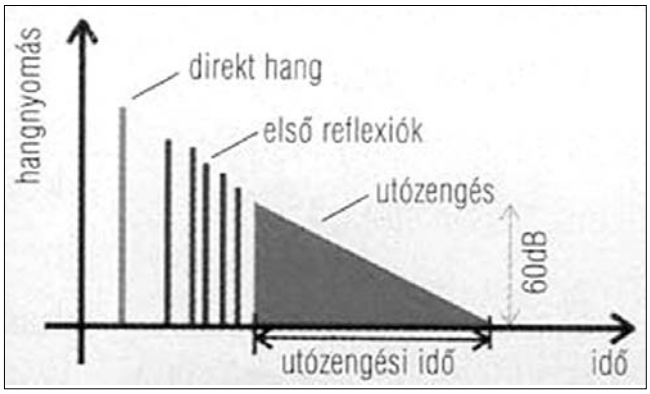


3. ábra
A későbbiekben bemutatott D1-es terem impulzusválasza

az impulzusválasz az egyik legfontosabb leíró függvényünk lett. Nincs más dolgunk, mint egy wave-fájlban rögzített impulzusválaszt a megfelelő szoftver segítségével egy mono stúdiófelvétellel konvolválni és végeredményünk olyan lesz, mintha az eredeti mono stúdiófelvétel az adott teremben szólna. A konvolúciós integrál alakja az alábbi:

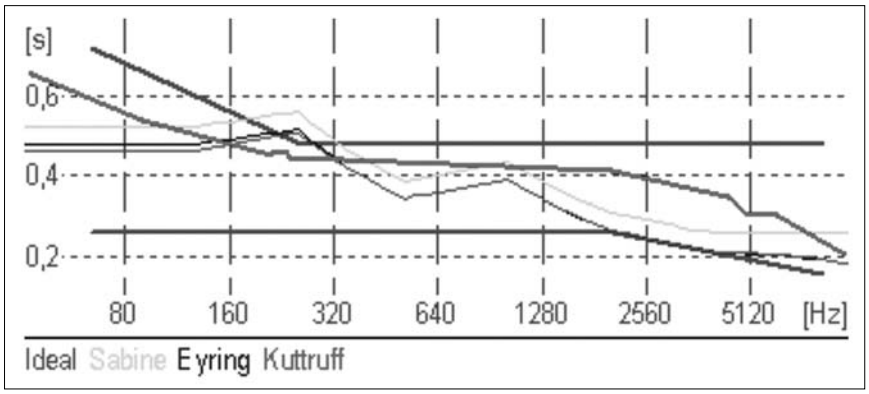
$$f * g = \int_0^t f(t-\tau)g\{\tau\}d(\tau)\,. \qquad (6)$$

Ehhez segítséget nyújtanak olyan speciális programok, mint az Altiverb [10], vagy az Adobe Audition, de MATLAB alatt is egyszerűen elvégezhető a művelet. Az így megmért teremátviteli függvényt (Room Impulse Response – RIR) tehát sok mindenre felhasználhatjuk. A CARA program ezt is megteszi számunkra.

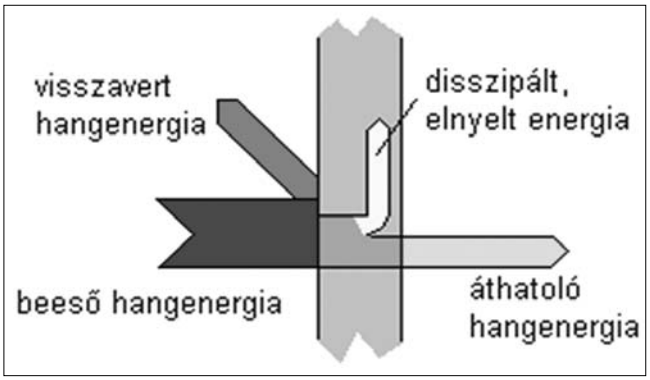


4. ábra Stilizált echogram

Az echogram tulajdonképpen egy terem impulzusának és a visszaverődéseknek a lekövetése az időtartományban.



6. ábra
Toleranciasémába illő, illetve nem illő utózengési idő diagramok

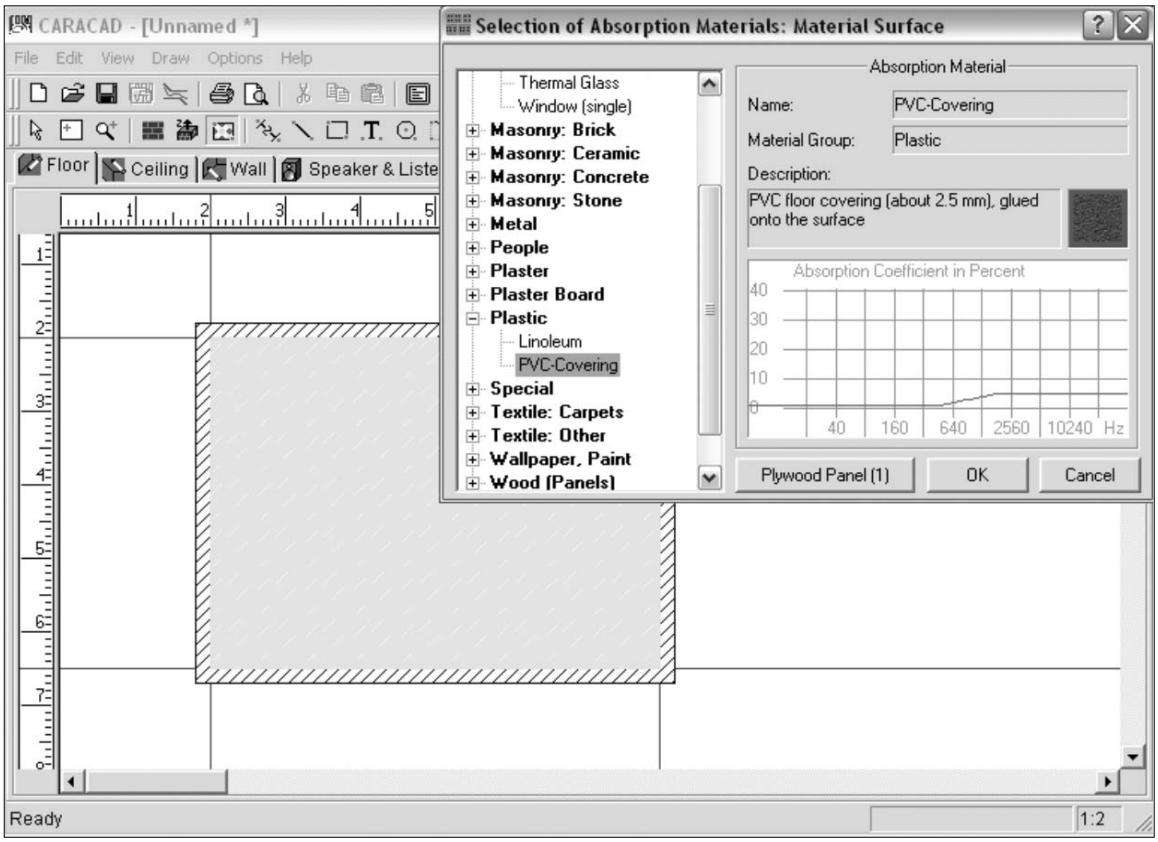


5. ábra
A beeső hangenergia útjai

2.3. Refrakció, diffrakció, reflexió

A hang terjedése során többféle hatásnak van kitéve. Ezek súlya attól függ, mekkora az akadály vagy lyuk mérete a hullámhosszhoz képest. A hang beeséskor jórészt visszaverődik, reflektálódik. Egy másik része elnyelődik, amely egyrészt kis mértékben hő formájában felszabadul, egy másik része pedig megmozgatva az akadályt átjut és ismételten lesugárzódik.

A hanghullámok emellett elhajlási jelenségeket is mutatnak, valamint szóródnak és árnyékba hatolnak. Egy adott hangterjedés esetén ezek szerepe és fontossága a frekvencia függvényében változik. Ne feledjük, hogy a hanghullámok a néhány centimétertől akár 17 méteres hullámhosszig is terjedhetnek. A mélyfrekvenciák lesugárzása, csillapítása, iránytól függő érzékelése lényegesen nehezebb probléma, mint a magasabb frekvenciáké.



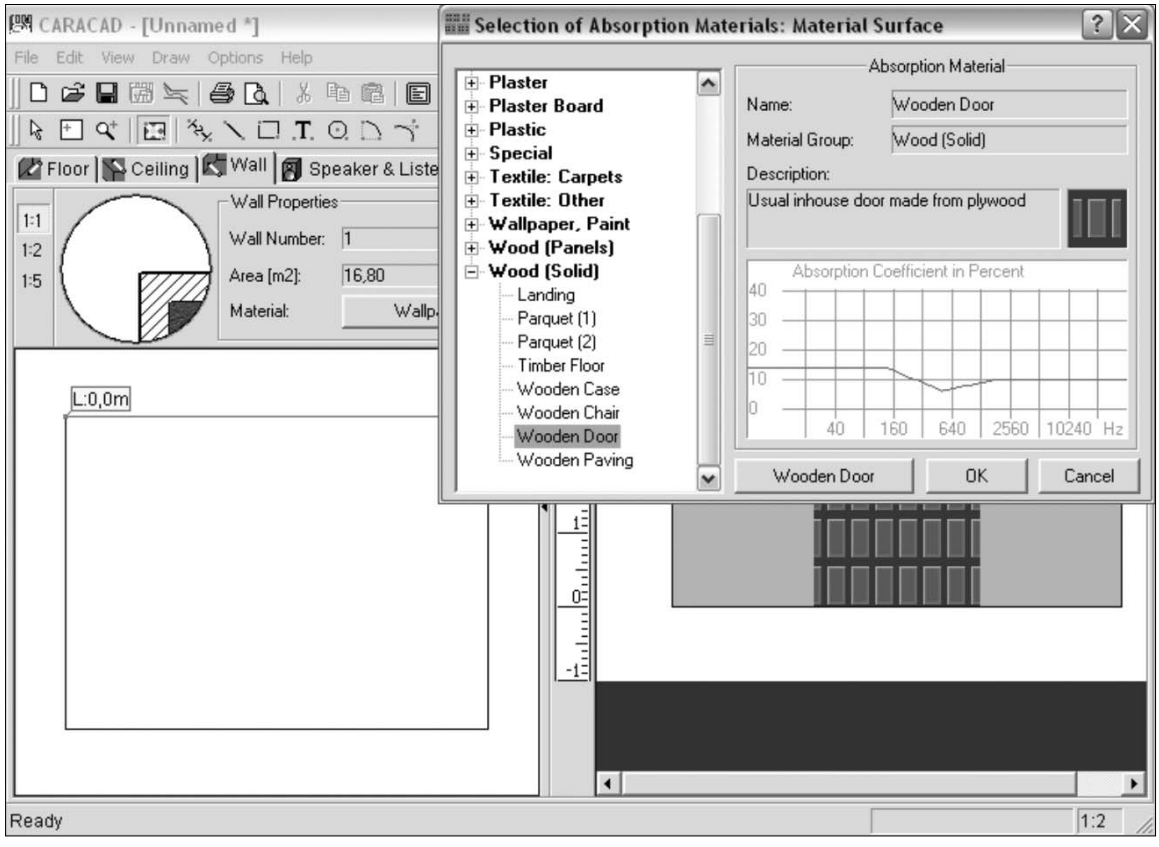
7. ábra
A padló virtuális beborítása PVC-vel és annak elnyelési görbéje a frekvencia függvényében

2.4. Egyéb paraméterek

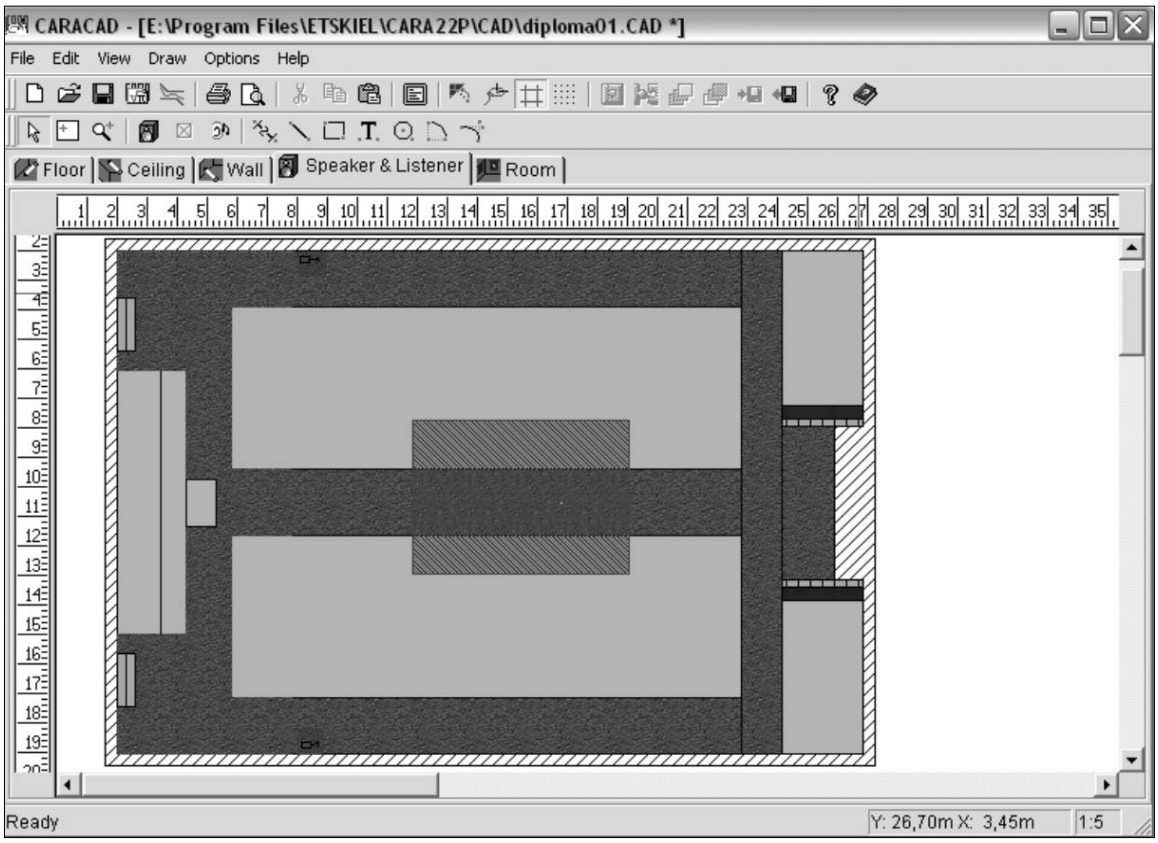
Néhány a fentiekkel rokon, azokból származtatható paraméterek is segíthetik munkánkat.

Az EDT (Early Decay Time) az első 10 dB-es eséshez tartozó idő. Ajánlatos, hogy ennek átlaga haladja meg a utózengési idő átlagát nagyszámú emberrel telített termekben.

Olyan paraméterek, mint a tisztaság (Clarity, C_{80}), vagy a 'Lateral Efficiency' (LE) a szoftverek segítségével meghatározhatók.



