Dialógustervezés VoiceXML-ben
(Dialog planning in VoiceXML)

Csapó Tamás Gábor <csapot AT tmit.bme.hu>
Tarján Balázs

2010. március 10.

1. Elméleti bevezető

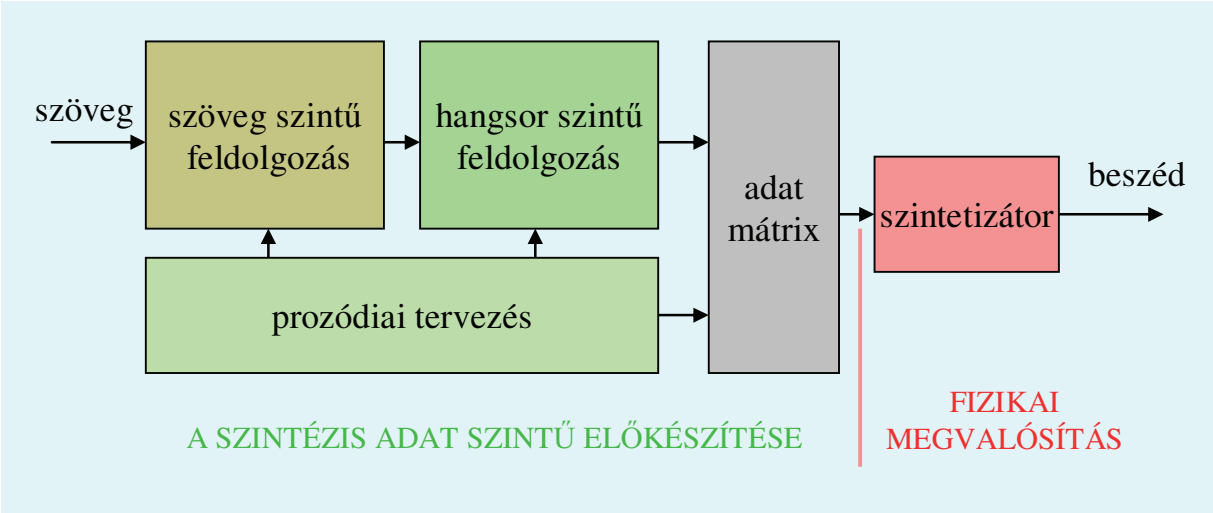
A mérés célja, hogy a hallgatók megismerkedjenek a VoiceXML programozási nyelv nyújtotta lehetőségekkel. A VoiceXML egy szabványos leíró nyelv, melynek célja, hogy a beszéd alapú ember-gép dialógusok fejlesztését meggyorsítsa és könnyítse. A VoiceXML a HTML nyelvhez hasonló elveken alapul, de előbbi a beszéd leírására szolgál, utóbbi pedig a vizuális tartalmak megjelenítését definiálja.
Az elméleti bevezetőben röviden ismertetjük a dialógustervezés, a beszédszintézis és a beszédfelismerés alapjait, melyek szükségesek a mérés elvégzéséhez.

1.1 Dialógus rendszer

A dialógus rendszerek célja, hogy az ember-gép kapcsolat megfelelő működését a beszédtechnológia eszközeivel segítsék. A párbeszéd, vagy más néven dialógus alapú rendszer létrehozásához egy olyan fejlesztőkörnyezet szükséges, amelynek segítségével kialakítható egy adott feladathoz az adott követelményeknek legjobban megfelelő dialógus rendszer. A rendszer beszéd generátor (más néven beszédszintetizátor) és robusztus beszédfelismerő motorok rendszerbe integrálását is támogatja [1].

1.1 Beszédszintézis

A beszédszintézis nem más, mint emberi beszéd előállítása mesterséges módon, tipikusan számítógép segítségével. Amennyiben a bemenet írott szöveg, szövegfelolvasóról (Angolul Text-To-Speech, röviden TTS) beszélünk. Ezt a szöveget a beszédszintetizátor különböző lépéseken keresztül alakítja át emberi beszéddé, amelyre az 1. ábrán látható példa. Általános szövegfelolvasó esetén ezek a lépések a bejövő szöveg feldolgozása, előkészítése a szintézishez, valamint a beszéd létrehozása [2]. Egy köztes lépés a prozódia tervezése, amely annyit jelent, hogy a szöveghez hozzárendeljük a dallamot, ritmust, a hangsúlyok helyeit és típusait. Ezeknek meghatározásához csak a bemeneti szöveg áll rendelkezésre, ami meglehetősen nehézzé teszi a lépést. A szintézis előkészítése után a tényleges beszédszintetizátor előállítja az adatokból a kimeneti beszédet.



