

Audiovizuális beszéd-adatbázis és alkalmazásai

BÁRDI TAMÁS, FELDHOFFER GERGELY, HARCZOS TAMÁS,
SRANCSIK BÁLINT, SZABÓ GÁBOR DÁNIEL

*Pázmány Péter Katolikus Egyetem, Információs Technológia Kar
{bardi, flugi, harczos, sraba, szasza}@digitus.itk.ppke.hu*

Kulcsszavak: beszédszintézis, hallássérültek, beszélőfej-animáció

Cikkünkben egy audiovizuális beszéd-adatbázis összeállításáról számolunk be, melyet hallássérült emberek számára készülő szájmozgás-animációs programunk fejlesztésében szeretnénk hasznosítani. Részletesen kitérünk a tervezési szempontokra, és arra a siketeket érintő vizsgálsorozatra, melyen alapulnak.

1. Bevezetés

Az utóbbi évtizedben a multimédia kommunikációs technológiai fejlődése következtében az emberi érzékelés kép és hang modalitásainak összefüggései, egymásra hatásai a tudományos érdeklődés homlokterébe kerültek, és megélnékültek a kutatások az audiovizuális beszédfeldolgozás területén.

A beszédhang, a kép és az elmondott szöveg hármán belül lehetséges konverziók közül két irány élvez nagyobb népszerűséget a kutatók között. Az egyik az audiovizuális beszédszintézis, ahol egy animált fej szájmozgásait és esetleg az arc mimikai mozgásait, valamint az ezzel szinkronban lévő beszédhangot képezük a szöveg alapján. A másik az audiovizuális beszédfelismerés, ahol a hagyományos beszédfelismerőhöz képest a hang mellett a képi információt is felhasználjuk a biztosabb felismerés érdekében [1].

Mi az elmúlt évben azonban egy harmadik iránnyal kezdtünk foglalkozni, mely a felismerésnek és a szintézisnek egy sajátos keveréke, a beszédhangból beszélőfej animáció képzése –, angolul: speech to animation conversion.

A felmerülő első kérdés: mire használható egy ilyen? Egyrészt a szórakoztató iparban rajzfilmek, animációs filmek, valamint játékszoftverek készítését teszi gyorsabbá és olcsóbbá, mintha képkockaként „kézzel” kellene őket megrajzolni. Másrészt hallássérült emberek hátrányos társadalmi helyzetének kiküszöbölésében is alkalmazható. A siketek és nagyothallók többsége jól tud szájról olvasni, így egy megfelelő minőségű szájmozgás-animáció „láthatóvá” tenné számukra a beszédet.

Ilyen, „megfelelő” minőségű speech to animation conversion alkalmazás azonban működő rendszerben ma még nem létezik, de reméljük, hogy jelenleg is folyó munkánk eredményeképpen megoldást fogunk adni a problémára.

Jelen cikkünkben beszámolunk az ezirányú kutatásokról: röviden a fontosabb nemzetközi eredményekről, valamint saját erőfeszítéseinkről és a jövőbeni törekvéseinkről.

2. Helyzetkép

2.1. WISDOM

A WISDOM (*Wireless Information Services for Deaf People*) egy EU kutatás-fejlesztési projekt, amely 2001-ben indult és 2003 végén fejeződött be. A projektben az Egyesült Királyság, Svédország, Spanyolország és Németország vettek részt. Kutatólaboratóriumok, mobil szolgáltatók, rendszerszállítók, egyetemek fogtak össze. A célközösség természetesen a siketeket jelentette az adott országokban.

Szándékuk az volt, hogy a kutatási eredményeikre alapozva olyan szolgáltatásokat lehessen nyújtani, melyek megoldást adnak személyek közötti élő beszélgetésnél, és információ lekérdezésénél nyilvántartó rendszerekből. Mindezeket egy hordozható, kép átvitelére is alkalmas berendezésre és a siketek jelnyelvére alapozták.

A WISDOM projekt kifejezetten a harmadik generációs mobil rendszerekre épül, mert ezekkel nyílt meg a mobil mozgókép-továbbítás lehetősége. Bluetooth képességekkel rendelkező eszközöket kapcsoltak össze videojel-felvétel, tömörítés, szövegkezelés, és jelnyelvi felismerés céljaira.

