DIPLOMATERV

VÁLTOZATOS PROZÓDIA MEGVALÓSÍTÁSA SZÖVEGFELOLVASÓ RENDSZEREKBEN

Készítette:
CSAPÓ TAMÁS GÁBOR
csapo@tmit.bme.hu

Konzulensek:

DR. NÉMETH GÉZA
nemeth@tmit.bme.hu

DR. FÉK MÁRK
fek@tmit.bme.hu

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
VILLAMOSMÉRNÖKI ÉS INFORMATIKAI KAR
TÁVKÖZLÉSI ÉS MÉDIAINFORMATIKAI TANSZÉK

2008.

BUDAPESTI MŰSZAKI ÉS GAZDASÁGTUDOMÁNYI EGYETEM
VILLAMOSMÉRNÖKI ÉS INFORMATIKAI KAR
TÁVKÖZLÉSI ÉS MÉDIAINFORMATIKAI TANSZÉK

DIPLOMATERV

Csapó Tamás Gábor

---

mérnökjelölt részére

Feladat:

Változatos prozódia megvalósítása szövegfelolvasó rendszerekben

A BME TMIT Beszédtechnológiai Laboratóriumában a témában végzett korábbi kutatások eredményeinek felhasználásával valósítsa meg az alábbi részfeladatokat:

- Tekintse át a szövegfelolvasó rendszerekben a változatos prozódia modellezésének irodalmát, különös tekintettel a szabály alapú ill. a gépi tanulási módszerekre
- Alakítson ki olyan prototípus rendszert a tanszéken fejlesztett Profivox diád/triád alapú szövegfelolvasó rendszer módosításával, amely a korábbi kutatásoknál felhasználtnál nagyobb szövegkorpuszon alapul és közel valós időben képes magyar nyelvű szövegből hangot előállítani
- Vizsgálja meg a változatos prozódia előállításának lehetőségeit korpusz-alapú rendszerekben
- Eredményeit és megállapításait szubjektív meghallgatásos tesztek alapján értékelje

Záróvizsga tárgyak:
Mobil infokommunikáció (Imre, BMEVIHI4380 dipl. tervhez kapcsolódó)
Beszédinformációs rendszerek (Gordos - Németh, BMEVITT3247)
Deklaratív programozás (Hanák - Szeredi, BMEVIFO2218)

A feladat beadásának határideje: 2008. május 16.

Tanszéki konzulens:
Dr. Németh Géza

Ipari konzulens:
Dr. Fék Márk

A tervezés bírálója:

Budapest, 2008. február

(Dr. Sallai Gyula)
egy. tanár, tanszékvezető

E feladatlap a diplomatervhez csatolandó.

Konzultációk:

Időpont			Észrevételek	Konzulens
év	hó	nap		

Az ipari konzulens véleménye:

A tanszéki konzulens véleménye:

Nyilatkozat

Alulírott *Csapó Tamás Gábor*, a Budapesti Műszaki és Gazdaságtudományi Egyetem hallgatója kijelentem, hogy ezt a diplomatervet meg nem engedett segítség nélkül, saját magam készítettem, és a diplomatervben csak a megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos értelemben, de átfogalmazva más forrásból átvettem, egyértelműen, a forrás megadásával megjelöltem.

Budapest, 2008. május 16.

Csapó Tamás Gábor
hallgató

Kivonat

Az információs társadalom kialakulása során elengedhetetlen az ember-gép kapcsolat folyamatos fejlesztése. A beszédtechnológiai alkalmazások, ezen belül a beszédszintézis elterjedése is ebbe a folyamatba illeszkedik, hiszen sok esetben a felhasználó és a gép között beszéd segítségével megvalósuló kommunikáció nélkülözhetetlen.

A beszédszintézis rendszerek minőségét az alapján ítélik meg, hogy az általuk keltett beszéd mennyire hasonlít az emberire. A létrehozott hang érthetősége a mai szövegfelolvasókban már nem jelent problémát. A jelenlegi rendszerek többsége egy szabályrendszer segítségével a nyelvi elvárásoknak megfelelő, adott szöveghez mindig azonos prozódiát (intonáció, hangsúlyozás, ritmus) 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. Az elmúlt években fokozatosan előtérbe került ez a kulcskérdés.

A dolgozatban először áttekintjük a beszédszintézis szakirodalmát, a prozódia modellezésének lehetőségeit részletesen ismertetve. Bemutatjuk az emberi beszéd változatosságának vizsgálatára és modellezésére tett kísérleteket. Megtervezünk és implementálunk egy módszert, amely alkalmas a korábbiaknál változatosabb mesterséges beszéd előállítására. Ezt oly módon végzi, hogy a bemeneti szöveghez természetes beszédből származó minták alapján határozza meg a dallamot. A változatosságot a dallamminta választásának bizonyos fokú véletlenszerűsége biztosítja. A megvalósított rendszer minőségét a különböző szempontok szerint elvégzett elemzések eredményei tanúsítják.

