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].

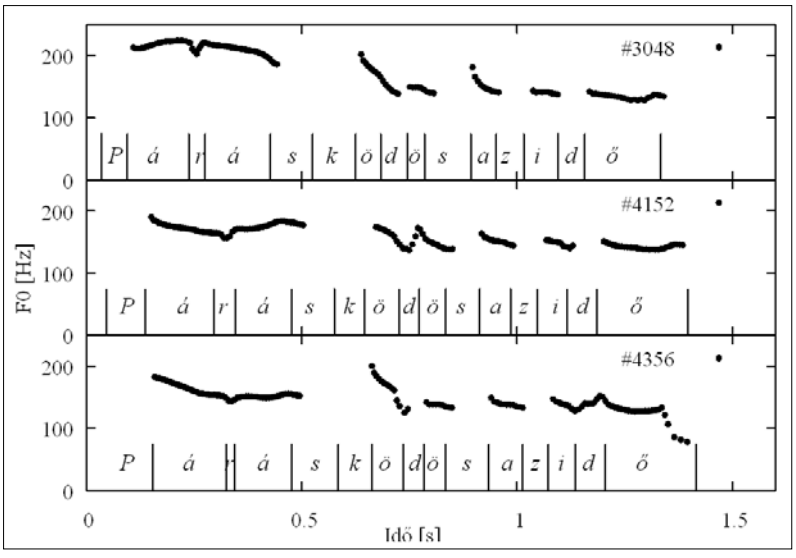
Cikkünk fő témája a szövegfelolvasó rendszerek egyik legfontosabb komponensének, a prozódia előállításának vizsgálata. A prozódia tervezésére sokféle modell ismert, úgymint a leíró jellegű, szabályalapú, gépitanulás-alapú, illetve szuperpozíciós modellek. A Profivox beszédszintetizátor első változata 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.

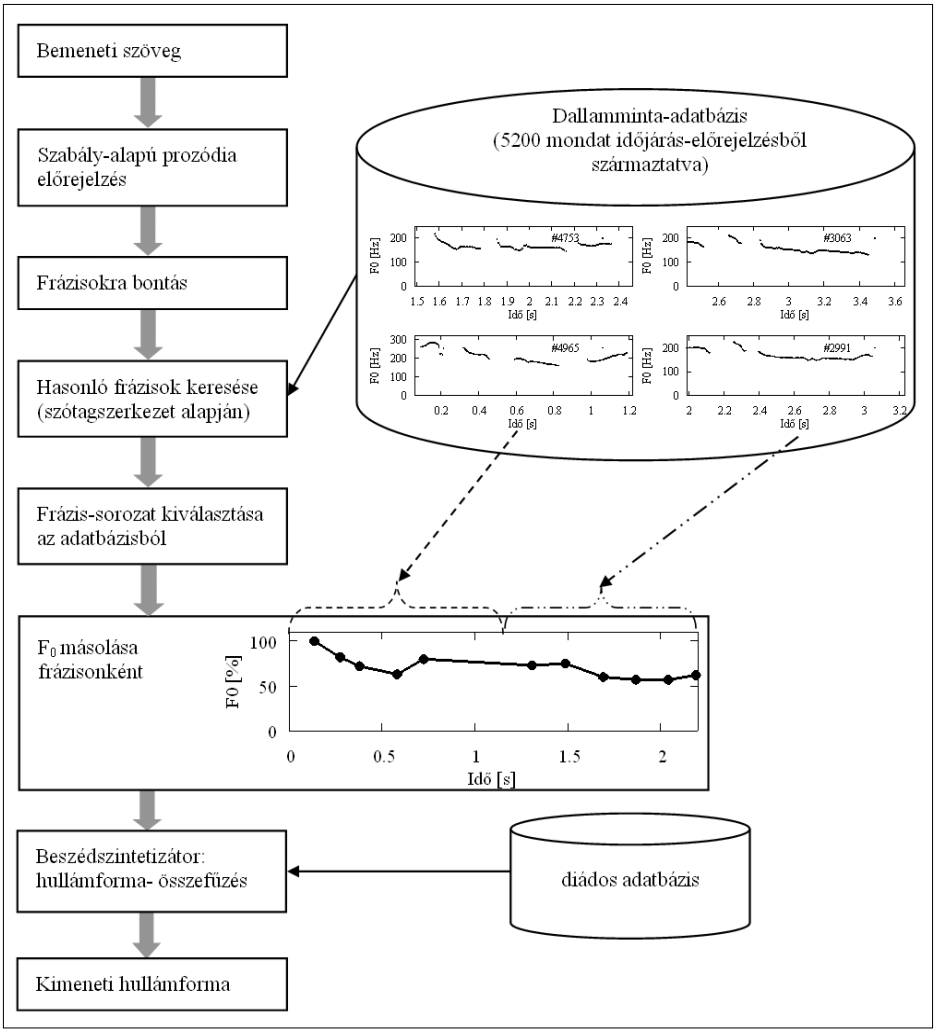
3. Dallammásolás frázisok alapján

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.