1. ábra Általános szövegfelolvasó megvalósítási sémája. A működés két fő lépésből áll: bemeneti szövegből szimbolikus információ létrehozása (bal oldal), majd ez alapján hangfájl szintetizálása (jobb oldal). Forrás: [2, 303.oldal].

A beszédszintetizátorok különböző generációit különböztetjük meg működésük alapján [3, 4].

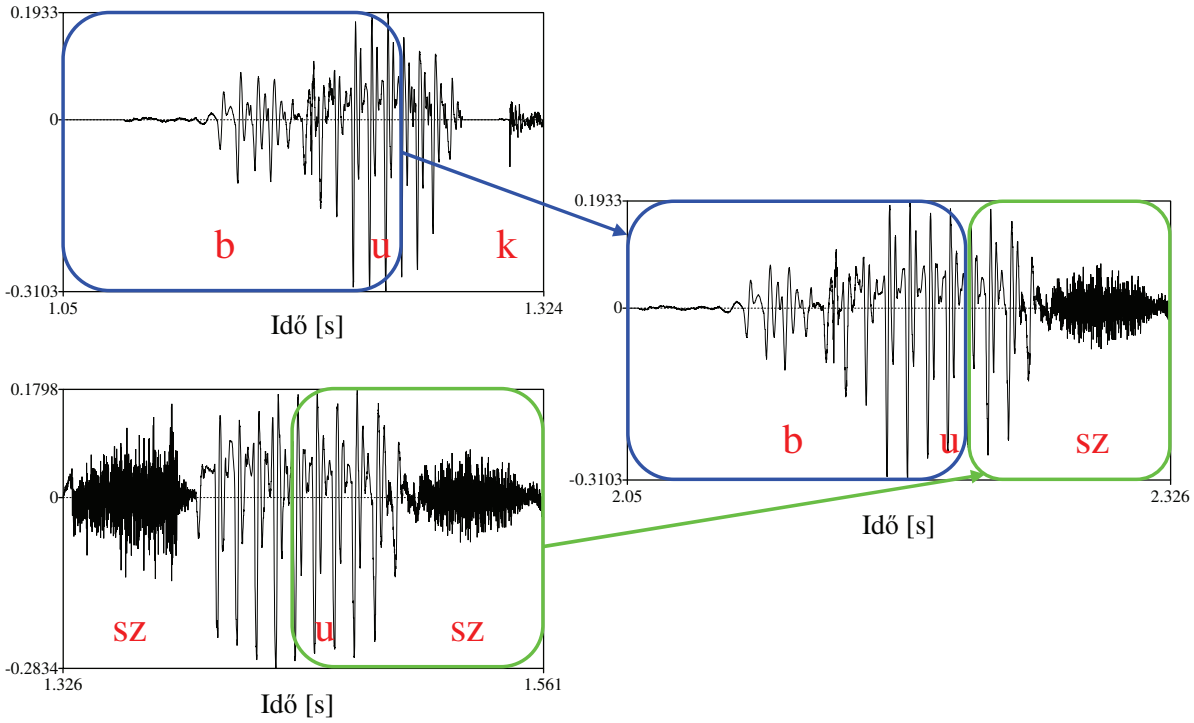
1.1.1 Formánsszintézis

A formánsszintézis volt az első olyan technológia, mellyel szöveget automatikusan érthető beszéddé lehetett alakítani. A rendszer az emberi beszéd formánsainak[1] modellezésével próbálja létrehozni a beszédhangot. Az ilyen rendszerek hangzása az érthetőség mellett meglehetősen „robotos", ami miatt ma már ritkán használják.