2.2. SYNFACE

A SYNFACE a stockholmi Királyi Műszaki Egyetem (KTH) projektje (2001-2003) [2,3]. A projektben fix telefontkészülékekhez köthető, valós idejű, laptopon futó rendszert állítottak össze. (Alapelveiben jelen cikk szerzői is egy, a fentihez hasonló rendszer kialakításában gondolkodnak.)

A SYNFACE projektben olyan moduláris megközelítést választottak, amelyben a feladatot gyakorlatilag kettévágták egy felismerő és egy szintézis részre. Rendelkezésükre állt egy 3D-s animált beszélőfej, amit TalkingHead néven fejlesztettek egy korábbi projektjükben [4].

A TalkingHead bemenetként fonéma kódokból álló sztringet vár és ehhez képezi a szájmozgás animációt. (Csak a modell szája mozog, az arc többi része, a szemkörnyék és a homlok merev. Az élő emberi arcon itt tör-

ténnek a hangsúlyozáshoz és érzelem kifejezésekhez tartozó finom mimikai mozgások.) Ehhez készítettek egy rejtett Markov-modelles (HMM) beszédfelismerőt, amely a beszédjelből előállítja a TalkingHead számára a fonéma sorozatot [5]. A felismerőjük nem szavakat ismer fel, hanem csak fonémákat. Így nagyobb a tévesztési arány, viszont a késleltetés kevesebb, mert nem szó hanem csak fonéma hosszúságú.

A beszéd fonémáknak megfelelő vizuális alapegységei a vízémák. Ezekből kevesebb van, mint a fonémákból, mert ha két fonéma képzése csak olyan jellemzőben különbözik, ami kívülről nem látszik, akkor ugyanaz a vízéma tartozik hozzájuk. Ilyen jellemzők például, hogy a hangszalagok rezegnek vagy sem, vagy a nyelvcsap enged-e levegőt az orrüregen át vagy sem. Ennek megfelelően a 'papa', 'baba', 'mama' szavak szájáról olvasva elvileg megkülönböztethetetlenek. (Vannak igen gyakorlott siketek, akik esküsznek rá, hogy ilyenkor is látnak különbséget.)

A projektben próbálkoztak fonéma felismerő helyett vízéma felismerővel, ami elvileg egy könnyebb osztályozási feladat, hiszen kevesebb osztályba kell sorolni az akusztikus kulcsokat. De a tévesztési arány így nagyobb volt, vélhetően azért, mert vannak akusztikusan nagyon hasonló beszédhangok, amik különböző vízémákhoz tartoznak, és vannak amelyek akusztikusan nagyon különbözőek, de ugyanahhoz a vízémához tartoznak.

A program a sok munka és a részletek színvonalas kidolgozottsága ellenére a teljesen siketek esetében nem bizonyult használhatónak. Viszont a nagyothalló alanyok esetében sikeres volt: a telefonon érkező beszédet jobban értették, ha láthatták hozzá az animált szájmozgást a laptop képernyőjén.

3. A választott megközelítés

A kutatási projektünk eredeti célja segédeszközök fejlesztése siketek mobiltelefonos kommunikációjának elősegítésére. A segédeszközök között szerepelnek a telefonhoz illesztett kiegészítő eszközök, vagy a mobiltelefonon futtatott olyan alkalmazások, amelyek megkönnyítik, vagy lehetővé teszik siket emberek hallókkal folytatott mindennapi beszélgetését.

A segédeszközök fejlesztése érdekében számos előzetes vizsgálatot végeztünk siket emberek kommunikációs képességét és szokásait illetően. Azokban a helyzetekben, amikor a siketek nem láthatják a beszélő személyek száját, arcát (például telefonos kapcsolatnál), akkor a beszédjelből olyan mozgóképek ábrázolását választottuk, melyek a megértéshez elegendő képi információt hordoznak a beszédjelről.

A spektrogramok „olvasását” is meg lehet tanulni kitartó gyakorlással. A szájról olvasást viszont nem kell tanulniuk és gyakorolniuk a siketeknek, mert sokéves gyakorlattal szinte művészetté tökéletesítették ezt a képességüket. Ezért gondoltunk olyan segédeszközben, ami ezt ki tudja használni. Úgy látjuk, hogy nem egy-

szerű, nem megoldott, de reálisan kitűzhető feladat a beszédjelből a szájmozgás képeinek gépi előállítását.

