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.

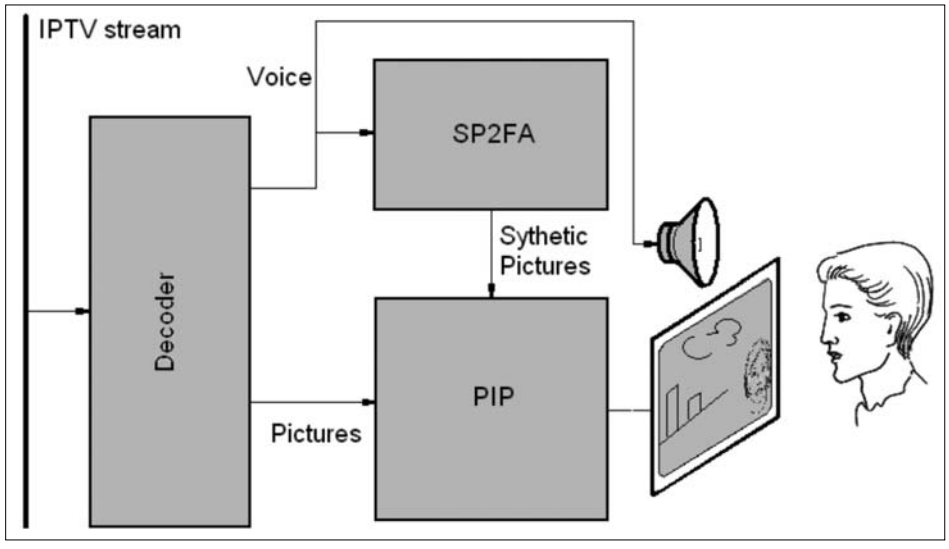
A (beszéd)hanggal vezérelt animált fej vagy száj ígéretes megoldás, mert:

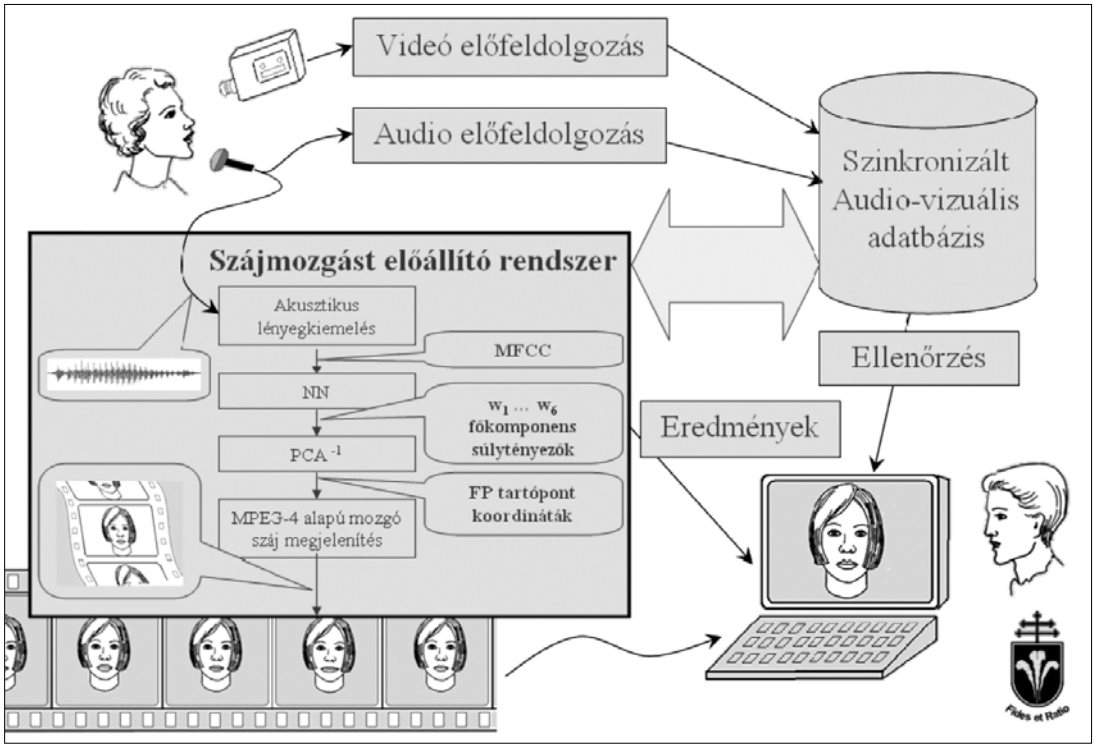
– a siketek deklarált fő igényeihez illeszkedik,
– nyelvfüggetlen megoldást sikerült megvalósítani alapszinten,
– teljesen automatizálható, szemben a feliratot készítő megoldással, tehát hatékony,
– ugyanaz a műszaki megoldás alkalmazható különböző technológiával továbbított vagy tárolt műsorok esetén 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