1.1.2 Elemösszefűzéses szintézis

Az elemösszefűzéses beszédszintézis során természetes beszédből kivágott hullámforma elemeket fűznek össze. Korábbi kísérletek kimutatták, hogy a mesterségesen előállított beszéd érthetőségéért a hangátmenetek természetessége felelős, nem maguk a beszédhangok. Emiatt kiemelten fontos a hangátmenetek megfelelő modellezése. Attól függően különböztetjük meg az elemösszefűzéses rendszereket, hogy mekkora a felhasznált elemek mérete. Természetesen ez a szükséges elemek számát is befolyásolja: míg a magyar nyelvű diádos[2] szintézishez szükséges elemek száma $38^2=1444$, addig triád[3] elemből $38^3=54872$ mintára lenne szükség. A gyakorlatban a teljes diád-lefedettség mellett a leggyakoribb 1000-2000 triád elem használatával már jó minőséget lehet elérni.

[1] A formáns az emberi beszédhangok jellegzetes hangszínét adó, rezonanciás úton felerősített felhangtartomány.
[2] A diád (angolul diphone) két félhang kapcsolata, vagyis egy hangátmenet (pl. ,,a-b'').
[3] A triád (angolul triphone) a környezetfüggő hangot jelenti (pl. ,,a-b-o'').



2. ábra Diád elemek (hangátmenetek) összefűzése: a „bu” és „usz” összefűzésével előáll a „busz” szó. Forrás: [5] alapján.

A 2. ábra a diádok összefűzésére mutat példát: két különböző hangkörnyezetből kivágott diád elem egymás után helyezésével jön létre a „busz” szó. Az elemek összefűzése után az előálló beszéd megfelelő prozódiájáról is gondoskodni kell jelfeldolgozási módszerek segítségével. Az ilyen módon létrehozott beszéd jól érthető ugyan, de messze van a természetes hangzástól.

1.1.3 Korpusz alapú, elemkiválasztásos szintézis

Az elemösszefűzéses technológia továbbfejlesztése az elemkiválasztásos beszédszintézis. Az újdonság itt egyrészt az, hogy nagyobb korpusz, vagyis beszédadatbázis áll rendelkezésre, amelyben egy-egy elem többször, többféle formában is előfordulhat. Másrészt ezek az elemek hosszabbak: szavak vagy akár szókapcsolatok is lehetnek.
A kimeneti beszéd létrehozása során a rendszer minél hosszabb olyan elemeket keres a korpuszban, amelyek a bemeneti szöveghez illeszkednek. A diádos és triádos rendszerekhez képest az elemek hosszabbak, így kevesebb összefűzési pont lesz az előállított beszédben. Mivel a korpuszban egy adott hangsorhoz tartozó beszédelem többféle formában (különböző dallammal, intenzitással) is előfordulhat, ezek közül a legtermészetesebbet választva javítható a szintetizált beszéd minősége. Ugyanakkor a rendszer minőségét az is befolyásolja, hogy a szintetizálandó szöveg és a beszédkorpusz témája mennyire van közel egymáshoz.

1.1.4 Rejtett Markov modell alapú szintézis

A statisztikai alapú, rejtett Markov modelleket[4] (HMM) alkalmazó beszédszintetizátor rendszerek egyre népszerűbbek lettek az elmúlt években (pl. HTS [6]). Az elemkiválasztásos rendszerek fő korlátja az, hogy a beszédkorpuszbeli hangsorozatokat használják. Így különböző beszédstílusok szintetizálásához egyre nagyobb adatbázis szükséges, amelynek előállítása meglehetősen költséges.
Ezzel szemben az új technológia alkalmazásához elég egy betanító korpusz, amelyből a rendszer környezetfüggő HMM-eket állít elő, a kimeneti hullámforma generálása pedig ezek alapján lehetséges. A betanítás a beszédfelismeréshez hasonlóan történik (hiszen a HMM-eket eredetileg erre használták), a tényleges szintézis eredménye pedig a hullámforma. Ezzel a módszerrel lehetővé válik különböző beszédstílusok, érzelmek modellezése a HMM paraméterek megfelelő módosításával.
Ez a technológia még nem teljesen kiforrott, jelen pillanatban is erőteljes kutatás és fejlesztés zajlik a statisztikai alapú TTS-ek területén.