Ezzel szemben egy beszédhangból jelnyelvi animációt képző program készítése ma még reménytelen feladatnak látszik, mivel egy ilyen programnak meg kellene „érteni” a beszélő mondanivalóját. A jeltolmácsok emberi intelligenciája, amit beszédre jelre vagy jelről beszédre fordításkor aktívan használnak, kérdéses hogy gépi módszerekkel kiváltható-e egyáltalán.

3.1. Előzetes vizsgálatok

Nem szeretnénk volna olyan eszköz fejlesztésébe kezdeni, amit a siket és nagyothalló emberek nem akarnak, vagy nem tudnak használni. Ezért előzetes „igényfelmérés” gyanánt interjúkat készítettünk, valamint a hallókétól eltérő képességeiket kísérletekben mértük. Ezekben a felmérésekben összesen 14 siket és erősen nagyothalló személy vett részt. Összehasonlításként mindegyiket elvégeztettük halló alanyokkal is.

Először interjúkat készítettünk velük, ahol konkrét történeteket meséltek el, melyekben a halló és siket emberek közötti párbeszéd valamiért elakadt, sikertelen volt, vagy félreértés történt. Ezután a hallókétól eltérő nyelvi képességeiket, szövegértésüket írásbeli feladatlapokkal vizsgáltuk.

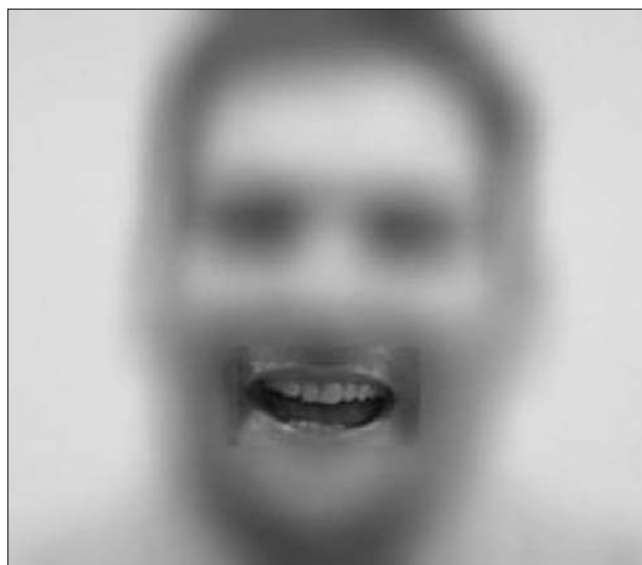
Szájáról olvasási képességüket különböző videó tesztekkel vizsgáltuk. A vetített felvételeken egy-egy ember beszélt szemből, közlőrlől a kamerába nézve, és a szájáról kellett leolvasni, hogy mit mond. A hangot minden esetben kikapcsoltuk.

A felvételek különböző minőségűek és felbontásúak voltak. A frame-rate 15, 25 és 30 kép volt másodpercenként. Jó minőségű kivetítőként projektort használtunk. A gyengébb minőségű kivetítő eszközt a Sony Ericsson P910-es mobiltelefon 208x320 pixeles színes kijelzője jelentette (40x61 mm).

A videókat különböző módon manipuláltuk:

- „blur”: a száj és közvetlen környékén kívül a képet elhomályosítottuk (1. ábra),

1. ábra „Blur” fedí az arc nagy részét





2. ábra Kékített videó

- kékítés: az RGB színek komponenseket eltoltuk a kék felé (2. ábra).
- binarizálás: a képet a pixelek világossága alapján kétszínűvé, fekete-fehérré konvertáltuk. Egy világosság-küszöböt beállítva fekete és fehér pixeleket kaptunk, a köztes szürke értékeket nem engedjük meg (3. ábra).



3. ábra Binarizált videó

Összehasonlításképpen a kékített és a kétszínű videóok mellett vetítettünk felvételeket az eredeti színekkel is. Ezeket a formák térbelisége jobban kivehető (4. ábra).

4. ábra Videó az eredeti színekkel



3.2. A vizsgálatok tanulságai

Az elvégzett vizsgálatok alapján a fejlesztési folyamatot is befolyásoló tanulságokat szűrtünk le.