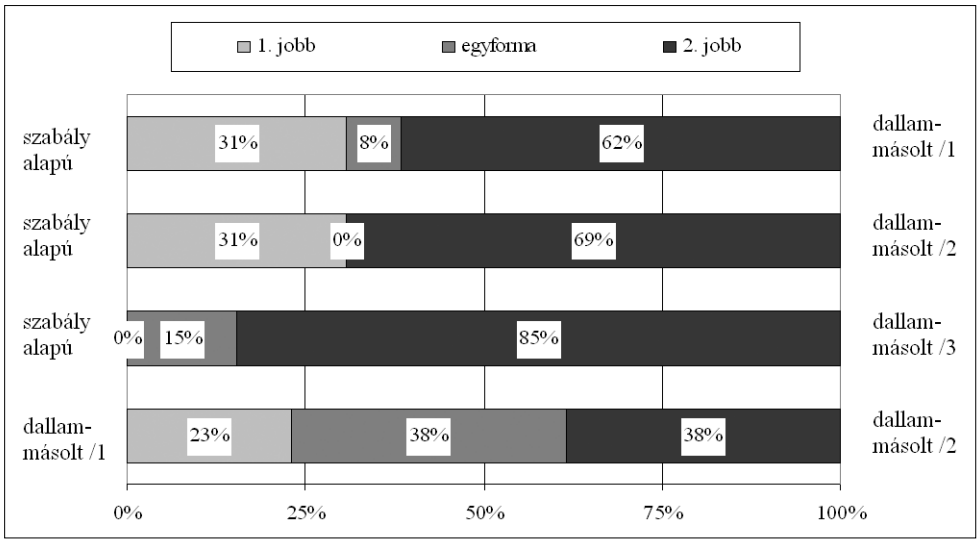
A tesztelőknek a *http://speechlab.tmit.bme.hu/csapo* oldalt meglátogatva egy rövid ismertetőt kellett elolvasniuk a teszt menetéről, majd néhány információt kértünk be róluk (becenév, életkor, nem). Ezután megkezdődhetett a mondatpárok meghallgatása. A szintetizált hangok meghallgatása után a tesztelők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

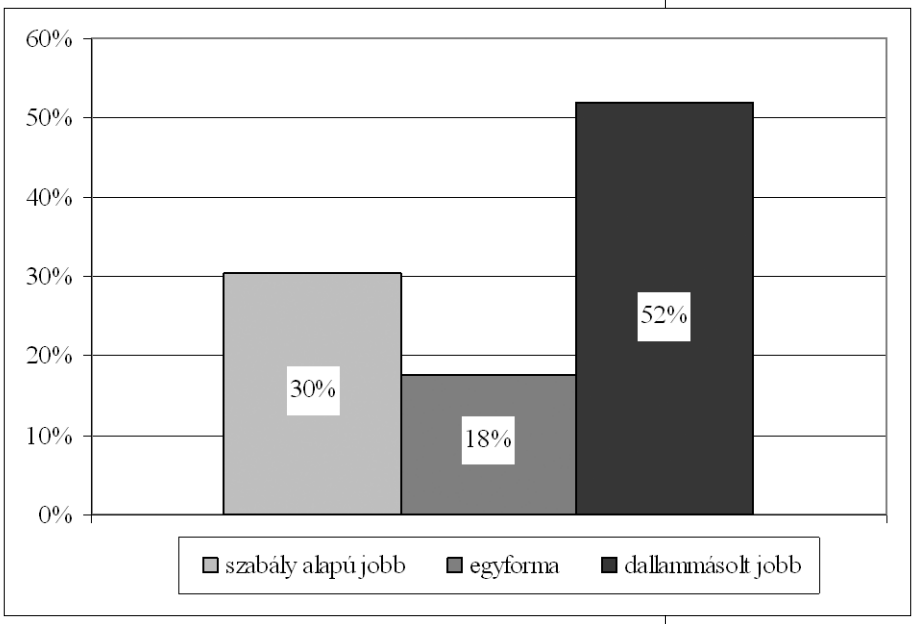
A *3. ábrán* egy mondat négy különböző változatának (egy szabályalapú és három dallammásolt) összehasonlítását láthatjuk, soronként egy mondatpár eredményeit ábrázolva. Észrevehető, hogy a dallammásolt változatokat a tesztelők természetesebbnek érezték,

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!