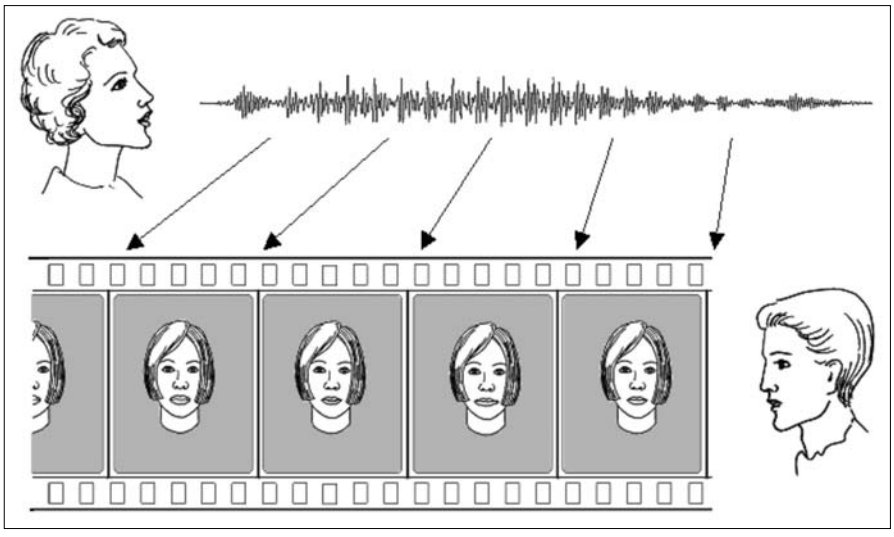
3. ábra
Az SP2FA rendszer felépítése

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

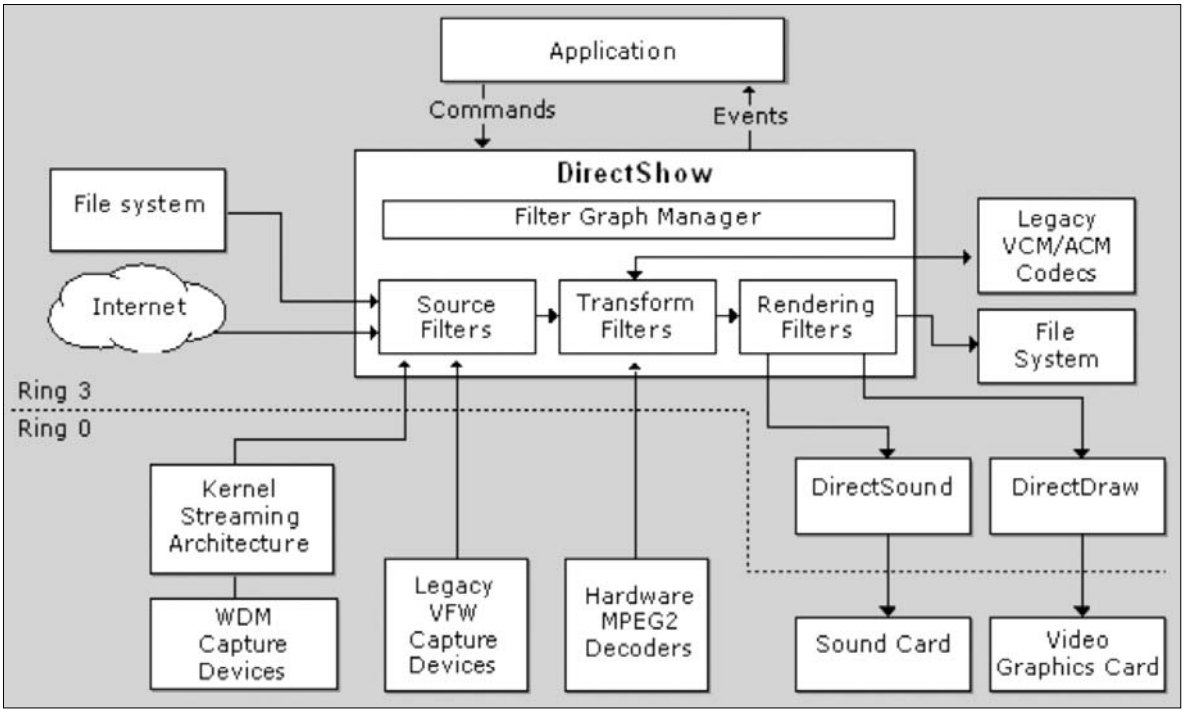
A DirectShow egy olyan környezet, amelyet a Windows operációs rendszer médiakezeléséhez fejlesztett a Microsoft. A rendszer a médiafeldolgozásban már jól ismert, hálózatba szervezhető alapvető funkciókra épül. Az alapvető feldolgozóegységeket jól megfogalmazott input-output rendszerbe helyezték, és lehetővé 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

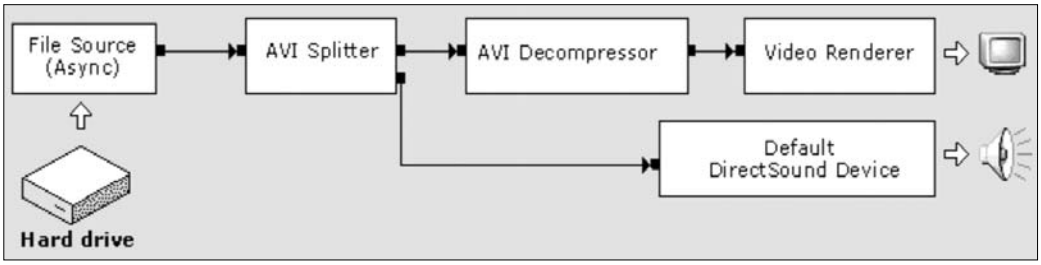
A kísérleti rendszer megvalósult. Szokványos PC erőforrásokon valós időben működik. Szélesebb körű tesztelésére és hosszabb idejű próbájára szükség lenne annak érdekében, hogy a célfelhasználók megtanulják, megszokják.

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