Az ily módon természetesebb hangzásúvá tett szövegfelolvasó rendszer számos gyakorlati alkalmazásban használható, mint például SMS-, e-levél-, könyv-felolvasó, vagy telefonos tudakozó. A változatosabb prozódia főleg hosszú szövegek felolvasása esetén előnyös, hiszen ekkor zavaró leginkább a beszédszintetizátor monotonitása.

Abstract

The continuous development of human-machine relationship is indispensable during the evolution of the information society. The spreading of speech technology applications and speech synthesis conform this process. The communication between user and machine aided by speech is often essential.

The quality of speech synthesis systems is judged on the basis of how they resemble human speech. The intelligibility of the produced voice is solved in the Text-To-Speech synthesizers used in our days. Most current systems assign suitable prosody (intonation, stressing and rhythm) to the input text by using predefined rules. For a given text this causes always the same speech. One of the key issues of last years is the modelling of human variability in synthesized speech: the same sentence should not sound the same when repeated.

In this thesis the scientific literature of speech synthesis is summarized first, including the detailed review of the possibilities of prosody modelling. Some experiments are introduced in connection with the modelling and investigation of the variability of human speech. A method is planned and implemented, that can produce more varied artificial speech than the former solutions. The intonation of the input text is determined using samples from natural speech. Variability is assured by the random selection from the available melody samples. The quality of the implemented system is evaluated by objective and subjective tests.

This Text-To-Speech synthesis system, that produces more natural sound than the previous solutions, can be used in numerous practical applications: SMS, e-mail, book reader software or directory inquiries. The monotony of synthesized speech is mostly disturbing while synthesizing extended passages; variable prosody is in these cases important.

Tartalomjegyzék

Ábrák jegyzéke

Táblázatok jegyzéke

Rövidítések

ANSI	American National Standards Institute
CART	Classification And Regression Tree
CBR	Case-Based Reasoning
CVC	Consonant-Vowel-Consonant
GToBi	German Tones and Break indices
HMM	Hidden Markov Model
ISCA	International Speech Communication Association
IViE	International Variation in English
JND	Just Noticeable Differences
ToBi	Tones and Break indices
TTS	Text-To-Speech
XML	Extensible Markup Language

Bevezetés

Napjainkban közösen éljük meg az információs társadalom kialakulását. Ehhez elengedhetetlen az ember-gép kapcsolat folyamatos fejlesztése, ugyanis széles rétegek számára kell elérhetővé tenni az új technológia által kínált lehetőségeket. Ebbe a folyamatba illeszkedik a beszédtechnológiai alkalmazások, ezen belül is a beszédszintézis elterjedése. A felhasználó és a gép között beszéd segítségével megvalósuló kommunikáció nélkülözhetetlen, ha a felhasználó keze és látása lekötött (pl. autóvezetés közben), illetve sérülés miatt nem használható (pl. látássérültek vagy olvasási problémákkal küzdők esetén), továbbá ha az igénybe vett szolgáltatás telefonvonalon keresztül érhető el (pl. intelligens tudakozó, hírfelolvasás mobil eszközön).

A beszédszintézis rendszerek minőségét az alapján ítélik meg, hogy az általuk keltett beszéd mennyire hasonlít az emberi beszédre. A létrehozott hang érthetősége a mai szövegfelolvasókban már nem jelent problémát, de az még élesen megkülönböztethető az emberi beszédtől. A jelenlegi rendszerek többsége egy szabályrendszer segítségével a nyelvi elvárásoknak megfelelő, adott szöveghez mindig azonos prozódiát (intonáció, hangsúlyozás, ritmus) 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. Az elmúlt években fokozatosan előtérbe került ez a kulcskérdés, vagyis a változatos prozódia megvalósítása.

A prozódiai változatosság alatt érthetjük az intonáció, a hangsúlyozás, vagy a ritmus variálását. A cél az, hogy a gépi megoldás jobban modellezze az emberi beszédet ebből a szempontból, hiszen a valóságban sincs két egyformán kiejtett mondat, mert a prozódia egy adott személy beszédében is folyamatosan változik. Természetesen a változatosság nem egyszerű véletlenszerűség, a beszéd létrehozását számos fizikai kényszer (pl. a levegő útja a tüdőnkből indulva) és mentális jellemző (pl. a fontosnak tartott gondolat hangsúlyozása) irányítja, aminek vizsgálata hosszabb kutatást igényel.

