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.

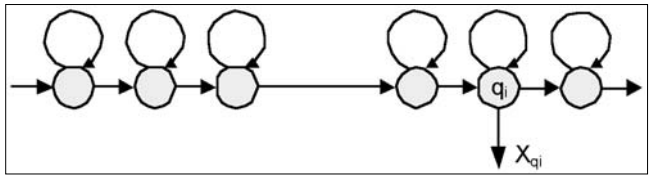
A HMM alapú beszédszintézis számos előnye miatt lett az elmúlt évek egyik legsikeresebb mesterséges beszédkeltési technológiája: kis (1,5-2 Mbyte) adatbázis méret mellett képes jó minőségű, érthető beszédet előállítani, amely hordozza a beszélő hangszínezeti tulajdonságait is. Továbbá lehetőség van több 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:

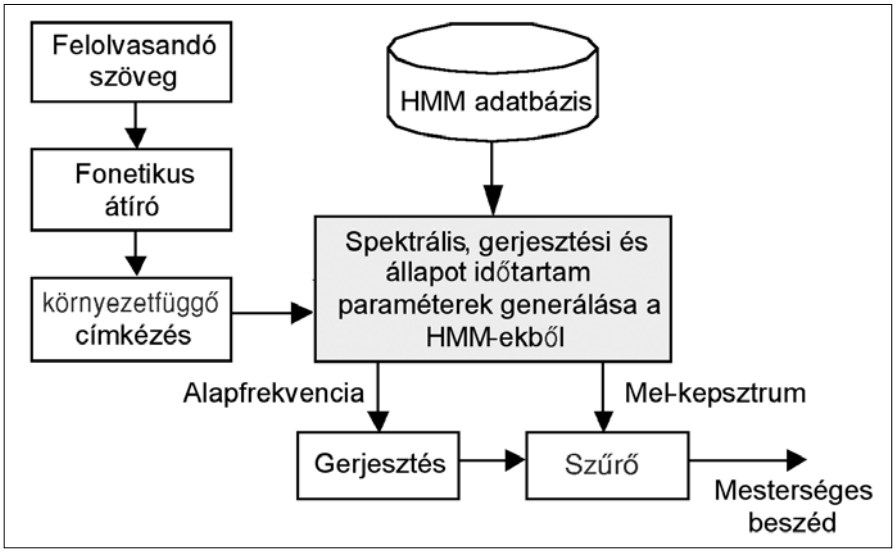
$$\underline{x} = ((x_{q_t})^T, (\Delta x_{q_t})^T, (\Delta^2 x_{q_t})^T)$$

3. A HMM alapú beszédszintézis

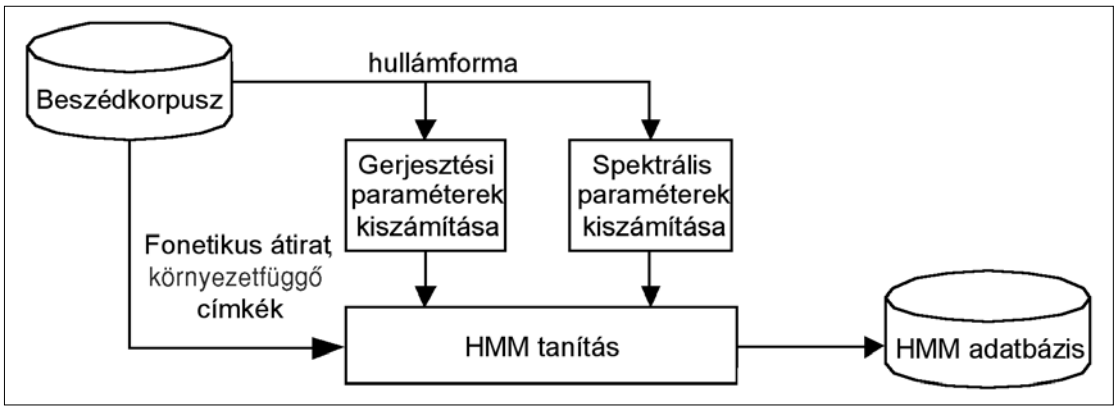
A HMM alapú beszédszintézist két részre oszthatjuk: a tanulás fázisára *(2. ábra)*, melynek során a HMM-eket tanítjuk be a beszédkorpuszunk alapján, illetve a beszédelőállítás fázisára *(3. ábra)*, amikor a betanított HMM-ekből kinyerjük a spektrális paramétereket, az időtartamokat és az alapfrekvenciát.

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

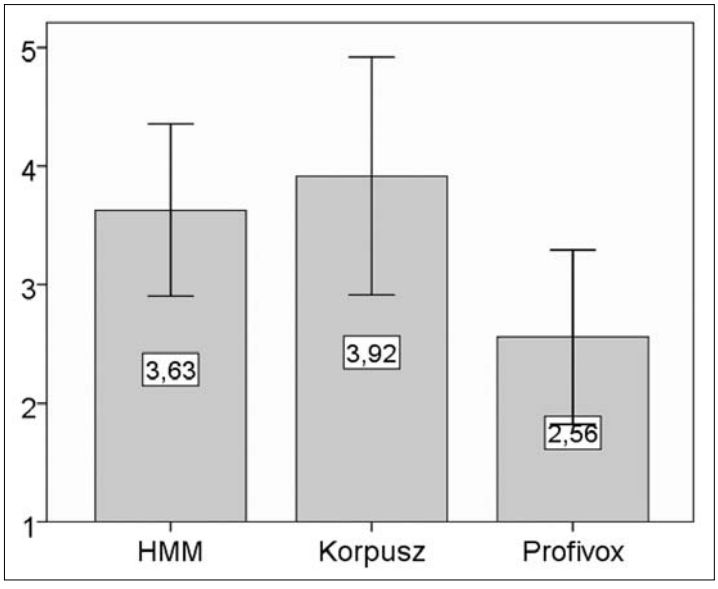
A teszt elején minden rendszertől 3-3 mondatot játszottunk le véletlenszerűen, amelyeket a tesztalanyok még nem értékelhettek. Ez azt a célt szolgálta, hogy az alanyok hozzászokjanak a mesterséges hangokhoz, és hallják előre, hogy nagyjából milyen minőségre számíthatnak.

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.