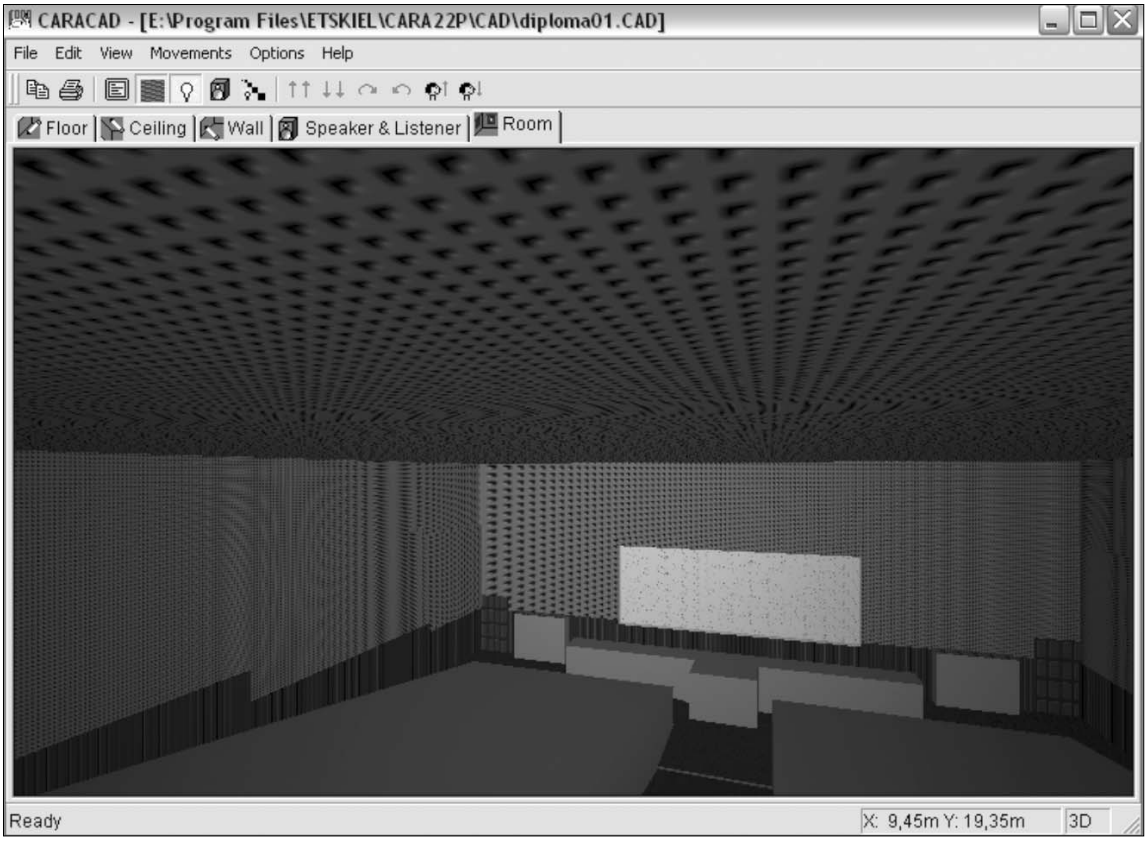
8. ábra
Faajtók elhelyezése és elnyelési tényezője



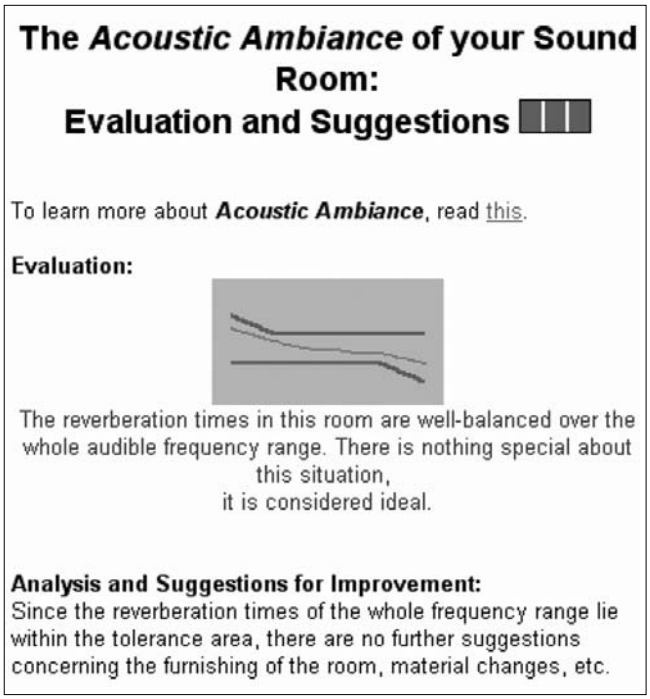
9. ábra
A kész terem 2D felülnézeti képe. Bal oldalon a katedra és a két faajtó. Középen a két nagy padsor és egy lehetséges lehallgatási zóna. Két hangszóró az oldalfalakon, a jobb oldalon a terem kijárata és még két padsor.

3. A számítógépes szoftver

A CARA lehetővé teszi a fenti paraméterek szimuláción történő becslését, meghatározását. Ennek első lépéseként a terem kialakítását, felépítését (room design) kell létrehoznunk. Ezután van lehetőségünk az akusztikai számításokra (room acoustics calculations); az eredmények 2D és 3D ábrázolására (presentation of results); valamint a hangsugárzó tervező modul kihasználásra (loudspeakers).



10. ábra
A kész terem 3D nézeti képe textúrázva. A hátsó sorokból látható a tanári asztal, a tábla és az ajtók. A padsorok egybefüggő fafelületűek.



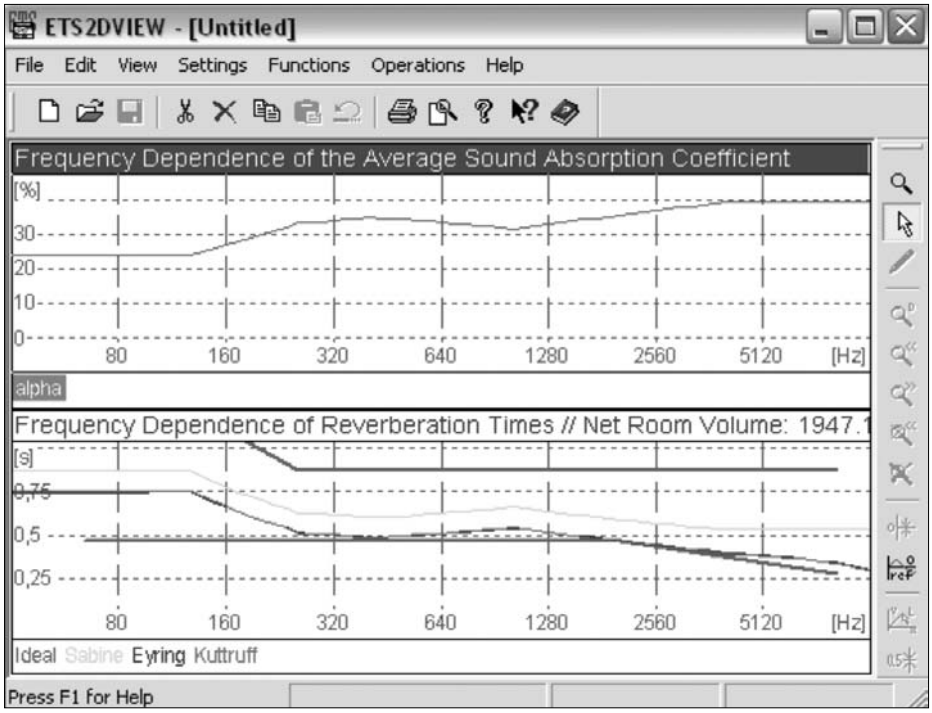
11. ábra
Az akusztikai vizsgálat eredménye

A tervezés első lépéseként a CARACAD-ben kell a termet létrehozni. Ebbe beletartozik a geometriai méret és alak, a falak, padlók, ajtók, ablakok burkolása, valamint a berendezési tárgyak elhelyezése. A tervezés centiméteres pontosságú. Itt lehet megadni a kívánt térhangzást is a sztereótól a 8.1-es rendszerekig, a hangsugárzók fajtáját, méretét, elhelyezkedését.

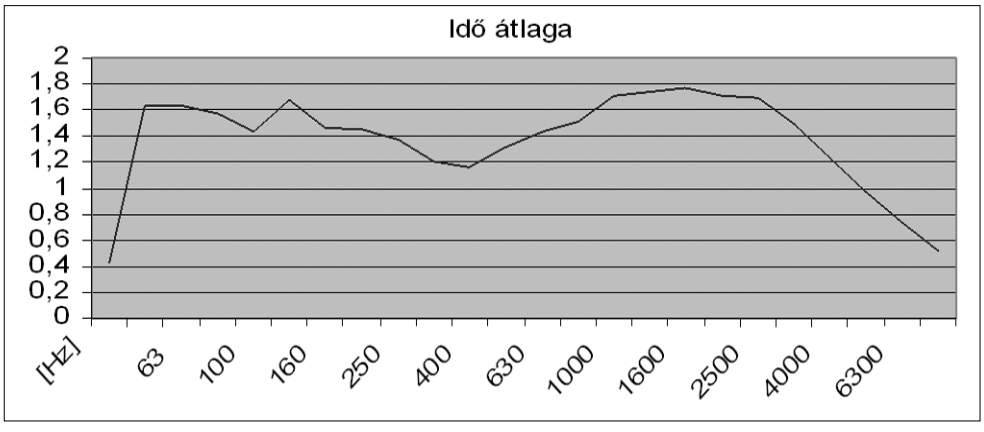
A nagyobb gyártók termékei megtalálhatók egy adatbázisban (mely az internetről frissíthető is), a hiányzókat pedig magunk létrehozhatjuk és elmenthetjük.

A tervezés után akusztikai ellenőrzést kell végeznünk (acoustic ambiance), mely az idő- és frekvenciaviszonyokat bemutatva kiszámítja és megjeleníti a toleranciasémát és a terem tulajdonságait (utózengési idők, reflexiók, elnyelődések).

Az akusztikai kalkuláció során a terem több ezer rácspontra lesz felosztva. Meghatározásra kerül az optimális hangsugárzó-elhelyezés a lehallgatási pont(ok) függvényében, néha több javaslat is előkerül. Az eredményeket 2D vagy 3D ábrázolásban is megtekinthetjük, bejárhatjuk.



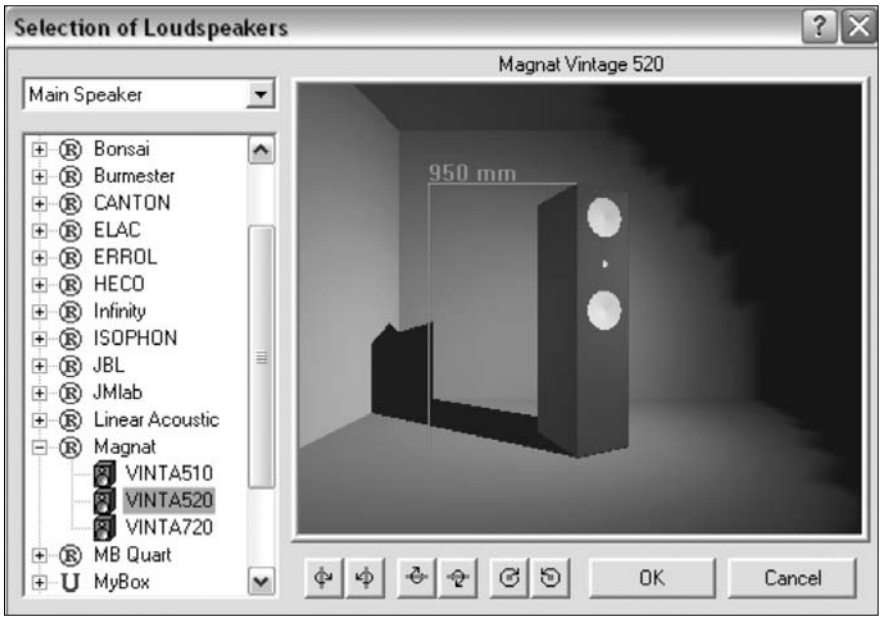
12. ábra
Felül az átlagos elnyelési tényező, alul az utózengési idő frekvenciafüggése a modell alapján számolva



13. ábra
Az utózengési idő mérése, Brüel-Kjaer 2260-as analizátorral, „lufidurrantásos” módszerrel

3.1. A D1-es előadó

A D1-es előadó a győri egyetemen a legnagyobb, falai vasbeton szerkezetűek. A terem teljes falterülete légrésekkel ellátott gipszkarton lemezzel van borítva, ami hang- és hőszigetelő. A mennyezet is a lámpák között ilyen szigetelő lapokkal van kitöltve. A terem lejtős, alján található két darab fa ajtó a tábla két



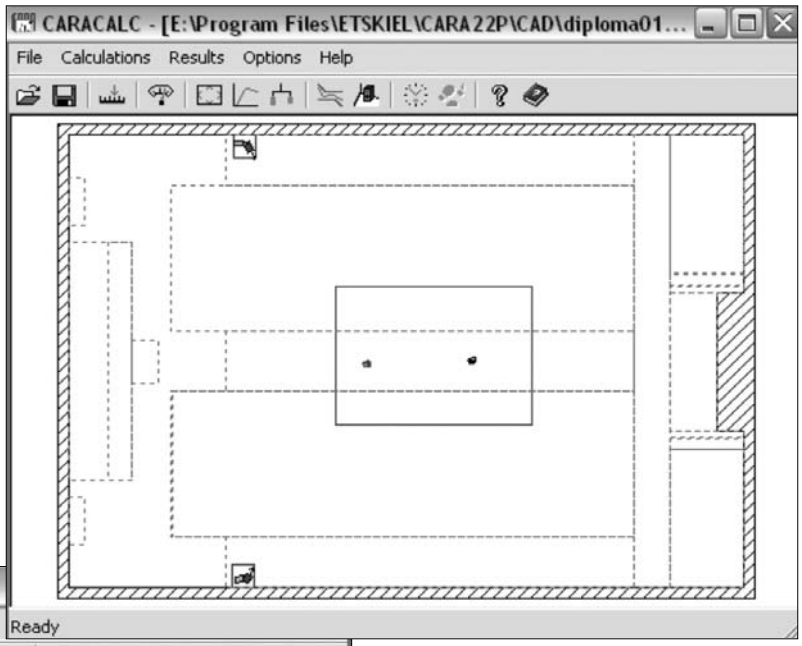
14. ábra
Magnat-hangsugárzó a listából kiválasztva

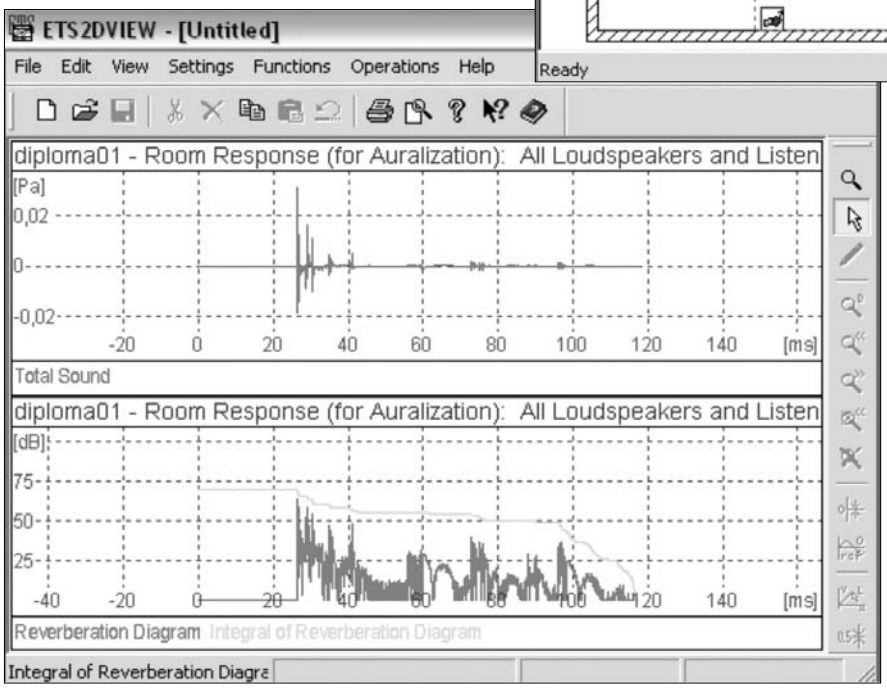
15. ábra
Az új hangsugárzó-elhelyezés és a hallgató helye a zónán belül. A hangszórókhoz közeli hallgatási pont a legjobb.

oldalán. Fent a bejárati dupla ajtó is fa. A padló borítása 8 mm-es PVC szőnyeg.