1.1.5 A beszédszintetizátor-technológiák összehasonlítása

	Előny	Hátrány
Formánsszintézis	kis erőforrásigény	"robotos" hang, sok paraméter
Elemösszefűzés	kis erőforrásigény, jól állítható prozódia	jelfeldolgozás miatt torzulás
Elemkiválasztás	közel természetes hangzás	nagy tárhelyigény, prozódia nem állítható
Rejtett Markov modell	beszédfelismerésben alkalmazott technológia	lassú tanítás

1. táblázat: A beszédszintetizátor-technológiák összehasonlítása

A beszédszintetizátorok fokozatos változáson mentek keresztül az elmúlt 25 évben. A legegyszerűbb technológiáktól eljutottunk a bonyolult modellt alkalmazó rendszerekig, amit az 1. táblázat összegez. A formánsszintetizátorokkal leginkább csak „robotos”-nak mondott hang hozható létre, igaz, kis erőforrás használata mellett. A diád és triád elemeket összefűző rendszerek kis adatbázis használata mellett is az emberihez hasonló beszédet tudnak előállítani. A korpusz alapú, elemkiválasztásos beszédszintézis segítségével már szinte teljesen természetes beszéd állítható elő. A legújabb, rejtett Markov modell alapú rendszerek pedig kis memóriaigény mellett is jó minőséget tudnak szintetizálni.

A mérés során a BME TMIT-en kifejlesztett Profivox [7] szövegfelolvasót használjuk a feladatok elvégzésére. A Profivox magyar nyelvű beszédszintetizátor, amelynek legújabb változata az 1444 diád mellett 6000 CVC[5] triád-elemet is tartalmaz. A rendszer több felolvasó hanggal rendelkezik, amelyek közül egy férfi változatot alkalmazunk.

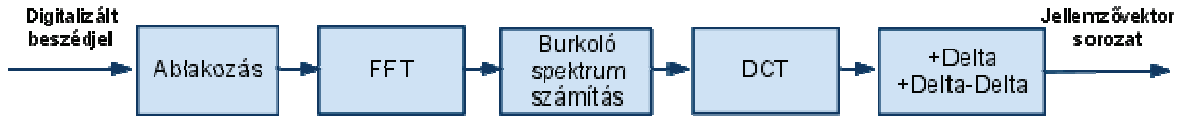
[4] A rejtett Markov modell (angolul Hidden Markov Model, röviden HMM) a beszéd egy valószínűségi modellje, amely diszkrét idejű, véges sok állapottal rendelkezik. A rejtett jelző arra utal, hogy csak a modell működésének eredményét ismerjük.
[5] Consonant-Vowel-Consonant, vagyis mássalhangzó-magánhangzó-mássalhangzó

1.2 Beszédfelismerés

A gépi beszédfelismerők célja az akusztikai beszédjelet szöveggé alakítani, ezzel lényegében a beszédszintézis inverz folyamatát valósítják meg. A felismerést két jól elkülöníthető fázisra szokás bontani. Az első a **lényegkiemelés** nevű jelfeldolgozási lépés, melynek során a beszédjelből kinyerjük a beszéd tartalmára jellemző paramétereket. A második ún. **mintaillesztési** szakaszban az előzőleg kapott paramétervektorokat illesztjük a nyelv egy tárolt modelljéhez. A teljes folyamat eredményeként a beszédfelismerő kimenetén a kérdéses beszédjelhez legjobban illeszkedő szót vagy szósorozatot kapjuk [8, 9].

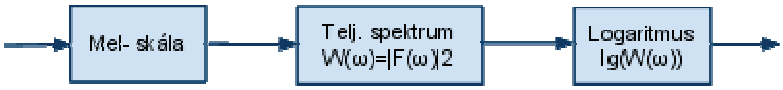
1.2.1 Lényegkiemelés

Az emberi beszéd igen komplex jel, ezért összetett feldolgozásra van szükség a tartalmát jellemző paraméterek kinyeréséhez.



3. ábra A lényegkiemelés folyamata

A jelfeldolgozást digitalizált beszédjelen végezzük, melyhez a beszéd időfüggvényét a feladattól függően mintavételezzük (~8-22 kHz) és kvantáljuk (8 bit, 16bit). Első lépésben a digitalizált jelet az emberi beszédhangok időtartamához illeszkedő szakaszokra (10-30ms) bontjuk átlapolt ablakfüggvények (pl. Hamming) segítségével. Jelenlegi tudásunk szerint az emberi füli harmonikus rezgéselemzést végez, így kézenfekvő, hogy a lényegkiemelés problémáját mi is frekvenciatartományban kezeljük. Az ablakozással keletkezett blokkokon belül a jel periodikusnak tekinthető, és spektruma előállítható FFT (Fast Fourier Transform) algoritmus segítségével. A spektrális összetevőkből jellemzővektor számításra többféle módszer is létezik. Itt röviden a legelterjedtebben használt mel-frekvenciás kepsztrális komponensek (MFCC, Mel Frequency Cepstral Coefficient) számítását mutatjuk be.