Ne legyen feltűnő

A legtöbb hallássérült nem szívesen hívja fel a figyelmet erre a tulajdonságára, mindaddig, amíg ez nem feltétlenül szükséges. A fehér botot használó vakokkal, vagy a kerekesszékekkel közlekedő mozgássérültekkel szemben a siketség (nagyothallás) első ránézésre egyáltalán nem látszik, és ők törekednek is rá, hogy amíg csak lehet, olyanak tűnjenek, mint a halló többség. Emiatt sokuk tartózkodik olyan technikai eszközök használatától, ami már messziről „elárulja” hallássérülésüket. Vannak akik még a hallókészüléket is szégyellik hordani, bár ma már ez alig látszik.

Ebből a szempontból egy mobiltelefonon futó program mint segédeszköz úgyszólván tökéletes. Mobilja szinte mindenkinek van, használója általában nem kelt feltűnést. Siketek is előszeretettel használják, jellemzően az üzenetküldő szolgáltatásokat. Emellett egy rendkívül kompakt eszköz, vihető bárhová és hosszabb ideig bírja akkumulátor feltöltés nélkül, mint egy laptop.

A szájról olvasás vizuális feltételei

A szájról olvasás még a leggyakorlottabb siketek számára is nehéz és fárasztó feladat, ami koncentrált figyelmet igényel. A beszélőnek és a szájról olvasónak egyenesen egymásra kell néznie közben, a fejüket nem forgathatják, és ügyelniük kell a megfelelő távolságra.

A beszéd szájról olvashatóságának két legfontosabb kritériuma a beszéd sebessége és artikuláltsága. Lassan és tisztán, látványosan jól artikuláltan kell beszélni. A hallók számára megszokott beszédtempó a szájról olvasáshoz túl gyors, nem érthető, ilyenkor gyakran előfordul, hogy a siket nem érti amit mondanak neki, csak bólogat. A halló emberek beszéd artikulációja nagyon változó, szájról olvashatóság szempontjából igen erős szórást mutat. Vannak, akik „lapos szájjal”, nagyon alulartikuláltan beszélnek, és a siketek akkor sem értik őket, ha külön a kedvükért lelassítanak, sőt még ismétlés után sem.

A legérthetőbben artikulálók jellemzően azok a siketek közt élő vagy dolgozó emberek (halló rokonok, jeltolemcsok), akiknek a kommunikációs és empatikus készségük egy folyamatos visszacsatolást biztosít, aminek köszönhetően javítani tudják a beszédjük szájról olvashatóságát. Ők egy idő után ösztönösen kihangsúlyozzák a szómozgásukban azokat az apró elemeket, melyek alapján a hasonló vizémák megkülönböztethetők. Jól választják meg a tempót, és a nehezebben érthető részeket az arckifejezésükkel hangsúlyozzák, és ott még lassabbra váltanak.

Azt tapasztaltuk, hogy ezek az artikulációs jellemzők sokkal többet számítanak, mint a kép felbontása vagy a frame-rate. Mindezek alapján arra a következtetésre jutottunk, hogy az aaudiovizuális beszéd-adatbázisunk felvételéhez jeltolemcsokat alkalmazunk. Így remélhetőleg olyan tanító adatokat kapunk, mely alapján az animált beszélőfejünk szómozgásában az említett artikulációs jellemzők érvényesülnek majd.

A szájról olvasás nyelvi feltételei

Szájról olvasáskor fontos szerepe van a nyelvi intelligenciának. Beszélgetéskor a halló emberek sem értenek tisztán minden hangot az elhangzott szavakból, hanem a hiányzó részeket a nyelvi tudásukra és a társalgás aktuális kontextusára alapozva kiegészítik agyuk. Az egyes beszédhangok azonosítása a szájmozgás alapján bizonytalanabb, mint hallás alapján. Emiatt szájról olvasáskor több a hiányzó vagy bizonytalan elem, amit a kontextus és a nyelv alapján kell kirakni fejben, mint a hallók beszédértésében.

A siketek nyelvhasználatukban aránylag kevés figyelmet fordítanak a ragozásra. Magyarban a ragozott szóalakok jellemzően csak egy-két nehezen leolvasható, könnyen összetéveszthető mássalhangzóban különböznek. Erre mindig tekintettel kell lenni a megfogalmazásban. Például annál a mondatnál, hogy *'Haza megyek'* jobban értik azt, hogy *'Én haza megyek'*.