A CARACAD-ben téglalap alapsémából kiindulva építhetjük fel a termet. Megadhatjuk a padló anyagát (plastic, PVC covering), a falak kialakítását, a plafon burkolatát. A teremben ablak nem található, csak ajtókat és egy táblát kell elhelyezni (wooden door, video screen).

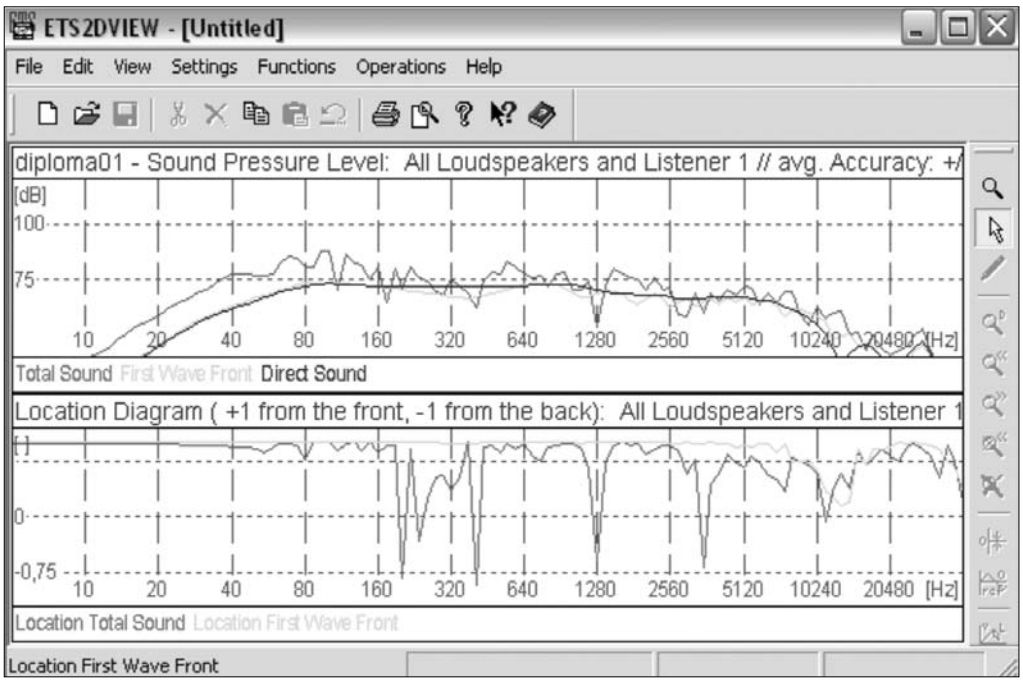
A tanári asztal, padsorok és radiátorok, mint 3D objektumok kerülnek be a modellbe. Ha ezek nem találhatók meg a sémában, kénytelenek vagyunk megépíteni őket a méretük alapján. A padok, mivel nem négyszögletesek, egyben a székekkel kerültek megtervezésre (fából). Ilyen jellegű termeknél a fapadok és -székek helyett „emberrel” is boríthatjuk a felületet, magyarán vizsgálhatjuk az üres és az emberekkel teli környezetet is.





16. ábra
A terem válaszfüggvényei.
Fent az impulzusválasz, alul az echogram.

Az akusztikai vizsgálat során az így kialakított termet vizsgálja a program. Ahogy az ábrán is látszik, elégedett az eredménnyel. Ezt a következtetést az utózengési idő és az átlagos elnyelési tényező ismeretében vonta le. Természetesen, ha a modellünk nem jó, vagy nem elég pontos, az eredmények hibásak is



17. ábra
A hangnyomásszint frekvenciafüggése a lehallgatási pozícióban az összes hangsugárzó működése esetén, valamint az úgynevezett hely-diagram, mely a teljes hang, illetve az első hullámfront elhelyezkedését mutatja (+1 szemből, -1 hátulról).

lehetnek. Ilyen esetben célszerű méréssel ellenőrizni azokat. Ha az utózengési idő gyanúsan alacsony értékű, méréssel ellenőrizve – különösen mélyfrekvencián – nagy eltérést tapasztalunk. Ennek oka az elégtelenül felépített modell lehet.

A továbbiakban a CARACALC segítségével optimalizáljuk a termet, elsősorban a hangsugárzók elhelyezése és irányítottsága a kérdés. A program futása során több ezer lehetséges pozíciót próbál ki és általában 12-16 optimális javaslattal áll elő.

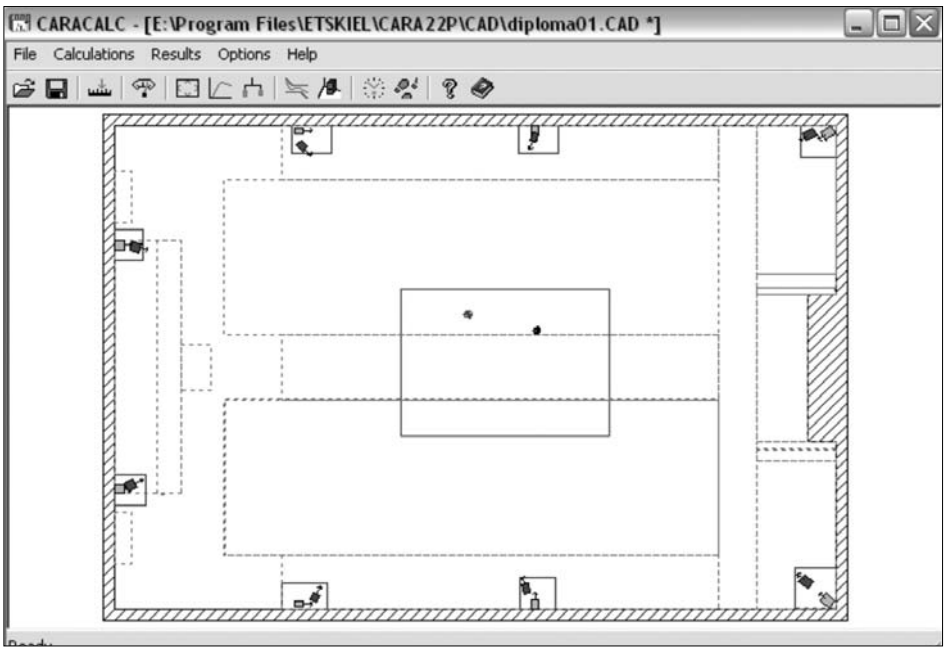
A hangsugárzók kiválasztása történhet a meglévő listából, vagy magunk is megszerkeszthetjük őket.

A program lehetőséget nyújt auralizációra is [11]. Lehet az összes hangsugárzóval egyszerre vagy egyesével is a szimulációt létrehozni. Wave-fájlba elmenthető a terem szimulált impulzusválasza, mellyel tetszőlegesen betöltött hangmintát, zenét színezhetünk. Egy utasítással összehasonlíthatjuk a hangzást optimalizálás előtt és után.

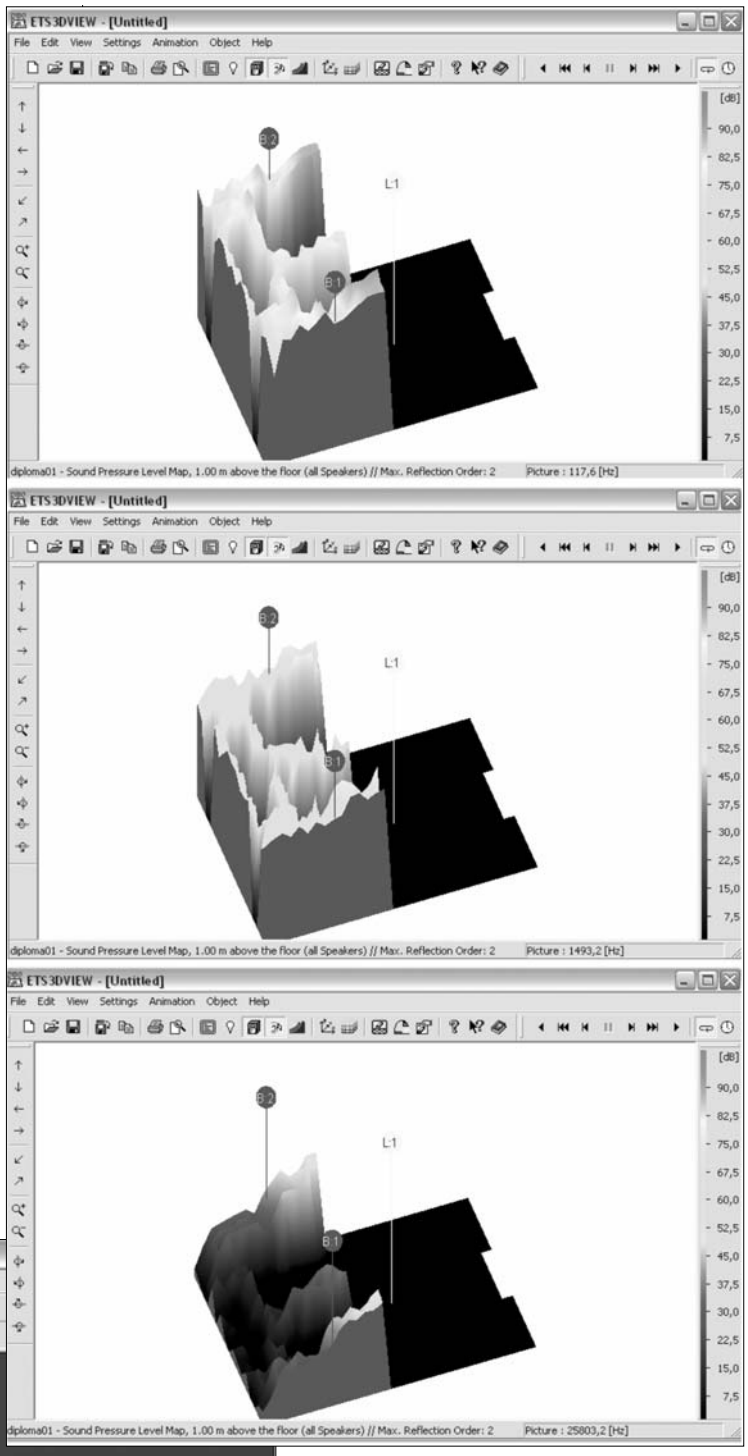
A program ezen túl színes, mozgó 3D ábrákkal szemlélteti a frekvenciában vagy az időben történő hangnyomásszintbeli ingadozásokat. Ezzel eloszlástérképeket és káros állóhullámokat kereshetünk. A mozgó ábráktól itt el kell tekintenünk, de néhány jellemző ábrát bemutatunk.

Hasonlóan, rögzített frekvencia mellett az időbeni hullámterjedést is felrajzoltathatjuk.

Végezetül két ábra egy optimalizált, nyolchangszórós elrendezésre 2D-ben és 3D-ben, ugyanazon terem számára.



19. ábra
Nyolc hangszóróra optimalizált terem 2D és 3D ábrázolása (jobbra, a túlodalon)



18. ábra
117 Hz, 1500 Hz és 25000 Hz-es kialakult hangnyomásszint-eloszlás a teremben. Sötéttel a hangsugárzók, világossal a hallgatók vannak jelölve.

3.2. Otthoni lakószoba

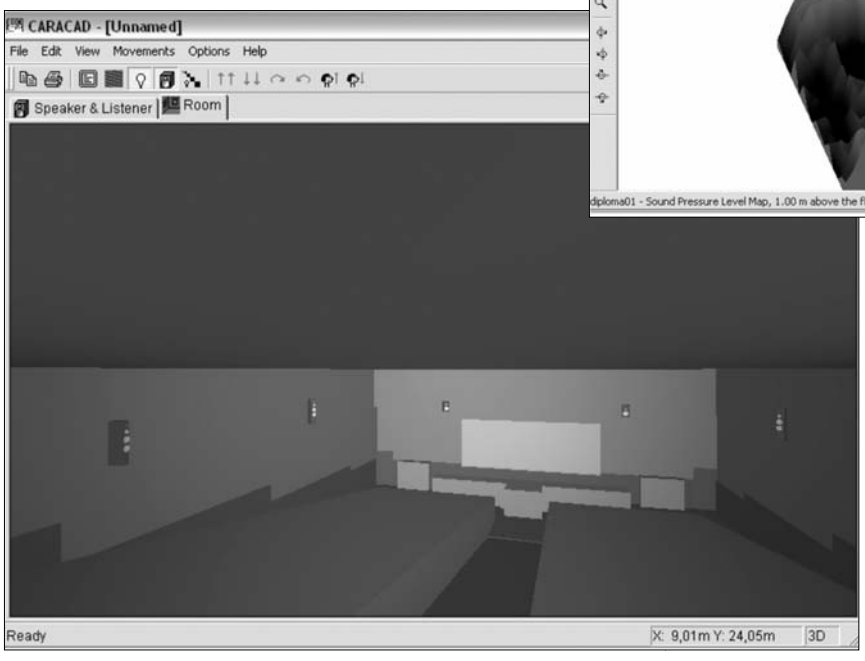
Utolsó rövid példánkban egy otthoni nappali házimozi-hangrendszer vizsgálatát láthatjuk. A *20. ábra* bal és jobb oldalán az 5.1 elrendezés optimalizált javaslata látható egy hallgató, illetve két hallgató esetén. Alatta a 3D megjelenítés azok elhelyezésére *(21. ábra)*, illetve az 50 Hz-es mélyhang eloszlása a mélynyomó (subwoofer) környezetében *(22. ábra)*.

4. Összefoglalás

A teremakusztikai tervezés számítógépes lehetőségei közül bemutatásra került egy költséghatékony, jól használható szoftveres megoldás. A program képes az alapvető akusztikai paraméterek becslésére, számítására, látványos 2D és 3D megjelenítésére.

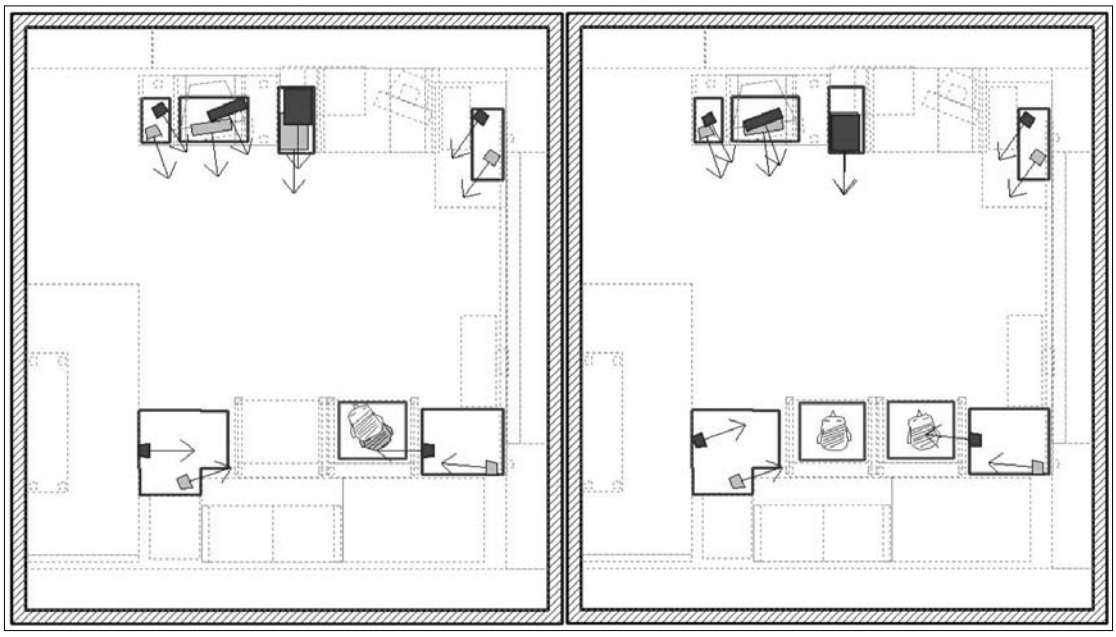
A győri egyetem előadójának szimulációja rámutatott a modell pontosságának és a szimulációk méréssel történő ellenőrzésének fontosságára. Ugyanakkor látható, hogy a mai számítási kapacitás lehetővé teszi az akusztikai tervezés és hangtérkialakítás alapvető lépéseinek felgyorsítását és vizualizálását.