4. ábra MFCC burkoló spektrum számítás

Tömörebb jellemzők kinyeréshez átlagoljuk az FFT spektrum komponenseit. Az emberi hallás egyik fontos jellemzője, hogy frekvencia felbontása a frekvencia növekedésével exponenciálisan csökken. Ennek következtében a spektrális komponensekre alkalmazott mel-skálás átlagolásban a komponensek összegzésére használatos ablak szélességét 1 kHz fölött exponenciálisan növeljük, így kompenzálva a kisebb információsűrűséget. A mel összegzett komponensekből teljesítmény spektrumot számítunk zajelnyomási megfontolásból, majd az így kapott értékeket logaritmizáljuk illeszkedve az inger és érzet között általában fennálló kapcsolathoz. Ezután következik a DCT (Discrete Cosine Transform) nevű eljárás, mely arra szolgál, hogy a jellemzővektor dimenziószámát csökkenteni tudjuk. Utolsó lépésben a vektorhoz hozzárendeljük a statikus elemek lineáris regresszióval becsült időbeli deriváltjait is (Delta és Delta-Delta). Ez utóbbiak használata kíséretileg igazoltan sokat javít a lényegkiemelés hatásfokán. Mindezek mellett még számos zaj és torzítás csökkentő elemet tartalmazhat a lényegkiemelő, ezekre azonban részletesen nem térünk ki.
A lényegkiemelés tehát egy jelfeldolgozási lépés, mely standardizált, a további lépések tekintetében optimális, diszkrét idejű jelet képez a beszédből.

1.2.2 Mintaillesztés

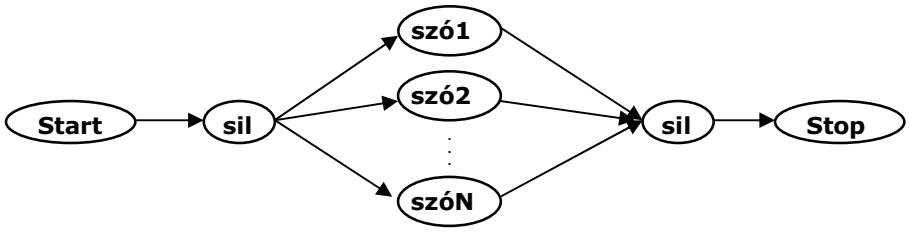
A mintaillesztés feladata a jellemzővektor sorozat leképzése egy szótári elemre vagy azok sorozatára (izolált szavas vs. folyamatos felismerő). Manapság legelterjedtebben **statisztikai** alapú felismerőket alkalmaznak, ahol a vektorsorozatot egy tanítóadatok alapján becsült, HMM alapú valószínűségi modell struktúrához illesztik. Ez a modell több hierarchiai szintre osztható (akusztikai modellek, nyelvi modell), melyek mindegyike egy-egy súlyozott, véges állapotú átalakítónak (WFST, Wighted Finite State Transducer) feleltethető meg a tanszéken alkalmazott megközelítés szerint [10]. A WFST-k a véges állapotú automatákból (FSA, Finite State Automata) származtathatóak az éleken elhelyezett súlyokkal (átmeneti valószínűségek) és kimeneti szimbólumokkal kiegészítve azokat, így alkalmasak arra, hogy a különböző hierarchia szintek lentről fölfelé egymás kimeneti címkéit dolgozzák fel. A mintaillesztés folyamatának végeredménye a legjobb illeszkedéshez tartozó útvonalon a felső modell (nyelvi modell) kimeneti címkéit, és elhangzásuk időpontját tartalmazza.