Mindezek fényében teljesen irreális volna azt várni, hogy tetszőleges témájú, nyelvileg bonyolult mondatokat szájról olvasás után szó szerint visszaadjanak. Ehelyett a témát előzetesen megadtuk, és csak egy-két szóval kellett kérdésekre válaszolniuk, amivel ellenőrizhettük, hogy mit értettek meg. A vizsgálatok egy részében tudhatták, hogy kétjegyű számokat kell leolvasniuk, ezzel a „keresési tartományt” jelentősen leszűkítettük.

A mobilkészülék kijelzője elegendő

Úgy találtuk, hogy a Sony Ericsson P910-es készülék kijelzőjének elég jó a felbontása a szájról olvasáshoz. A jól artikulált felvételek esetén a mobiltelefon kijelzőjéről és a jó minőségű projektor képéről egyaránt 80-90%-os felismerési aránnyal olvasták le a kétjegyű számokat. A felismerés a kéékített és a binarizált felvételeken is ugyanilyen megbízható volt.

2D vagy 3D

Az emberi agy több módszert is bevet a látott kép térbeli rekonstrukciójában. Egy arc nézésekor leginkább a gömbölyded felületek folytonos színáttnései és a finom fényárnyék hatások játszanak szerepet. A kéékített és a binarizált felvételeken viszont ezek jelentősen torzultak. Úgy gondoltuk, ha ezekről tudnak szájról olvasni, akkor elég 2 dimenziós fejmodellben gondolkozni. A felmérések eredményei megerősítették várakozásainkat. Ennek alapján 2D-s audiovizuális adatbázis felvétele mellett döntöttünk. Így a több kamera képéből számított 3D-s mozgás követés technikai problémáival nem kell törődnünk.

4. Adatbázis-összeállítás és -felvétel

Egy rendszer tanításához és ellenőrzéséhez olyan adatbázisra van szükség, ami megfelel a teljességi, a pontossági, és a statisztikai elvárásoknak. Az adatok felvétele előtt mondatokat és diádokat állítottunk össze úgy, hogy a lehetséges fonéma együttállásokat a felvétel lefedje, ehhez a lehetséges diádok felsorolását választottuk. A példamondatok a magyar fonémakészletet a nyelv átlagos eloszlása szerint használták.

Ezek a követelmények beszédfelismerési feladatokban már ismertek és jól dokumentáltak, az újdonságot a videó anyag rögzítése jelentette. Olyan videó anyagra van szükség, ami pontos és használható információt tartalmaz. Ha egyszerű, jól megvilágított arcot fényképezünk le, akkor abból nehéz absztrakt információt kivonni, ezért kitüntetett pontok követését tűztük ki célul.

Az arc teljes körű modellezését már elvégezték és szabványosították MPEG-4 néven. Ez a szabvány leír többek között 84 feature pointot (FP) az arcon. Úgy döntöttünk, hogy amennyire lehetséges, e szabvány szerint dolgozunk. A képeket úgy vettük fel, hogy az arcon megjelöltük néhány FP-t. Ez a jelölés bőrszemleges textílfestékkel történt, ami könnyen lemosható, nem irritálja a bőrt, és a kiszérelésének köszönhetően kényelmesen felvihető 2-3 mm vastag foltokban. Ezzel az eljárással az a probléma, hogy az ajkak belső oldalára nem lehet pontot tenni, mert az elkenődne beszéd közben, így csak bizonyos pontokat vettünk fel. Nem jelöltük meg a siketek által nem figyelt pontokat, csak a száj környékén szemből látható területen törekedtünk teljességre (5. és 6. ábra).

A megvilágítást úgy választottuk meg, hogy a későbbiekben automatikusan tudjuk a pontokat megkeresni a felvételen. Mivel a beszéd közben ezen pontok mozgását kell megállapítani, fontos, hogy a felvétel felbontása minél jobb legyen. A kameráink a piacon elterjedt digitális kamerákhoz hasonlóan 720x576 képpont felbontásúak. Két kamerával dolgoztunk, amik egyszerre indultak el. Az egyik kamera a teljes arcot fényképezte, 90 fokos elforgatással, hogy a vízszintes képet jobban kihasználjuk, a másik kamera csak a száj környékét, hogy a számunkra legfontosabb FP-k mozgását minél pontosabban meghatározhatjuk. A két kamera és a hang szinkronitását ellenőriztük, az esetleges elcsúszások nem haladják meg az egy frame-t. Azzal, hogy az egyik kamera csak a száj körüli pontokat figyelte, elértük, hogy egy FP mozgását átlagosan vízszintesen 40-60 pixel, függőlegesen 80-140 pixel felbontással lehessen rögzíteni.