Segítségével képet kaphatunk a terem hangképéről, követhetjük a javaslatokat és az optimalizálási stratégiákhoz ötleteket meríthetünk. A végső szót úgyis a hallgató, a tesztalany, a nézősereg hozza meg szubjektív benyomásai alapján.



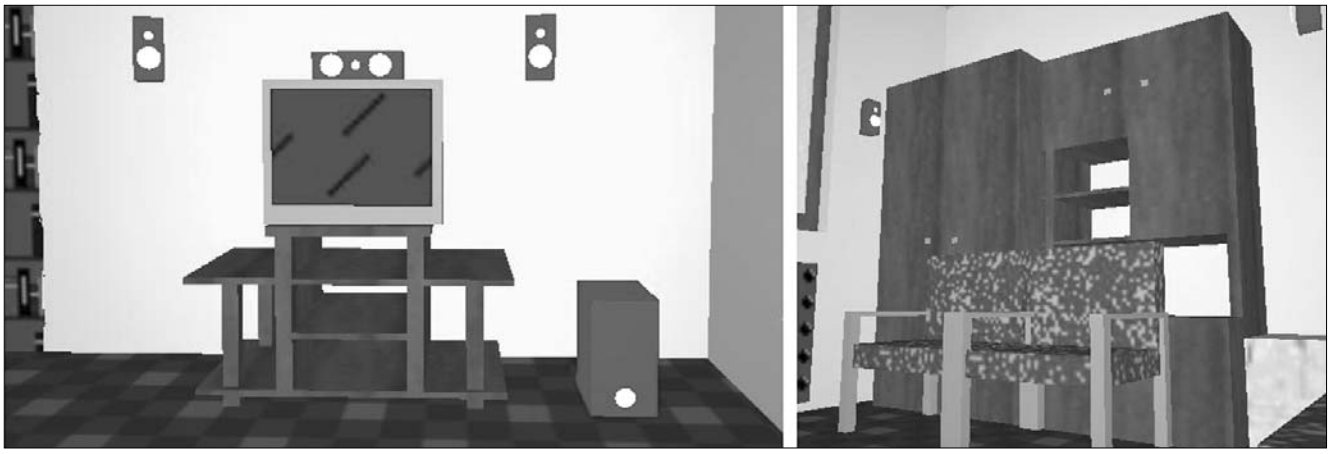
A szerzőről

Wersényi György a Budapesti Műszaki Egyetem villamosmérnöki szakán szerzett egyetemi diplomát 1998-ban. Négy évig doktoranduszként dolgozott a Távközlési és Telematikai Tanszéken, majd PhD fokozatot szerzett a Brandenburgische Technische Universitat (BTU) Cottbus-tól, Németországban, 2002-ben. Jelenleg a győri Széchenyi István Egyetem Távközlési Tanszékének oktatója, egyetemi docens beosztásban. Szakterülete a hang- és képtechnika, a telekommunikáció, akusztika. Aktuális kutatási témái: emberi térhallás vizsgálatok, lokalizáció, virtuális valóság szimulátorok, műfejes méréstechnika. Tagja az Audio Engineering Society-nek (AES), a Hírközlési és Informatikai Tudományos Egyesületnek (HTE), az OPAKFI-nak és az International Community for Auditory Display-nek (ICAD).



20. ábra
Egy, illetve két hallgatóra optimalizált elrendezés

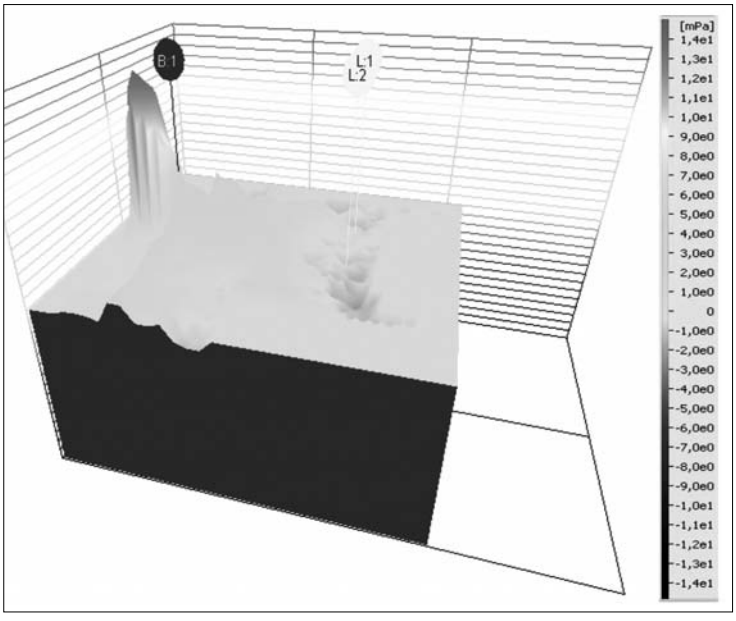
21. ábra
3D megjelenítés



Irodalom

[1] Tarnóczy T.: Hangnyomás, hangosság, zajosság. Akadémiai Kiadó, Budapest, 1984.
[2] Tarnóczy T.: Teremakusztika I-II. Akadémiai Kiadó, Budapest, 1986.
[3] http://www.catt.se/
[4] http://www.cara.de/
[5] Wersényi Gy.: Műszaki Akusztika, egyetemi jegyzet, 2004.
[6] Kotschy A.: Egy hangversenyterem akusztikai tervezése – tervezett és kész állapot. Akusztikai Szemle, VI. évf. 2., 2005, p.19–21.
[7] Tarnóczy T.: Akusztikai tervezés I-II. Akadémiai kiadó, Budapest, 1986.
[8] http://www.isover.hu/termekeink/owa/owacoustic/hangelnyeles.html
[9] http://www.isover.hu/acoustic/absorption/acoustics/contentframe.html
[10] http://www.audioease.com/Pages/Altiverb/AltiverbMain.html
[11] M. Kleiner, B.I. Dalenbäck, P. Svensson: Auralization – an overview. J. Audio Eng. Soc., Vol. 41, 1993, pp.861–875.

22. ábra
50 Hz-en a maximális hangnyomásszint-eloszlás a mélynyomó által kibocsátva



Prozódiai információ felhasználása a beszédfelismerés hatékonyságának növelésére

SZASZÁK GYÖRGY, VICSI KLÁRA

BME Távközlési és Médiainformatikai Tanszék
{szaszak, vicsi}@tmit.bme.hu

Lektorált

Kulcsszavak: beszédfelismerés, prozódia, szóhatár-detekció, prozódiai szegmentálás

Cikkünkben lényegében a mondat-, tagmondat- és szóhatárok prozódia alapú detektálását mutatjuk be rejtett Markov modelles módszerrel. Az így elkészült prozódiai szegmentálót beszédfelismerőbe építve a felismerési hipotéziseket újrasúlyozzuk annak alapján, hogy mennyire illeszkednek a detektált dallammenetre. Ultrahangos leletező alkalmazásban egyszerűsített nyelvi modellel a felismerési hatékonyság 3,82%-os javulását értük el.

1. Bevezetés

A prozódia – vagy szupraszegmentális szerkezet – az emberi beszéd szerves részét képezi, így például a hallgató számára segíti a közlés értelmezését azáltal, hogy a beszédet tagolttá teszi, kiemeli a fontos vagy új információt tartalmazó részeket. E funkcióján túlmenően hordozza a modalitást (kijelentő, kérdő stb.), illetve lehetőséget ad a beszélőnek érzelmei kifejezésére is, ami jelentős részben szintén a szupraszegmentumok révén valósul meg.

A beszédtechnológia műszaki oldaláról közelítve ma már elképzelhetetlen jó minőségű beszédszintézis a megfelelő prozódia – azaz a megfelelő hangsúlyozás és a természetes dallammenet – modellezése nélkül. A beszédfelismerésben azonban korábban szinte egyáltalán nem foglalkoztak a prozódiával, jóllehet a szupraszegmentumoknak nem csak a jelentést tagoló vagy árnyaló, hanem bizonyos esetekben a jelentés egy részét illetően egyfajta „hordozó” szerepük is van. Mindezt a gépi beszédfelismerés során is figyelembe kell vennünk, ha nem szeretnénk a közlésből lényeges információt elveszíteni. Így például bizonyos beszédinformációs rendszerekben alapvető követelmény lehet, hogy különbséget tudjunk tenni kérdések és kijelentések között akkor is, ha adott esetben a kérdő- és a kijelentő mondat ugyanazon szóláncból épül fel, különbség közöttük csak modalitásukban van [1]. Hagyományos beszédfelismerő rendszerrel – amely csupán az elhangzott beszédhang-szekvenciára koncentrál – ez a feladat megoldhatatlan.

Ezen a „magától értetődő” felhasználási területen túl a szupraszegmentális jellemzők alakulásának nyomon követése a beszédfelismerési feladat során egyéb haszonnal is járhat, amennyiben a beszédfolyam prozódiai eszközökkel történő tagolása segítségünkre van a beszédfelismerésben. A mondatok, tagmondatok, szószerkezetek vagy akár az egyes szavak határainak ismerete hasznos lehet a keresési tér csökkentésében, ha ugyanis biztosak vagyunk benne, hogy valahol a beszédfolyamban szóhatárt találunk, akkor azokat a felismerési hipotéziseket, amelyek az adott ponton nem tartalmaznak szóhatárt, kizárhatjuk. Ezáltal gyorsul és pontosabb lesz a felismerés, amelynek során – különösen a toldalékoló nyelvek esetében, amilyen a magyar nyelv is – sokszor problémát, de legalábbis korlátozó tényezőt jelent a valós idejű működés követelménye. További segítséget adhat a prozódia a lényeges információ automatikus kiemeléséhez a közlésből, illetve segítheti a szintaktikai elemzést is [3].

A nemzetközi porondon többen próbálkoztak már a prozódiai információ felhasználásával a beszédfelismerésben, elsősorban angol és német nyelven. Veilleux és Ostendorf például olyan algoritmust dolgoztak ki [10], amelyek az egyes hipotézis-gráfok közül az N darab legvalószínűbbet újrasúlyozzák (úgynevezett N-best rescoring) a prozódiai információk ismeretében, majd ezután az újrasúlyozott gráfokkal számítják a felismerés végeredményét a hagyományos módon. Hasonló munka készült a német nyelvre is [2]. Gallwitz és munkatársai ennél tovább lépve integrált gépi beszédfelismerőt készítettek [1], amely a beszédláncra vonatkozó, illetve a prozódiai információt egységesen kezeli és követi a felismerés során. A szerzők is végeztek már kísérleteket magyar nyelvre a prozódia beszédfelismerésben történő felhasználhatóságát illetően [12].

2. A prozódiai információ kinyerése a beszédből

A prozódiai jellemzők közül az alapfrekvencia, az energiaszint, és az időtartamok objektív mérésére van lehetőség. Ezen paraméterek közül az alapfrekvencia és az energiaszint értékeit találtuk jellemzőnek a hangsúly detektálása szempontjából korábbi vizsgálataink alapján [8]. E két prozódiai tényező értékeit a beszédjelből a Snack programcsomag [7] segítségével nyertük ki, majd előfeldolgozásnak vetettük alá.

Az alapfrekvencia számításához a Snack programban implementált AMDF-alapú algoritmust használtuk 25 ms ablakmérettel, 10 ms keretidővel. Ezt követően okta-

vugrást korrigáló szűrőt alkalmaztunk, amelyet 5 pontos átlagoló (mean) szűrő követett, majd az alapfrekvencia-értékek logaritmusát alapul véve lineáris interpolációt végeztünk annak érdekében, hogy az alapfrekvencia görbe többé-kevésbé folytonos legyen. Nem történt interpoláció olyan zöngétlen szakaszokon, amelyek hossza a 250 ms-ot meghaladta, illetve akkor sem, ha a zöngétlen szakasz utáni első zöngés keret alapfrekvenciája meghaladta a zöngétlen rész előtti 3 utolsó keret alapfrekvencia-értékei átlagának 1,1-szeresét.

Minderre azért volt szükség, hogy a 250 ms-nál hosszabb, ezért igen nagy valószínűséggel beszédszünetet tartalmazó szakaszokon az alapfrekvenciát ne interpoláljuk, mivel egyrészt a szünetet magát is szeretnénk a későbbiekben detektálni, másrészt ilyen hosszú szakaszon az interpolálás már túl durva közelítés lenne. Az alapfrekvencia-érték zöngétlen szakasz utáni emelkedését a zöngétlen szakasz előtti utolsó három érték átlagánál pedig azért nem engedjük magasabbra, mert ekkor valószínűbb, hogy a kérdéses szakaszon mondat, tagmondat vagy szószerkezet határa volt, és emiatt indít magasabbról az alapfrekvencia. E fenti értékeket tapasztalati úton állítottuk be, de a jövőben célszerű lehet ezeket a beszélőtől (beszédtempó, artikulációs sebesség stb.) függően meghatározni, ehhez azonban további vizsgálatok szükségesek, így a továbbiakban ezzel egyelőre nem foglalkozunk.

Az energiaszint-értékeket szintén a Snack programmal számítottuk, a keretidő ismét 10 ms volt, az alkalmazott ablakméret 25 ms, melyet szintén átlagoló (mean) szűrés követett. Mivel az energiaszint folytonosan számítható a beszédjelre, ezért itt interpolációra értelemszerűen nem volt szükség.

Ezután mind az alapfrekvencia, mind az energiaértékekhez első és másodrendű deriváltjaikat is kiszámítottuk. A deriváltak közelítésére alkalmazott (1) regressziós képletben a figyelembe vett környezetet három lépcsőben fokozatosan növelve valójában 3-3 első és másodrendű deriváltat képeztünk, rendre ±10, ±25 és ±50 keretnek megfelelően ablakolt (*W* az (1) képletben) minták alapján, így a véglegesen kapott jellemzővektor összesen 14 elemet tartalmazott: az eredeti, feldolgozott alapfrekvencia- és energiaértéket, és ezek mindegyikéhez 3-3 első- és másodrendű deriváltat. A deriváltak számítására használt képlet ([9] alapján):

$$d_t = \frac{\sum_{i=1}^{W} i(c_{t+i} - c_{t-i})}{2\sum_{i=1}^{W} i^2}, \qquad (1)$$

ahol d_t a t időpontban értelmezett derivált, c_{t-i} és c_{t+i} az eredeti (deriválandó) együtthatók, W pedig az ablakméret keretszámban.

3. A prozódiai információ felhasználása a beszédfelismerésben

A prozódia vizsgálatával és modellezésével célunk a beszéd durva felszegmentálása mondat- és tagmondat-határokon, illetve a szavak, szószerkezetek határainak minél pontosabb meghatározása. Az így nyert információt ezután a beszédfelismerőben felhasználva a felismerés hatékonysága javítható, illetve új, a beszédfelismerő felhasználási lehetőségeit kibővítő – a bevezetőben már említett – egyéb funkciók is megvalósíthatóak lesznek.