Akusztikai modell

A teljes felismerő hálózatban általában három akusztikai réteget különböztetünk meg, melyeket azonban tekinthetünk együtt egyetlen a jellemzővektor sorozatot fonéma sorozattá alakító WFST-nek. Az akusztikai modell tanítása címkézett hanganyagon történik. A rendelkezésünkre álló tanító hanganyag és a szöveges leirata alapján megbecsüljük a jellemzővektorok feltételes eloszlását az elhangzó fonémákra nézve. Az így létrejött statisztikai modellel becslést tudunk adni egy jellemzővektor sorozat és egy fonéma sorozat közötti akusztikai illeszkedés mértékére.

Nyelvi modell

A nyelvi modell határozza meg, milyen módon és milyen valószínűséggel kapcsolódhatnak össze a felismerő szótári elemei. Struktúráját döntően befolyásolja, hogy milyen feladatra szánjuk a felismerő rendszert. Izolált szavas felismerő esetén általában párhuzamos struktúrát használunk, ahol minden szótári elemen keresztül azonos valószínűséggel haladhatunk keresztül a mintaillesztés folyamán. Ilyen esetekben az akusztikus modell által szolgáltatott súly alapozza meg a döntésünket (5. ábra). Ezzel szemben folyamatos beszéd felismerésekor tanítószöveg alapján becsüljük meg a szótári elemek kapcsolódási valószínűségét, mely bár jóval összetettebb modell struktúrához vezet, lehetőséget teremt az akusztikai modell becslési bizonytalanságainak kivédésre. A nyelvi és akusztikai modell összekapcsolásából létrejövő WFST-t felismerő hálózatnak nevezzük.



5. ábra Egy izolált szavas felismerő nyelvi modelljének sémája (sil – a szünetmodell)

Dekódolás

A mintaillesztés folyamata egy a felismerési hálózatban történő, a jellemzővektorok által vezérelt optimális útvonalkeresésként fogható fel. Mivel egy kimerítő keresés számításigénye túl nagy lenne, a gyakorlatban elterjedten használják a dinamikus programozáson alapuló Viterbi-algoritmust, mellyel meghatározható minden időpillanatban az odáig tartó legjobb

útvonal. Emellett a folyamat további gyorsításához időről-időre el szokás hagyni kevésbé valószínű részútvonalakat.

A mérés során a BME TMIT-en kifejlesztett WFST alapú VOXerver beszédfelismerőt használjuk a feladatok elvégzésére.

1.3 Felhasznált irodalom

[1] Chetan Sharma & Jeff Kunins, VoiceXML: Strategies and Techniques for Effective Voice Application Development with VoiceXML 2.0, Wiley 2002

[2] Olaszy Gábor –Kovács Magdolna – Nikléczy Péter – Gósy Mária: Magyar nyelvi beszédtechnológiai alapismeretek. (600 oldal CD-ROM-on). http://alpha.tmit.bme.hu/pub/beszinf/start.html, 2002.

[3] Csapó Tamás Gábor, „Változatos prozódia megvalósítása szövegfelolvasó rendszerekben", BME TMIT, Diplomaterv, 2008.

[4] Fék Márk – Pesti Péter – Németh Géza – Zainkó Csaba: Generációváltás a beszédszintézisben. LXI. évf. (2006) 3. sz., Híradástechnika, 21–30. p.

[5] Sprachsynthese. Technische Universität Dresden, Institut für Akustik und Sprachkommunikation, http://www.ias.et.tu-dresden.de/sprache, 2008.

[6] Heiga Zen – Takashi Nose – Junichi Yamagishi – Shinji Sako – Takashi Masu ko – Alan W. Black –Keiichi Tokuda: The HMM-based speech synthesis system (HTS) version 2.0. In SSW6-2007 (konferenciaanyag). 2007, 294–299. p.

[7] Gábor Olaszy – Géza Németh – Péter Olaszi – Géza Kiss – Géza Gordos: PROFIVOX – a Hungarian professional TTS system for telecommunications applications. 3. évf. (2000. december) 3/4. sz., International Journal of Speech Technology, 201–216. p.

[8] Mihajlik Péter, „Spontán magyar nyelvű beszéd gépi felismerése nyelvspecifikus szabályok nélkül", PhD értekezés.

[9] Fegyó Tibor, Mihajlik Péter, „Gépi beszédfelismerés", oktatási segédanyag, 2009.

[10] M. Mohri, F. Pereira and M. Riley, "Weighted Finite-State Transducers in Speech Recognition," *Computer Speech and Language*, 16(1):69-88, 2002