A kamerák képét közvetlenül videószalagra készítettük, majd a felvételek után DV AVI formátumban digitalizáltuk. A pontosság érdekében ennél jobban nem tömörítettük. Összesen négy jeltolmács beszélt a felvételeken, 76 perccel. A teljes kép- és hanganyag 36 GB-ot foglal.

5. Az MPEG-4 FAP pontok kinyerése a videó felvételekből

A videó felvételen előszűrést végeztünk. A megvilágítástól függően olyan intenzitásintéket állítottunk be, amik kiemelik a megfestett pontokat. Egy vágás után a pontok egy-egy összefüggő foltot jelentenek, amin eróziót hajtunk végre, hogy megkapjuk a foltok közepét pixel pontossággal. Ez az erózió akkor pontos, ha a folt konvex, ami sokféleképpen sérülhet. A legegyszerűbb

eset a gyors mozgásnál fellépő pontatlan interlaced kép, ami egy folt helyett vízszintes vonalakat ad. Ennek elkerülése érdekében elmostuk a képet, mielőtt a vágás megtörtént volna. Így már csak ritkán volt probléma a pontok követésével, ez pedig abból ered, hogy bizonyos szájtartásoknál az ajak lebiggyed, és az általunk festett pontot eltakarja, vagy árnyékot vet rá. A számításokat MATLAB programmal végeztük, mert a videó fájlokhoz való hozzáférést ez nagyon könnyűvé tette.

6. Továbbtekintés

A továbbiakban az adatbázis képanyagából rendelkezésünkre álló pont koordinátákat és a beszédhangból nyert sajátság vektorokat (MFCC) neurális hálók tanításához fogjuk felhasználni. Az így betanított neurális hálók fogják majd a beszédhez tartozó animációs paramétereket automatikusan előállítani. Az adatbázis feldolgozása és a tanítási folyamat erősen számítás igényes, ehhez a ma kapható csúcskategóriás PC-ket kell használnunk. De a létrejövő rendszer várhatóan egy jobb mobiltelefonon is futtatható lesz. [6] Reméljük, hogy olyan minőségű szájmogás animációt sikerül ezzel elérni, amelyet siketek az élő arcot közelítő megbízhatósággal le tudnak olvasni. Ilyen tudomásunk szerint a világon ma még nincs.

Köszönetnyilvánítás

A szerzők hálás köszönetüket szeretnék kifejezni az NKTH-nak a projekt támogatásáért, a közreműködésért a SINOSZ-nak, valamint a siketeknek és jeltolmácsok-

nak, akik részt vettek a programban. A T-Mobile-nak köszönjük, hogy a telefon készülékeket a rendelkezésünkre bocsátotta. Külön köszönjük Dr. Takács Györgynek az értékes segítségét és a témavezetést.

Irodalom

- [1] Czap L.:
Audiovizuális beszéd felismerés és szintézis,
PhD értekezés, 2005.
- [2] B. Granström, I. Karlsson, K-E Spens:
„SYNFACE – a project presentation”
Proc. of Fonetik 2002, TMH-QPSR, 44: 93–96.
- [3] M. Sheard, N. Thomas:
„The SYNFACE project: development and evaluation
of a talking face telephone”
The HCI 2003 –
Designing for Society Conference in Bath, England
- [4] J. Beskow:
Talking Heads – Models and Applications for
Multimodal Speech Synthesis. Doctoral Thesis.
Department of Speech, Music and Hearing,
KTH, Stockholm (2003)
- [5] G. Salvi:
„Truncation error and dynamics in very low latency
phonetic recognition”
Proc. of ISCA workshop on
Non-linear Speech Processing (2003)
- [6] Feldhoffer G., Bárdi T., Jung G., Hegedűs I. M.:
„Mobiltelefon alkalmazások siket felhasználóknak”,
Híradástechnika, 2005.

5. és 6. ábra

Megfestett referencia pontok az arcon, felvétel közben és a referencia pontok kinyerése a felvételtől