Munkánk során nagyban támaszkodunk arra a tényre, hogy a magyar nyelv kötött hangsúlyozású, azaz a hangsúly mindig a hangsúlyos szó első szótagjára esik [2]. Mindez azért rendkívül fontos, mert így lehetőségünk nyílik a prozódiai információ egységes kezelésére anélkül, hogy a beszédet szavak vagy beszédhangok szintjén is ismernünk kellene a prozódiai struktúra értelmezéséhez. Természetesen végcélunk ezzel együtt az, hogy a beszédhangok sorozatát jellemző spektrális, illetve a prozódiát befolyásoló szintaktikai információt egységesen kezeljük és használjuk fel a beszédfelismerésben.

3.1. Automatikus prozódiai szegmentáló betanítása

A prozódiai szegmentálás során az egyes mondatépítő szintaktikai elemek dallamtípusát szeretnénk felismerni és járulékosan a szintaktikai egységek határát a lehető legnagyobb pontossággal meghatározni. A szintaktikai egységek határai egyben szóhatárok is lesznek, amelyek egybeeshetnek tagmondatok vagy mondatok határaival is.

A felismerni szándékozott dallamtípusok kiválasztásánál tekintettel kell lennünk arra, hogy az egyes dallamok egymástól élesen elkülöníthetőek legyenek és felöleljék a leggyakrabban előforduló dallamváltozatokat. Az éles különbségtétel követelménye miatt mindössze 6 alapvető dallamtípust különíttettünk el a prozódiai felismeréshez. A szünet adja a 7. felismerendő „dallamtípust". A felhasználandó dallamtípusokat az *1. táblázatban* foglaltuk össze.

A prozódiai szegmentáló betanításához a tanítómintákat a BABEL beszédadatbázis [6] szöveganyagából választottunk ki (22 beszélő által bemondott 1600 mon-

1. táblázat
A felismerésre kiválasztott dallamtípusok

Címke	Dallam	Megjegyzés
me	változó	Mondat eleje.
fe	(emelkedő-) eső v. eső-ereszkedő	Erősen hangsúlyos szintaktikai egység.
fs	eső-ereszkedő	Mellékhangsúlyos szintaktikai egység.
mv	ereszkedő	Mondat vége.
fv	emelkedő	Folytatást jező szintaktikai határ.
s	ereszkedő vagy lebegő	Hangsúlytalan szakasz. Szünetet is kitölthet az F_0 interpolációja miatt.
sil	–	Szünet.

dat). Ezt a szöveganyagot kézzel, majd félautomatikusan felszegmentáltuk a táblázatban szereplő dallamtípusok szerint. A kézi szegmentálás az alapfrekvencia- és az energiakontúr alapján történt, a lehallgatás során kapott szubjektív ítéletet is figyelembe véve.

Az elkészült prozódiai szegmentáló rejtett Markov-modell alapú, a keretidő a már említett 10 ms, a Markov-modellek lineárisak, állapotai száma (optimalizálás után) 11. A prozódiai szegmentálót a HTK szoftvercsomag [9] felhasználásával valósítottuk meg.

3.2. Az automatikus prozódiai szegmentálás menete

Az automatikus prozódiai szegmentálás menete a beszédfelismerésben is használt algoritmusokkal az ott ismeretes lépések szerint történik, azaz a lényegkiemelést (előfeldolgozást) követi a dekódolás. Az előfeldolgozás a 2. szakaszban megismertek szerint történik, azaz az alapfrekvencia- és az energia-jelet az ott ismertetett módon nyerjük ki és dolgozzuk fel. A dekódolás során a Viterbi-algoritmus fut, amely a „rövid" jellemző vektorok, a modellek csekély száma és az alkalmazott nyelvtan miatt igen gyors.

A dekódolás során ugyanis a szintaktikai egységeknek megfelelő dallamtípusokra vonatkozóan szigorú nyelvtant vezetünk be, amely megadja, hogy azok milyen sorrendben követhetik egymást. Az ily módon létrehozott megszorítások tapasztalataink szerint a szegmentálás minőségét lényegesen javítják, ugyanakkor ehhez képest elhanyagolható azoknak az eseteknek a száma, amikor a szigorú, nem minden kivételes esetet leíró nyelvtan miatt történik tévesztés.

A nyelvtan a HTK-ban is alkalmazott jelöléseket alapul véve (vö. [9], p.163.)

$$Phrase = [sil] < [me]\{fe \mid fv[s]\}[mv][sil] > sil \quad (2)$$

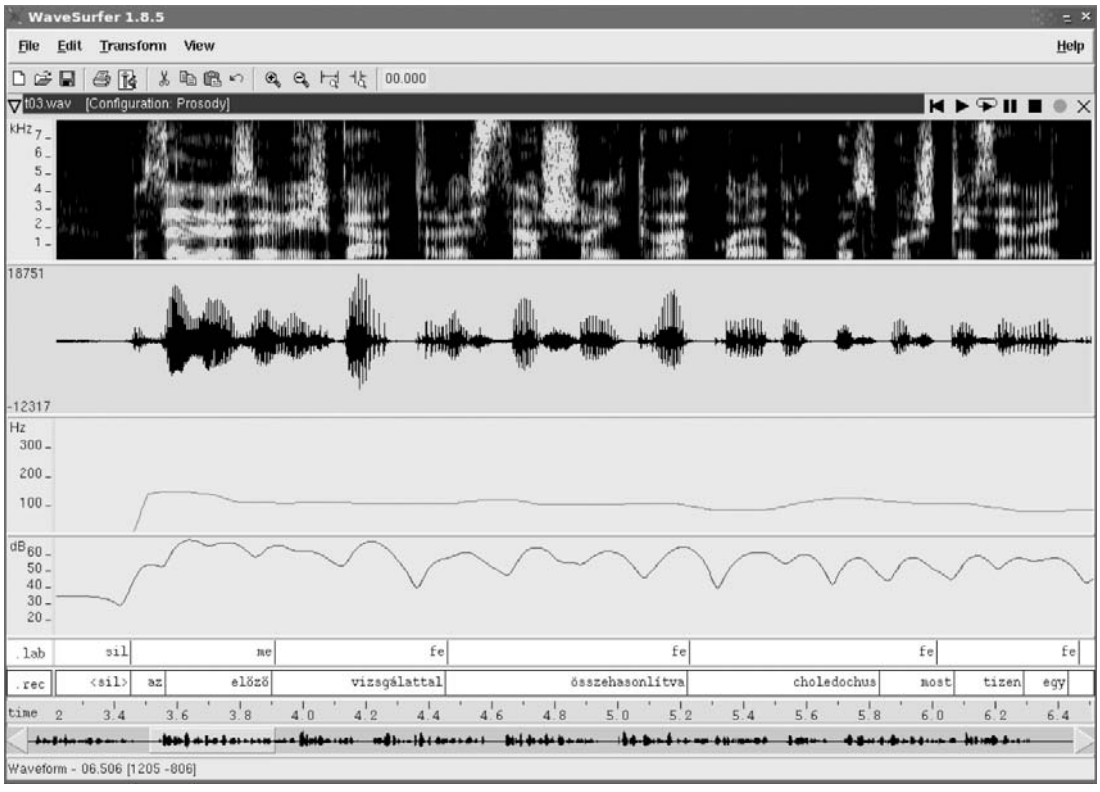
összefüggéssel írható le, melyben a '<>' szimbólumok egy vagy több, a '{}' nulla, egy vagy több ismétlődést jelölnek. A '|' szimbólum kizáró vagy kapcsolatot, a '[]' opcionálisan elmaradó eseményeket jelöl. Az ily módon formalizálva lejegyzett sorozatot tekintjük a prozódiai mondatmodellnek. A prozódiai szegmentálás eredményeként a felismert dallamtípusok kezdő- és végidőpontjukkal együtt ismeretté válnak. Az *1. ábrán* látható példa egy így nyert prozódiai szegmentálást jelenít meg.

3.3. Prozódiai szegmentáló beépítése beszédfelismerőbe

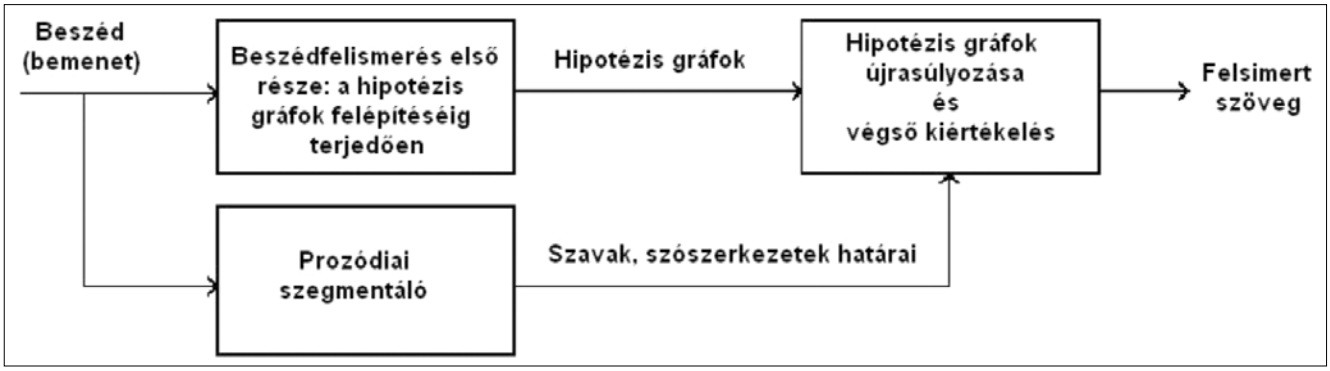
A prozódiai szegmentáló kimenetét felhasználhatjuk a beszédfelismerőben a keresési tér csökkentésére, így jobb felismerés és gyorsabb működés remélhető a beszédfelismerőtől. Ehhez a beszédfelismerés folyamatába kell avatkoznunk. Erre egy lehetőség az a pont, ahol a dekódolás részeként a hipotézis-gráfok felépítése – azaz voltaképpen a felismerés lehetséges eredményeinek felmérése és valószínűségeik kiértékelése – történik.

A prozódiai szegmentálás ismeretében a hipotézisgráfok újrasúlyozhatók, a felismerés további folyamatába pedig már az újrasúlyozott gráfok kerülnek, így a felismerés végeredményének kiértékelését már a prozódia alapján nyert információ is befolyásolja. Utolsó lépésként az újrasúlyozott hipotézis gráfon kell a maximális pontszámú utat megkeresnünk, amihez gyors keresőalgoritmusok állnak rendelkezésünkre.

1. ábra A prozódiai szegmentálás kimenete
„Az előző vizsgálattal összehasonlítva a choledochus most 11 milliméteres..." mondatrészletre.



A felső sávban a spektrogram, az alatta levő sávokban rendre az időfüggvény (hullámforma), az interpolált alapfrekvencia, az energia görbéje, a prozódiai szegmentálás, végül az elhangzott szöveg látható bejelölt szóhatárokkal.



2. ábra Prozódiai szegmentálóval kiegészített beszédfelismerő felépítése

A felismerés menetét a *2. ábrán* követhetjük végig.

3.4. A hipotézis-gráfok újrasúlyozása

Mint az előzőekben láttuk, a hipotézis-gráfok újrasúlyozása a prozódiai szegmentálás alapján történhet. Az alapötlet az, hogy azokat a szavakat és szóláncokat (a hipotézis-gráfból kinyerhető szósorozatokat, ezek egy-egy lehetséges felismerést írnak le), amelyek esetén a szavak határai időben egybecsengenek a prozódiai szegmentálás által jelzett határokkal, valamilyen módon részesítsük előnyben a felismeréskor, azaz a hozzájuk rendelt valószínűségi súlyt növeljük. Hasonlóképp, azokban az esetekben, amikor a prozódiai szegmentáló által megadott határok szavak belsejébe esnek, az eredetileg hozzárendelt súlyokat csökkenthetjük.

Problémát okoz azonban, hogy a prozódiai szegmentáló sem működik hibamentesen, azaz bizonyos százalékkal téves ítéletet hoz a szintaktikai egységek határait illetően. Spontán beszédben még gyakoribbak azok a jelenségek, amelyek az automatikusan futó algoritmust megzavarhatják (gyakoribbak a szótévesztések, javítások, előfordulhat, hogy a prozódia eltorzul, ha a beszélő „mondat közben meggondolja magát” és máshogyan folytatja közlendőjét, a hevesebben kifejezett érzelmek is befolyásolhatják a prozódiát stb.). A prozódiai szegmentáló teljesítményének kiértékelésével korábban részletesen foglalkoztunk [8], jelen esetben pedig a hibaszázalék pontos számszerű ismerete nélkül is beláthatjuk, hogy az újrasúlyozás során valamennyire a prozódiai információt is fenntartással kell kezelnünk.

Ügyelnünk kell továbbá arra is, hogy a prozódiai információ – éppen szupraszegmentális jellegéből adódóan – időben kevésbé pontosan lokalizálható, mint az egyes beszédhangok – így akár az egyes szavak – határai. Gondoljunk például arra, hogy ha egy adott szintaktikai egység utolsó beszédhangjaként zöngétlen hangot (különösen is, ha zöngétlen réshangot) találunk, a dallamban számunkra az utolsó „biztos” támpontot a legutóbbi magánhangzó jelenti. Ez máris egy beszédhanghossznyi bizonytalanságot jelent, amit a prozódiai szegmentáló a beszédhangsor ismeretének hiányában nem tud feloldani.

Éppen ezért a prozódiai szegmentáló által megjelölt határokat intervallummá transzformáljuk, azaz megengedünk bizonyos ΔT csúszást a prozódiai szegmentáló által megállapított határhoz (t_B) képest. Az intervallumon belül a ténylegesen előrejelzett határtól való távolság függvényében értelmezzük a határ adott időpontban történő elhelyezkedésének valószínűségét, pontosabban egy azzal arányos pontszámot (L_B) az alábbiak szerint:

$$L_B(t) = \begin{cases} A\cos(\frac{\pi}{2\Delta T}t) + C, ha\ t \in [t_B - \Delta T, t_B + \Delta T] \\ 0 \quad egyébként \end{cases}, \tag{3}$$

ahol A és C konstansok. A kísérleteinkben ΔT értéke 10 keret, azaz 100 ms volt. A cosinus függvényt az egyszerűség kedvéért választottuk, de lényeges, hogy minél távolabb van a határ az előre jelzettől, annál kisebb pontszámot rendelünk hozzá.

Mindezek után rátérhetünk a hipotézis-gráfok tényleges újrasúlyozásának algoritmusára, amelyet az alábbiakban mutatunk be. Elöljáróban annyit jegyezzünk még meg, hogy a hipotézis-gráf éleihez szavak vagy szóláncok, csomópontjaihoz pedig a megfelelő kezdő- és végidőpontok vannak rendelve. Az újrasúlyozáshoz minden, a gráfban található szót vagy szóláncot kigyűjtünk, majd kezdő- és végpontjaira pontszámot számítunk, amely annál nagyobb (lásd (3)), minél közelebb van a prozódiai szegmentáló által jelzett határhoz a szó eleje, illetve vége:

$$Sc_{remum} = w_a L_B(t_{start}) + w_b L_B(t_{end}), \tag{4}$$

ahol t_{start} a szó gráf szerinti kezdő, t_{end} a szó gráf szerinti végpontjának felel meg (az időben), w_a és w_b pedig súlyok.

Ezt követően a szó valamennyi i keretére – az első és utolsó k darab keret kivételével – összegezzük $L_B(t_i)$ értékeket, ahol t_i az aktuális keretidő:

$$Sc_{punish} = \sum_{i=k+1}^{N-k-1} L_B(t_i), \tag{5}$$

A fenti képletben N a szóhoz tartozó összes keret száma, k=ΔT=100 ms pedig ésszerű választásnak kínálkozik.

A gráf éléhez tartozó új $Sc_{rescored}$ pontszám értéke pedig:

$$Sc_{rescored} = w_O Sc_{orig} + w_P (Sc_{remum} - Sc_{punish}), \tag{6}$$

ahol Sc_{orig} a gráf éléhez eredetileg tartozó, most felülbírált pontszám (élsúly), w_O és w_P pedig súlytényezők.

4. Kísérlet a prozódiai szegmentáló beszédfelismerő rendszerbe való illesztésére

A prozódiai szegmentáló elkészítésével célunk a beszédfelismerés hatékonyságának növelése volt, így a továbbiakban egy olyan kísérletről számolunk be, melynek során a prozódiai szegmentálót automatikus gépi beszédfelismerőbe építettük be. A 3.3. szakaszban már szó esett a beépítés módjáról, az előzőekben (3.4. szakasz) pedig áttekintettük azokat az algoritmikus változtatásokat, amelyeket a felismerés folyamatában szükséges eszközölnünk.

A kísérlethez magyar nyelvű, folyamatos beszédfelismerőt választottunk ki, amely az orvostudománybeli radiológiai leletezés területét, azon belül is a hasi és kismedencei ultrahangos vizsgálatok leletezésénél használt szótárkészletet öleli fel. A szótár elemszáma viszonylag csekély, mintegy 4000 szó. A területre bigram nyelvi modell is készült, azonban jelen kísérletben a bigram nyelvi modellt binarizáltuk, azaz csak azt tüntettük fel benne, milyen szavak után milyen szavak előfordulása megengedett a szövegben. Ennek célja egyúttal annak kipróbálása is, hogy képes-e a prozódiai információ minimális nyelvtani információ mellett a felismerés hatékonyságát javítani. Ezzel egyúttal a későbbiekben tervezett nagyszótáras alkalmazások felé is tekintünk, ugyanis nagy szótárméret esetén a nyelvi modell elkészítéshez rendkívül nagy szövegadatbázis kell, a nyelvi modell használata pedig rendkívül műveletigényes. Különösen igaz ez az agglutináló nyelvekre – így a magyarra is – amelyek esetén viszonylag kis tématerület esetén is relatíve nagy az előforduló szóalakok száma a toldalékoló jelleg miatt.

A felismerő HTK környezetben implementált, felépítését tekintve a „klasszikus” 39 MFCC együtthatót alkalmazó, a kibocsátási valószínűségeket 32 Gauss függvény szuperponálásával leíró, 10 ms keretidejű rendszer. A felismerő betanításához az MRBA adatbázis [11] mintegy 8 órányi, részben beszédhang szinten felszegmentált anyagát használtuk fel, összesen 37 beszédhang modell készült.

Ebbe a felismerőbe építettük bele a prozódiai szegmentálót, és vizsgáltuk a felismerési eredmény változását. A (4) és a (6) képletekben megadott súlyok értékeit tapasztalati úton az alábbiakra állítottuk be: w_a=0,5, w_b=0,5, w_O=1, w_P=2,5.

4.1. Eredmények

A kísérleti rendszerrel hasi és kismedencei ultrahangos leletek felismerését vizsgáltuk összesen 20 darab leletre. (Egy lelet nagyságrendileg kb. 10-20 mondatot tartalmaz.) A felismerést azonos körülmények között azonos (rögzített, majd visszajátszott) leletekre először az alaprendszerrel, majd a prozódiai szegmentálóval kibővített rendszerrel végeztük el. Az eredményeket a *2. táblázatban* mutatjuk be 6 darab, a teljes tesztanyag tekintetében reprezentatívan kiválasztott leletre.

A táblázatban megjelenített mérőszámok a helyesen felismert szavak aránya, illetve a szótévesztési arány javulása, mindkettő százalékosan értendő.

A táblázatból látható, hogy a prozódiai szegmentálóval kibővített rendszer teljesítménye összességében 3,82%-kal javult. A javulás mértéke leletenként változó, egyes esetekben 10% fölötti eredményt is kaptunk (lásd pl. 03-as azonosítójú lelet), ugyanakkor előfordul (lásd pl. 16-os azonosító), hogy a felismerés nem javul, hanem éppenséggel romlik a prozódiai információ figyelembe vételekor.

Az egyes leletbemondásokat megvizsgálva arra a következtetésre jutottunk, hogy a prozódiai szempontból jobban – ezzel együtt a „megszokott hétköznapi”, általánosan elvárható kiejtésnél nem gondosabban – bemondott leletek felismerése a prozódiai információ figyelembe vételekor jelentősebb mértékben javul. Azokban az esetekben, amikor a felismerés a kibővített rendszerrel nem javult, hanem romlott, a hibát jellemzően a prozódiai szegmentáló tévesztése okozta, ami a hipotézis gráfok újrasúlyozásakor eltorzította a felismerést. A hiba forrása esetenként a prozódiailag gondatlan beszéd, esetenként az alapfrekvencia-detektor tévesztése volt. Ez utóbbi történhet például egy kissé rekedt hangú beszélőtől származó lelet esetében.

Általános tapasztalatunk, hogy a prozódiai szegmentálás esetenként időben kevésbé olyan pontos, mint ami a szóhatárok helyének biztosabb megállapításához szükséges lenne. Úgy gondoljuk, ez utóbbi probléma legalább részben orvosolható, amennyiben a prozódiai szegmentálóban figyelembe vesszük az adott szót felépítő fonémasorozatot.

Korábban már említettük, hogy a prozódia követésének szempontjából számos szóvégi zöngétlen hang máris hosszának megfelelő, durván 50-100 ms, de akár 150 ms nagyságrendjébe eső bizonytalanságot is eredményezhet a szóhatár megítélésében, mivel az alapfrekvencia menetére ekkor általában nem támaszkodhatunk. Amennyiben lehetőségünk van figyelembe venni az aktuálisan elhangzó beszédhangokat is, időben korrigálni tudjuk a prozódiai alapján előrejelzett és a tényleges szóhatárok eltérését, vagy legalábbis tud-

2. táblázat
A helyesen felismert szavak arányának alakulása az alaprendszer és a kibővített rendszer esetén, illetve a szótévesztési arány javulása.

Lelet-azonosító	Helyesen felismert szavak [%]		A szótévesztési arány változása (relatív) [%]
	Alap rendszer	Kibővített rendszer	
03	71,2	78,9	10,9
07	78,8	80,6	3,6
08	84,6	84,6	0,0
10	70,8	72,2	2,0
16	68,3	66,7	-2,4
19	83,8	90,5	8,1
Összes lelet (20)	**75,99**	**78,89**	**3,82**

juk, hogy az adott helyen mennyiben támaszkodhatunk a prozódiai szegmentálás pontosságára. További kutatásaink során mindenképpen szeretnénk ilyen jellegű vizsgálatokat végezni, ezáltal a rendszert teljesebbé tenni.

Megjegyezzük továbbá, hogy a prozódiai szegmentáló által meg nem jelölt szintaktikai vagy szóhatárok a beszédfelismerést a hipotézis gráfok újrasúlyozási algoritmusából kifolyólag nem rontják, jóllehet érdekünkben áll minél több szóhatárt megtalálni, ezáltal több lehetőséget adva a prozódiai információt nem használó felismerés hatékonyságának növelésére. A prozódiai szegmentálótól nem várhatjuk el, hogy valamennyi szóhatárt megtaláljon (erre még gyakorlott szakember sem vállalkozhat pusztán az alapfrekvencia és az energiaértékek ismeretében), ugyanakkor bebizonyosodott, hogy a megtalált szóhatárok alapján a felismerés hatékonysága javítható.

5. Összegzés

Írásunkban azt vizsgáltuk, hogyan használhatók fel bizonyos szupraszegmentális jellemzők a beszédfelismerés segítésére. Bemutattunk egy automatikus prozódiai szegmentálót, amely az alapfrekvencia és az energiaszint értékei alapján próbálja meg felismerni a dallam 6 kiválasztott alaptípusát, illetve a szünetet. A prozódiai szegmentálót beszédfelismerőbe építve arra használjuk, hogy a dallamtípusok felismerése révén megtaláljuk az egyes szintaktikai egységek közötti határokat, amelyek egyúttal szóhatárokat is jelentenek.

Az elvégzett kísérletek tanúsága szerint a szóhatárok ismerete révén a felismerés hatékonysága a hipotézis gráfok újrasúlyozásával javítható (arról nem is beszélve, hogy egyes írásjelek, így a vessző kitételében is nagy segítséget adhat a szintaktikai határok ismerete). Mindezek mellett a prozódiai szegmentáló akár automatikus szintaktikai elemzőkben is felhasználható lehet.

A szerzőkről

Szaszák György 2002-ben végzett a Budapesti Műszaki és Gazdaságtudományi Egyetem Villamosmérnöki és Informatikai Karán. Ez évtől kezdődően a BME-TMIT Beszédakusztikai Laboratóriumában dolgozik, főbb kutatási területe a gépi beszédfelismerés, ezen belül beszédadatbázisok konstrukciója és feldolgozása, beszédfelismerés rejtett Markov-modellekkel, ejtésvariáció modellezés, szupraszegmentális jellemzők felhasználása a beszédfelismerésben, érzelmek felismerése akusztikai beszédjel alapján. Doktorjelöltként PhD dolgozata megvédésére készül.

Vicsi Klára a BME TMIT Beszédakusztikai Laboratórium vezetője. Beszédfelismerési témában írta meg PhD-jét 1992-ben. A MTA Mérnöki Tudományok Doktora lett 2004-ben, BME habilitációja pedig 2007-ben volt. Számos korábbi hazai és nemzetközi kutatás témavezetője és jelenleg is aktiv projektvezető a beszédakusztika, a gépi beszédfelismerés, beszéd adatbázis készítés és a pszichológiai akusztika területén. Előszeretettel foglalkozik beszédsegítő eszközök létrehozásában nagyothalló és beszédhibás személyek részére. A lektorált hazai és nemzetközi folyóiratokban, nemzetközi konferenciakiadványokban több mint 65 publikációja jelent meg. Szerzője számos beszédkutatással foglalkozó könyvrészletnek.

Irodalom

[1] Gallwitz, F., Niemann, H., Nöth, E., Warnke, V.: Integrated recognition of words and prosodic phrase boundaries. Speech Communication, Vol. 36, 2002, pp.81–95.

[2] Kassai Ilona: Fonetika. Tankönyvkiadó, Budapest, 1998.

[3] Kompe, R.: Prosody in Speech Understanding Systems. LNAI 1307, Springer Verlag, Berlin-Heidelberg, 1997.

[4] Kompe, R., Kiessling, A., Niemann, H., Nöth, H., Schukat-Talamazzini E.G., Zottman, A., Batliner, A.: Prosodic scoring of word hypothesis graphs. In: Proceedings of the European Conference on Speech Communication and Technology, Madrid, 1995, pp.1333–1336..

[5] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A.: Stochastic pronunciation modelling from hand-labelled phonetic corpora. In: Modeling Pronunciation Variation for ASR. Rolduc, 1998, pp.109–116.

[6] Roach, P.S. et al.: BABEL: An Eastern European Multi-language database. Int. Conference on Speech and Language, 1996.

[7] Sjölander, K., Beskow, J.: Wavesurfer – an open source speech tool. Proc. of the 6th International Conference of Spoken Language Processing in Beijing, China, Vol. 4, 2000, pp.464–467.

[8] Szaszák György, Vicsi Klára: Folyamatos beszéd szószintű szegmentálása szupra-szegmentális jegyek alapján. III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2005, pp.360–370.

[9] Young, S. et al.: The HTK Book (for version 3.3). Cambridge: Cambridge University, 2005.

[10] Veilleux, N.M., Ostendorf, M.: Prosody/parse scoring and its application in ATIS. Human Language and Language and Technology. Proc. of the ARPA Workshop, Plainsboro, 1993. pp.335–340.

[11] Vicsi K., Kocsor A., Tóth L., Velkei Sz., Szaszák Gy., Teleki Cs. , Bánhalmi A., Paczolay D.: A Magyar Referencia Beszédadatbázis és alkalmazása orvosi diktálórendszerek kifejlesztéséhez. III. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, 2005, pp.435–438.

[12] Vicsi, K., Szaszák, Gy.: Automatic Segmentation of Continuous Speech on Word Level Based on Supra-segmental Features. International Journal of Speech Technology, Vol. 8., No.4, 2005, pp.363–370.

Multi-modális gépi sakkozó – Török-2

KOVÁCS GYÖRGY, SAJÓ LEVENTE, FAZEKAS ATTILA

Debreceni Egyetem Informatikai Kar, Debreceni Képfeldolgozó Csoport
{attila.fazekas, sajolevente, gykovacs}@inf.unideb.hu

Lektorált

Kulcsszavak: multimodális ember-gép interakció, arci érzelem felismerés, arc detektálás

Az információs rendszerek használatának egy új módját jelenti a multi-modális ember-gép kommunikáción alapuló technikák használata. Ebben a cikkben egy ilyen technikán alapuló gépi sakkozó megvalósítását ismertetjük. Az ember-gép kommunikáció multi-modális mivolta abban nyilvánul meg, hogy billentyűzet nélkül, a verbális és a gesztusnyelv csatornáit is felhasználva sakkozhatunk a géppel.

1. Bevezetés

Az információs társadalom egyik alapvető igénye az információkhoz való hatékony hozzáférhetőség biztosítása. Az elmúlt évtizedben a technológia fejlődése egyre hatékonyabb eszközöket adott a kezünkbe az információ tárolására, rendszerezésére és lekérdezésére, továbbá az élet számos területén megjelentek olyan eszközök, amelyek feladata a tárolt információk lekérésének biztosítása. Ezen eszközöket gyűjtőnéven információs rendszereknek nevezzük. Ebbe a kategóriába tartoznak az általános célú számítógépek, de azok a speciális számítógépek is, amelyek valamilyen jól definiált célra készültek, mint például a menetrendekkel kapcsolatos információk tárolása, visszakeresése és az esetleges foglalások, illetve jegyek on-line vásárlásának lebonyolítására.

Annak ellenére, hogy az információs rendszerek életünk számos területén jelen vannak, a használatukkal szemben egyfajta ellenállás figyelhető meg. Ez egyrészt a technológia használatához szükséges ismeretek hiányának, másrészt az egyes rendszerek használati módja közötti különbözőségeknek tudható be. Ez azt jelenti, hogy életünk minden pillanatában újabb és újabb ismereteket kell elsajátítanunk az információs rendszer használatához. Ráadásul az információs rendszerek egy-egy új generációjának megjelenése egyre rövidebb idő alatt következik be és életünk egyre több területén hódítanak teret maguknak.

Az információs rendszerek használatának módja egyfajta kommunikációs nyelvnek a használatához hasonlít. Ha sok, különböző információs rendszerrel való kommunikációhoz szükséges nyelvet ismerünk, akkor az egy-egy újabb ilyen kommunikációs nyelv elsajátításához szükséges idő egyre kevesebb lesz. Továbbá, ha egy információs rendszer kommunikációs nyelvét napi szinten használjuk, akkor általában könnyebb lesz elsajátítanunk egy újabb generációjának a kommunikációs nyelvét.

A fentiek alapján teljesen nyilvánvaló, hogy jelenleg minden esetben a felhasználónak kell megtanulnia az adott információs rendszer kommunikációs nyelvét, azaz azt, hogy milyen formában adhatunk utasításokat a rendszernek, illetve milyen formában kapjuk meg a rendszerben tárolt információkat. A multi-modális ember-gép kommunikációval kapcsolatos kutatások alapgondolata az, hogy találjuk meg annak a módját, hogy a jövő információs rendszerei képesek legyenek a felhasználóval a számukra legtermészetesebb módon kommunikálni, azaz lehetővé tenni az emberi nyelv használatát. Egy felhasználónak nem kell újabb és újabb kommunikációs nyelvet elsajátítania, hanem elegendő „szóba elegyednie” az adott rendszerrel.

Minden szempont alapján ideális multi-modális ember-gép kommunikációs rendszer még nem készült el. Ezzel is magyarázható, hogy az ipari fejlesztések területén a kivárás a jellemző. A jelenleg hazánkban elérhető technológiák képességeinek bemutatása érdekében döntöttünk úgy, hogy elkészítjük a multi-modális gépi sakkozónkat, amelyet – tisztelegve Kempelen Farkas sakkozógépe előtt – Török-2-nek neveztünk el.

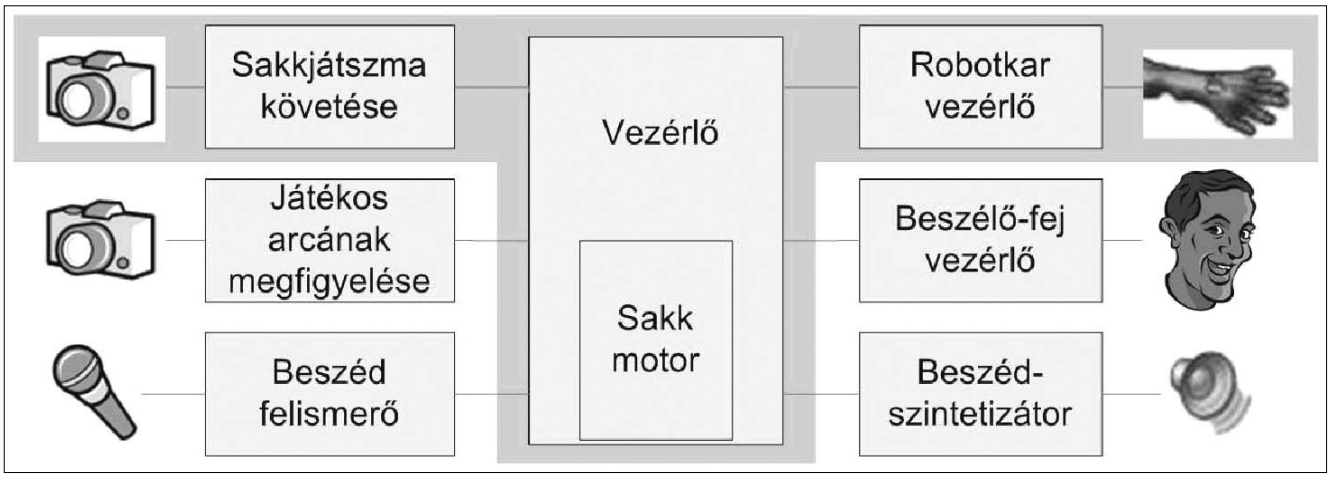
E cikk célkitűzése az, hogy az olvasó számára röviden vázolja a Török-2 felépítését és az egyes komponensek funkcionális szerepét. Áttekintjük az alapkoncepciókat, a rendszer működését, majd a rendszert alkotó egyes komponensek működését, funkcióját mutatjuk be.

A multimodális ember-gép kapcsolaton alapuló rendszerek felhasználói értékelése interdiszciplinális feladat. A rendszer használata közben készült videókat nyelvészek és pszichológusok értékelik ki a felhasználó kommunikációs tevékenységére, gesztusaira, a rendszerbeli elmélyülésére összpontosítva.

2. A Török-2 általános felépítése

A Török-2 rendszer komponenseit és a komponensek közötti kapcsolatot a következő oldali, *1. ábra* szemlélteti [3].

A rendszer koncepciója, hogy egy virtuális játékost valósítsunk meg, aki a játék szempontjából a lehető legtöbb tekintetben emberként viselkedik. A virtuális sakkozó rendszert funkcionálisan két fő részre bonthatjuk:



1. ábra A Török-2 rendszer általános felépítése

a sakkjátszma lebonyolítását biztosító modulra, valamint az ember-gép kommunikációt szolgáló felület modulra.

A sakkjátszmáért felelős modul szintén több kisebb komponensből épül fel. Mivel a játszma egy valódi sakktáblán zajlik, szükségünk volt egy eszközre, amely a virtuális játékos karját helyettesíti, azaz képes sakklépéseket végrehajtani. Ezt egy négy szabadsági fokkal rendelkező robotkar végzi. A modul bemeneti interfésze egy webkamera, amely a sakktábla fölött helyezkedik el. Ez a kamera felelős a játékállás felismerésért, a játszma követéséért.

Az ember-gép kommunikációt megvalósító komponens több emberi kommunikációs csatornát felhasználva, multi-modális kapcsolatot tesz lehetővé. A bemeneti interfészek hardver elemei: a webkamera, amely az emberi játékos arcát figyeli és egy mikrofon a játékos hangjának rögzítésére. A webkamera interfészt képfeldolgozási módszereket megvalósító szoftver egészíti ki, mely képes felismerni, hogy ül-e játékos a sakktábla előtt, vagy sem, felismeri a játékos nemét, életkorát és a játék alatt a játékos arcán megjelenő érzelmeket figyeli. A rögzített hangot a beszédfelismerő szoftver dolgozza fel, és a játékmenet vezérlésével kapcsolatos kulcsszavakat detektál.

Az ember-gép kommunikáció kimeneti interfészének hardver egysége a hangszóró mellett a monitor, a szoftver komponense pedig a szövegfelolvasó szoftver mellett az érzelmek kifejezésére is képes, animált beszélő fej [4].

2.1. A sakkozógép felépítése

A sakkozógépünket a mechanikus robotkar, mint kimeneti interfész, a sakkállás-felismerő, mint bemeneti interfész, egy sakkmotor, valamint a vezérlő egység alkotják.

2.1.1. A vezérlő

A vezérlő realizálja a virtuális játékos tudatát, vagyis a játék aktuális állapotának megfelelően vezérli és szinkronizálja az egyes komponenseket: sakkozógép esetén a robotkar és a sakkállás-felismerő megfelelő időzítése kulcsfontosságú.

2.1.2. A sakkállás-felismerő

A sakkállás-felismerő modullal szemben támasztott elsődleges követelményünk volt, hogy valós időben működjön. Ehhez úgy alakítottuk ki a fizikai környezetet, hogy az ideális legyen a sakkállás optikai felismeréséhez anélkül, hogy a sakktáblát, vagy a figurákat megváltoztatnánk. Ezzel sikerült elkerülnünk a költséges számításokat.

A webkamera a tábla fölött, középen helyezkedik el, olyan magasságban, hogy a perspektívából adódó geometriai torzítások elhanyagolhatóak legyenek. Mivel a figurák felülnézetből nem különböztethetők meg, a sakkállás-felismerő csak egy kiindulási állapothoz relatívan bekövetkező változásokat tudja detektálni.

A sakkfigurák elhelyezkedésének robosztus felismerésére két különböző módszert használunk, az első módszerrel a képet élképpé alakítjuk, azaz kiemeljük a lokális intenzitáskülönbségeket, majd ezt követően az úgynevezett Hough-transzformációval a sakkfigurák a priori ismert méretének megfelelő köröket keresünk az egyes mezőkben. A második módszerrel az egyes mezők lokális hisztogramjai alapján következtetünk arra, hogy van-e figura az adott mezőn, vagy sem. A két módszer együttesen megbízhatóan működik. [2]

Miután tudjuk, hogy mely mezőkön van figura, meghatározzuk azok színét az őket tartalmazó mezők középső régióinak világosságértékeiből.

A sakkállás-felismerő komponens a sakkállás felismerésen kívül képes érzékelni, azt, hogy a játékos mikor nyúl be a sakktábla fölé. Ezt az információt a virtuális játékos kezeli a következő módon: amint az emberi játékos benyúl a sakktábla fölé, a beszélő fej a képernyőn abbahagyja a „bámészkodást” és a sakktáblára néz, kíváncsian várva a lépést.

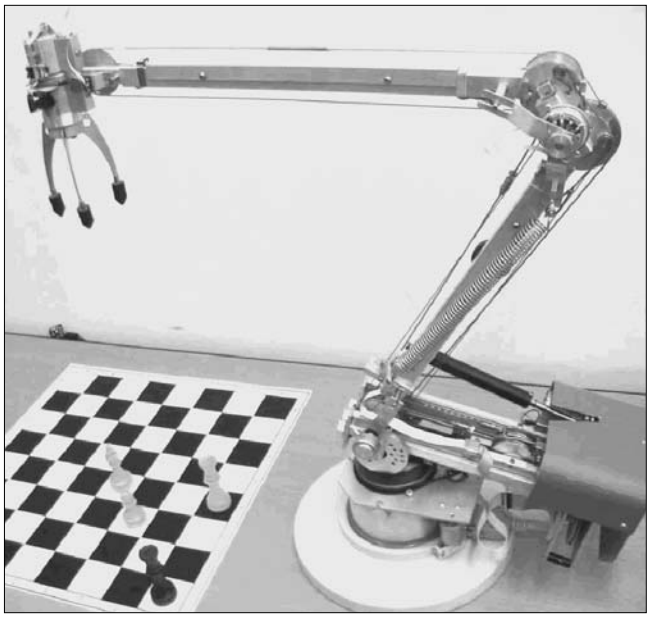
2.1.3. A robotkar

A robotkar kimondottan ehhez a rendszerhez lett kifejlesztve, ennek megfelelően a következő követelményeknek kell megfelelnie:

– tudja elérni a tábla legtávolabb eső mezejét is,
– a megfelelő figurához a környező bábok érintése nélkül férjen hozzá,

– tudjon megfogni bármilyen alakú sakkfigurát,
– a bábot függőlegesen emelje föl, és tegye le,
– a bábok elhelyezését megfelelő időn belül hajtsa végre,
– lehetőség legyen a sakktáblán kívülre történő pozícionálásra is.

A robotnak négy szabadsági foka van: kettőt a vállizület, egyet-egyet pedig a könyök és csuklóízület valósít meg. A robot működési területe egy negyed gömb. Technikai és anyagi okok miatt a robot elektromos és mechanikai alkatrészekből áll, a meghajtást villanymotorok, az erőtovábbítást pedig bowden huzalok végzik *(2. ábra)*.



2. ábra A robotkar

2.2. A virtuális játékos

A virtuális játékos komponens az ember-gép kommunikáció megvalósítására szolgál. Bemeneti csatornája az emberi kommunikációban is használt beszéd, valamint az arci gesztusok felismerése.

2.2.1. Érzelemfelismerés

Az emberi arc önmagában is egy információhalmaz, amelyből mi, emberek bármikor ki tudjuk nyerni az életkort, nemet és érzelmi állapotot. Ahhoz azonban, hogy mindezt fel tudjuk használni a számítógéppel történő kommunikációban, a felismerést számítógéppel kell végeznünk, ami összetett képfeldolgozási feladat. A fent említett információk (érzelem, nem, életkor) kinyerésére statisztikai tanuló algoritmusokat alkalmazunk (Support Vector Machine). A gyakorlatban azonban számos előfeldolgozási lépést kell tennünk ahhoz, hogy az arcokból információt nyerhessünk ki: meg kell találnunk az emberi játékos arcát a képen (arcdetektálás), valamint követnünk kell azt mozgás során (arckövetés), ráadásul mindezt, beleértve a tanuló algoritmusok osztályozását is, valós időben kell végeznünk.

A rendszer jelenlegi állapotában másodpercenként 2-3 érzelemdetektálást tud végrehajtani [1]. Mivel az érzelmek nem váltakoznak gyorsan, a videófolyamot felhasználva a korábbi detektált érzelmek alapján átmeneti valószínűségeket figyelembe véve a módszer még robosztusabbá tehető.

2.2.2. Beszédfelismerés

A sakkjáték tulajdonságai miatt nincs szükség tetszőleges szöveg felismerésére, elegendő egyes izolált kifejezéseket azonosítani, amelyeknek az ellenfél kiválasztása illetve a játék újrakezdése esetén van szerepe.

2.2.3. Beszélő fej

A virtuális játékos komponens kimeneti interfésze a beszélő fej, amely az emberi játékossal szemben lévő képernyőn jelenik meg. Két kimeneti csatornát használunk: a szintetizált beszédet, valamint a beszélő fejen megjelenő gesztusokat. A beszédszintetizálás a ProfiVox TTS rendszerrel történik [5], míg a megjelenő animált fej a CharToon rendszerre épült [6].

A beszédszintetizátor minden verbális megnyilvánulás esetén előállít egy hanghullám-állományt, amely a szintetizált beszédet tartalmazza, valamint egy időparaméterekkel ellátott fonéma-szekvenciát, amely azt az információt tartalmazza, hogy melyik pillanatban milyen fonéma hallható. A beszélő fej szájának animálásához definiáltuk a magyar nyelv egyes fonémáinak kiejtésekor megjelenő arcokat, ezeket a fonémához tartozó vizémának nevezzük. A beszéd animálása tehát a képernyőn megjelenő fejhez definiált vizémák lináris interpolálása olyan módon, hogy a hangállomány párhuzamos lejátszásakor a fonéma és vizéma párok a megfelelő pillanatban jelenjenek meg.

A másik csatorna a beszélő fej esetén az érzelmek kifejezése. Jelenleg 4 érzelmi állapotot tudunk megjeleníteni: természetes, szomorú, vidám, unott állapotokat *(3. ábra)*. Mindemellett a fej véletlenszerűen kisebb arcgesztusokat tesz (pislogás, szájhúzogatás), ezzel is élethűbbé téve a fej viselkedését. A véletlenszerű folyamatok vezérlésére a fej esetén a Perlin-zaj függvényt használjuk, mivel az más zajfüggvényeknél jobban közelíti a valós világban előforduló természetes viselkedésmintákat.

3. A rendszer működése

Ha nincs játékos a táblánál, a rendszer felismeri azt, és a beszélő fej embereket szólít meg, hogy üljenek le, és játsszanak vele. Ha valaki úgy dönt, hogy leül, azt a korának és nemének megfelelő köszöntéssel üdvözli.

3. ábra A beszélő fej érzelmi állapotai

Ezt követően játékost választhatunk. A virtuális játékos kiválasztása után kezdődik a játék. Az emberi játékos kezd. Lépését a sakkállás-felismerő modul felismeri és a sakkmotor meghatározza a következő lépést, melyet a robotkar elvégez. Mindeközben a játékossal szemben lévő webkamera képét feldolgozó szoftver érzelmeket detektál, ezen érzelmeknek és a sakkjátszma állapotának függvényében a beszélő fej az arcán érzelmeket kifejezve, az érzelmi állapotának megfelelő megnyilvánulásokkal kommentálja a játékot. A virtuális játékos igyekszik kapcsolatban maradni és kommunikációt folytatni az emberi játékossal annak lépése során is, sőt ha kell egy kicsit sürgetni lépése megtételére. A játék természetesen valamelyik játékos győzelméig tart, ami után egy újabb játékot ajánl fel a rendszer.

4. Összefoglalás

A rendszer tesztelése folyamatban van, eredményekről még nem tudunk számot adni. Összefoglalva tehát van egy rendszerünk, amely a multimodális ember-gép kapcsolatok lehetőségeit demonstrálja. A rendszer komponensei (érzelem-, nem-, kor-felismerés, beszélő fej, robotkar, beszédfeldolgozás) függetlenek egymástól, a vezérlő komponens módosításával újabb alkalmazásokat alakíthatunk ki.

A sakkozó tesztelésével párhuzamosan lassan elkészül a multimodális dámajátékos. A meglévő komponensek felhasználásával és újabb modulok (például kézgesztusok felismerése) fejlesztésével szeretnénk megvalósítani nem táblás játékokat is, például barkochba, kő-papír-olló.

A szerzőkről

Fazekas Attila egyetemi docens a Debreceni Egyetem Informatikai Karán, ahol digitális képfeldolgozással és alakfelismeréssel kapcsolatos kutatásokat folytat 1992 óta. PhD fokozatát Matematika és Számítástudomány területén 1999-ben kapta meg. Végzés után tudományos ösztöndíjasként (1992-1995), majd egyetemi tanársegédként (1995-2000), illetve egyetemi adjunktusként (2000-2005) dolgozott a Kossuth Lajos Tudományegyetemen (2000-től jogutódján a Debreceni Egyetemen). Dr. Fazekas elnöke az International Association for Pattern Recognition magyarországi tagszervezetének, azaz a Képfeldolgozók és Alakfelismerők Társaságának. Közel 60 publikációja jelent meg a szakma folyóirataiban és konferenciák kiadványaiban.Közel 40 alkalommal tartott előadást különböző nemzetközi és hazai fórumokon kutatásainak eredményéről. Munkáját számos alkalommal elismerték, többek között 1997-ben a Kalmár László Alapítvány díját, 2002-ben a Magyar Tudományos Akadémia Bolyai János Kutatási Ösztöndíját, 2004-ben a Debreceni Egyetem Informatikai Karának díját, 2006-ban a Matsumae International Foundation kutatási ösztöndíját kapta meg. Számos konferencia szervezője volt, szakmai folyóiratok bírálója.

Sajó Levente jelenleg a Debreceni Egyetem Informatikai Karának másodéves PhD hallgatója multimodális ember gép kapcsolat és arci érzelemfelismerés témákban. Egyetemi tanulmányait 2000-ben kezdte Programtervező Matematikus szakon. Már hallgató korában bekapcsolódott a Képfeldolgozó Csoport munkájába, az elért eredményeket TDK dolgozat formályában publikálta. Sikeresen pályázott meg egy Finnországi tanulmányi ösztöndíjat, melynek következtében 2005-ös évben a tanulmányait külföldön folytathatta. Hazatérése után szerezte meg MSc diplomáját (2006), sikeresen felvételizett PhD-ra és tovább folytatta az egyetemi hallgató korában megkezdett munkát. Az eddigi eredményeit az Universitas Alapítvány ösztöndíjával ismerték el.

Kovács György a Debreceni Egyetem Informatikai Karának elsőéves PhD hallgatója. Egyetemi tanulmányait 2002-ben kezdte programtervező-matematikus szakon. Már hallgató korában bekapcsolódott a kar munkájába, demonstrátorként tevékenykedett, nyári szakmai ösztöndíjat több ízben nyert el, köztársasági ösztöndíjas volt. Több kutatási projektben is részt vett, majd miután megszerezte MSc. diplomáját (2007), PhD-hallgatóként folytatta tanulmányait képfeldolgozás területen.

Irodalom

[1] A. Fazekas, Gy. Hingyi, L. Sajó,
Multi-modális gépi sakkozó – Török 2,
Proc. of KÉPAF'07,
25-27 January 2007, Debrecen, Hungary, pp.173–181.

[2] A. Fazekas, A. Nagy, L. Sajó,
Török 2 gépi sakkozó,
Proc. of KÉPAF'07,
25-27 January 2007, Debrecen, Hungary, pp.182–189.

[3] A. Fazekas, L. Sajó,
Multi-modal human-computer chess player: The Turk 2,
Proc. of ITI'07,
25-28 June 2007, Dubrovnik, Croatia, pp.29–30.

[4] Gy. Kovács, Zs. Ruttkay, A. Fazekas,
Virtual Chess Player with Emotions,
Proc. of 4th Hungarian Conference on Computer Graphics and Geometry,
13-14 November 2007, Budapest, pp.182–188.

[5] Olaszy, G., Németh G., Olaszi, P., Kiss, G., Gordos, G.,
"PROFIVOX – A Hungarian Professional TTS System for Telecommunications Applications",
International Journal of Speech Technology,
Vol. 3, No.3/4, December 2000, pp.201–216.

[6] Zs. Ruttkay, H. Noot,
Animated CharToon Faces.
Proc. of NPAR 2000 – 1st International Symposium on Non-Photorealistic Animation and Rendering,
June 2000, pp.91–100.

A Huszty Dénes Alapítvány
a hazai akusztikai szakma fejlődéséért

HUSZTY GÁBOR

az Alapítvány Kuratóriumának titkára

Az Alapítvány

A közhasznú alapítványt a család, az Entel Műszaki Fejlesztő Kft., a Hírközlési és Informatikai Tudományos Egyesület, valamint az Optikai, Akusztikai és a Film- és Színháztechnikai Tudományos Egyesület hozták létre 2001-ben. Az induló pénzügyi eszközöket, valamint az emlék-plaketteket az első két alapító bocsátotta rendelkezésre. A 2001. november 15-én a Fővárosi Bíróság által bejegyzett Huszty Dénes Alapítvány célja, hogy az akusztika, vagy elektroakusztika területén tevékenykedő fiatal szakemberek, felsőfokú tanulmányaikat éppen befejező, vagy már végzett fiatalok – a pályázat beadásakor 35. életévüket még be nem töltött fiatal akusztikusok – olyan kiemelkedő eredményeit jutalmazza, melyek hozzájárulást jelentenek az akusztika egyetemes fejlődéséhez. Az Alapítvány további célja, hogy emléket állítson Huszty Dénes munkásságának, aki az 1950-1979 közötti időszak kiemelkedő akusztikai szaktekintélye volt.

Huszty Dénes (1927-1979) a II. világháború utáni Magyarországon az elektroakusztika egyik legjelentősebb szakmai vezetője, kutatója, szerzője és nemzetközileg is nagyra becsült szaktekintélye volt. Maradandót alkotott a tudományos kutatásban, műszaki fejlesztésben, a szakmához kapcsolódó gyártástechnológiában, a hírközlésben és stúdiótechnikában, a nemzetközi és hazai szabványosításban, az elektroakusztika menedzselésében és üzletpolitikájában. A 30 éven át tartó gazdag szakmai tevékenység, amely az ORION Rádió és Villamossági Vállalattal, a VIDEOTON elődjével (Vadásztölténygyár), az Elektroakusztikai Gyárral (BEAG), a Magyar Rádióval és az MTA Akusztikai Kutató laboratóriummal, valamint a bolgár elektroakusztikai iparral kapcsolódott össze, 1979-ben szakadt meg. Munkássága valamennyi gyártónál megalapozta a hangszórók, hangsugárzók tömeggyártását. A 70-es évek elején hazánkban évente már mintegy 1,3 millió darab hangszórót gyártottak, elsősorban az ő fejlesztéseinek eredményeként. A hangátvitel és az ahhoz kapcsolódó technológiai területeken közel 100 szabadalom bizonyítja mérnöki tehetségét. A nemzetközi szabványosításban ma is több élő, bevált kezdeményezését sikerült elfogadtatni. A stúdiótechnikai OIRT szabványosításnak elfogadott és elismert szaktekintélye volt. A tudományos kutatásban iskolaalapító, intenzív munkásságát 68 közlemény bizonyítja. Az MTA Akusztikai Komplex Bizottságban, az OPAKFI-ban, a Híradástechnikai Tudományos Egyesületben és az Audio Engineering Society-ben aktív szakmai-társadalmi tevékenységet fejtett ki. A számára életpályát jelentő akusztikus szakmát a fiatalok számára is vonzóvá igyekezett tenni. Munkásságát Petzval József-díj és Békésy-díj fémjelzi.

Az Alapítvány eddigi eredményei

Az alapítás óta eltelt hét évben a Kuratórium összesen 8 díjat adott ki. A díjazottak neve és pályamunkáik rövid ismertetője megtalálható az Alapítvány honlapján (http://www.huszty.org). Támogatóink sorába időközben belépett a Magyar Rádió és a magánszemélyek 1%-os adótámogatásán kívül néhány további vállalkozás is segíti munkánkat. Az Alapítvány mérlegei és közhasznú jelentései is megtekinthetőek a honlapon. A tevékenységünket meghatározó szabályok szerint a díjak összegét a mindenkori vagyon kamataiból lehet kifizetni, így támogatóink jóvoltából a díjak összege az elmúlt években sem csökkent.

A 2007. évi pályázat

A pályázati felhívást a Kuratórium a tárgyévet megelőző évben, szeptember során írta ki és tette közzé a HTE és az Opakfi lapjaiban, az Interneten és az oktatási intézmények hirdetési helyein. A pályázaton végzett, elsősorban mérnökök, fizikusok saját önálló munkájuk összefoglaló dolgozatával, szakirányú lapban megjelent cikkeikkel, vagy új dolgozataikkal, mint pályaművekkel vehettek részt. A pályázóknak lehet más diplomájuk is, de tevékenységüket az akusztika területén kell, hogy kifejtsék.

A 2007. évi pályázat kiemelt témaköre a következő volt: „Időszerű akusztikai problémák korszerű megoldásai és a tartalom-előállítás és reprodukció akusztikai vonatkozásai". A Kuratórium döntése alapján a Huszty Dénes emlékdíjat – mely emlékplakettből és 200 000 Ft-os pénzjutalomból állt – Mihajlik Péter nyerte el közeljövőben beadandó doktori disszertációjának eredményeivel, „Spontán magyar nyelvű beszéd gépi felismerése nyelvspecifikus szabályok nélkül" című dolgozatával. A díjkiosztó ünnepséget 2008. március 11-én, a HTE Stúdiótechnikai Szakosztálya és az OPAKFI Akusztikai Szakosztálya közös ülése keretében tartották meg, melyen a témaválasztás aktualitásáról Dr. Gordos Géza professzor úr tartott bevezető előadást. A pályázatot a Kuratórium Elnöke, Dr. Illényi András értékelte.

A 2008. évi pályázatok kiírásával kapcsolatban az első hirdetmények májusban várhatóak. Az Alapítvány működésére vonatkozó kérdésekkel kapcsolatban a HTE és az OPAKFI tikársága áll az érdeklődők rendelkezésére.

Irodalom

Kép és Hangtechnika XXV/5. szám (1979. október), Elektroakusztikai szám Huszty Dénes emlékére.

Hidden Markov-Modell based text-to-speech synthesis method applied to the Hungarian language
Keywords: speech synthesis, text-to-speech (TTS), Hidden Markov-Modell (HMM)

The Hidden Markov-Model synthesis has numerous favorable features: it can produce high quality speech from a small database, furthermore, theoretically it allows us to change characteristics and style of the speaker and emotion expression may be trained with the system as well.

Increasing the naturalness of text-to-speech synthesizers
Keywords: speech synthesis, prosodic model, prosodic variability, F_0 transplantation

This paper briefly introduces the prosody models used in current speech synthesis systems and one of their shortcomings: the lack of prosody variation. Our approach for decreasing monotony in extended synthesized passages is described in detail. The method uses a database of natural sample sentences. The solution is based on copying the prototype of the fundamental frequency curves. Finally, the evaluation of sentences produced by our method is described.

An analysis of extension possibilities of Hungarian limited domain corpus-based synthesis systems to unlimited vocabulary
Keywords: corpus-based speech synthesis, databases

The naturalness of corpus-based speech synthesizers can be very high in limited domains. In this paper, we discuss the possibilities of extending Hungarian limited domain synthesis to unlimited vocabulary. Enlargement the speech database in order to get optimal coverage, can not be realised for Hungarian for practival reasons. More than 100 hours of speech would be needed in the speech database. The analysis of speech quality was based on a listening test. The final conclusion was that combining the corpus-based technique with the prosody models of diphone-triphone concatenation synthesisers may result in better speech quality for unlimited synthesis.

Increasing the efficiency of research and development by better processing of speech databases
Keywords: speech database, correction of sound boundaries, labeling, corpus-based speech synthesis

Creating large spoken databases has become necessary in the last decades to support speech research and the development of speech recognition and text-to-speech systems as practical applications. These databases serve their applications well if their inner labeling (sound boundaries etc.) are correct. In this paper, the creation of accurate labels/markers is discussed for databases that contain many sentences from the same speaker. Such databases are used principally for speech synthesis. Labeling these speech databases needs software automation support, exclusively manual work is not feasible. The goal however is to place the labels/markers as precisely as possible. A new hybrid solution is presented that results in a practically error-free marker set in the database. The method uses phonetically based software algorithms and human correction, too.

Speech enhancement based on the reconstructed phase space
Keywords: speech enhancement, signal subspace, reconstructed phase space, dimension embedding

The speech enhancement method, presented in this paper, is based on the concepts of reconstructed phase-space and dimension embedding. The proposed algorithm separates the speech from noise using a non-linear transformation in a transformed domain. Our recent results in case of uncorrelated, additive noise are presented in this paper.

Visible speech in IPTV for deaf users
Keywords: face animation, speech-to-facial conversion, visible speech, Direct Show system

In Hungary the dubbing of English speaking films was absolute common from the very beginning of TV broadcasting and normal users need it even nowdays. For the growing community of deaf and hard of hearing people, such sound of films is not intelligible so they need additional information. A new method was elaborated to convert the speech signal into animation of a speaking face for lip reading at the corner of the TV screen. Our system operates real time and can be applied to traditional analogue TV signals, to DVD or IPTV. It has no language specific elements so can be adapted for any language. The Windows Direct Show system was used for the implementation.

Computer-based room acoustics simulation for sound field optimalization
Keywords: room acoustics, CARA, CAD, reverberation, sound field optimalization

Room design, reverberation time calculations and sound field optimalization are the most important tasks in room acoustics. Using simple geometric calculations and some restrictions these tasks are optimal for computer design (CAD applications). The Computer Aided Room Acoustics software is able to handle various types of rooms and textures, 3D objects, it calculates reverberation time, sound pressure levels and even optimizes room layouts, loudspeaker and listener positions. Our demonstration is based on the newly designed and reconstructed D1 lecture room at the Széchenyi István University. Furthermore, some results are shown for an optimized 5.1 home theater system in a normal living room.

Using prosody for the improvement of automatic speech recognition
Keywords: speech recognition, Hidden Markov-Modell, prosody, word boundary detection,

This article presents sentence, phrase and word boundary detection based on prosodic features, implemented in a Hidden Markov Model-based prosodic segmentation tool. Integrated into a speech recognizer, an N-best rescoring is performed based on the output of the prosodic segmenter, which determines the prosodic structure of the utterance. In an ultrasonography task, we obtained 3,82% speech recognition error reduction using a simplified bigram language model.

Turk 2 – Multimodal chess player
Keywords: multimodal human-computer interaction, facial gesture recognition, face detection

The use of multimodal human-computer communication based technique is a new approach to interact with information systems. In this paper the realization of a chess player machine based on these technique is presented. The multimodal essence of human-computer communication means that you can play chess without keyboard, using only the channels of the verbal and gesture language.

Summaries • of the papers published in this issue

Contents

Szerkesztőség
HTE Budapest V., Kossuth L. tér 6-8.
Tel.: 353-1027, Fax: 353-0451, e-mail: info@hte.hu

Hirdetési árak
Belív 1/1 (205x290 mm) FF, 120.000 Ft + áfa
Borító II-III (205x290mm) 4C, 180.000 Ft + áfa
Borító IV (205x290mm) 4C, 240.000 Ft + áfa

Cikkek eljuttathatók az alábbi címre is
Szabó A. Csaba, BME Híradástechnikai Tanszék
Tel.: 463-3261, Fax: 463-3263
e-mail: szabo@hit.bme.hu

Előfizetés
HTE Budapest V., Kossuth L. tér 6-8.
Tel.: 353-1027, Fax: 353-0451
e-mail: info@hte.hu

2008-as előfizetési díjak
Közületi előfizetők részére: bruttó 32.130 Ft/év
Hazai egyéni előfizetők részére: bruttó 7.140 Ft/év
HTE egyéni tagok részére: bruttó 3.570 Ft/év

Subscription rates for foreign subscribers:
12 issues 150 USD,
single copies 15 USD

www.hte.hu
Felelős kiadó: NAGY PÉTER
Lapmenedzser: DANKÓ ANDRÁS

HU ISSN 0018-2028

Layout: MATT DTP Bt. • Printed by: Regiszter Kft.