A diplomaterv kiírás elemzése

A diplomaterv feladatot a BME TMIT Beszédtechnológiai Laboratóriumában a témában végzett korábbi kutatások eredményeinek felhasználásával tervezem megoldani.

A szövegfelolvasókhoz kapcsolódó szakirodalom rendkívül sokrétű, számos nemzetközi és hazai konferencia (pl. az évente megtartott *Interspeech* konferenciasorozat, az ISCA által rendezett *Speech Synthesis Workshop*, valamint a magyar nyelvű *Beszédkutatás* sorozat), tudományos folyóirat (pl. *Speech Communication*, *International Journal of Speech Technology*, a magyar nyelvű *Híradástechnika* folyóirat évente megjelenő, beszédtechnológiával foglalkozó különszáma), könyv és internetes forrás (pl. a német *Technische Universität Dresden* honlapja) foglalkozik a beszédtechnológia ezen témakörével. Ezekből kell kiválasztani a mélyreható irodalomkutatás során a változatos prozódia modellezésével kapcsolatos tanulmányokat.

A szakirodalom áttekintése után vizsgálni fogom az emberihez hasonló, változatos prozódia létrehozásának lehetőségeit szövegfelolvasó segítségével. A korpusz alapú rendszerekre külön is kitérek, hiszen a jelenleg legtermészetesebb hangzásúnak tartott beszédszintetizátor-technológia alkalmazásával jó eredményeket lehetne elérni a mesterséges beszéd monotonitásának csökkentésében. A lehetőségek közül egyet kiválasztva megtervezem annak működését.

A változatos beszédet megvalósító rendszer tervezése után implementálni fogom az általam kiválasztott alternatívát. A tanszéken alkalmazott szövegfelolvasó rendszert először részletesen megismerem, majd létrehozok egy ehhez illeszthető prototípus rendszert, amely képes változatos mesterséges beszédet előállítani.

Végül a megvalósított rendszer minőségét elemezni fogom. Először automatikus tesztekben vizsgálom a megfelelő működés követelményeinek teljesülését. A beszédkutatásban a gépi szimuláción kívül mindenképpen szükség van szubjektív értékelésre is, hiszen a létrejövő rendszert emberek fogják használni, hallgatni. Megállapításaimat ezért meghallgatásos tesztek segítségével is fogom ellenőrizni.

A diplomaterv felépítése

A dolgozatban legelőször ismertetem a későbbiekben használt fogalmakat (1. fejezet). Részletesen bemutatom az emberi beszéd prozódiájának tulajdonságait, összetevőit szubjektív és objektív szempontok szerint vizsgálva. A beszédszintetizátorok, más néven szövegfelolvasók az elmúlt évek során sokat fejlődtek, több különböző technológiát fejlesztettek ki mesterséges beszéd létrehozására, melyek működését röviden áttekintem. Az ezen rendszerekben alkalmazott prozódiai modellek is bemutatásra kerülnek. Külön elemzem az egyes modellek

előnyeit és hátrányait. A diplomaterv kiírásban szereplő változatos prozódiamodellezés irodalmának bővebb ismertetését az 1.3. alfejezet tartalmazza. A szabály alapú és gépi tanulási modellek nagyobb hangsúlyt kaptak ezen összefoglalásban. A modellek egy részével részletesebben is foglalkozom, így a korpusz alapú prozódiai modellekről bővebben olvashatunk. A fejezet végén az emberi beszéd változatosságáról írok. Betekintést nyerhetünk abba is, miért nem foglalkoztak eddig az emberi beszéd ezen tulajdonságának átültetésével a mesterséges beszédre.

A 2. fejezetben a változatos prozódia modellezésének hiányát próbálom orvosolni egy erre kifejlesztett szövegfelolvasó rendszer tervezésével. Először azt gyűjtöm össze, hogy a megvalósítandó rendszernek milyen tulajdonságokkal kell feltétlenül rendelkeznie. Bemutatom, milyen alternatívák jöhetnek szóba ennek megoldására: vizsgálom az alkalmazható beszédszintetizátor-technológiát és a felhasználható prozódiai modellt. A korpusz alapú rendszerek segítségével megvalósítható változatos prozódia-előállítás lehetőségeit külön is tárgyalom a 2.2.1. részben, a kiírás szerint. Ezután kiválasztok egy konkrét technológiát, és ismertetem a megvalósítandó rendszer tervezett működését. Azzal is foglalkozom, hogyan lehet a módszert általános szövegfelolvasóba illeszteni, és így széles körben használni. Az utolsó alfejezetben bemutatom a megvalósítandó rendszer azon előnyeit, hátrányait és korlátait, amelyek már a tervezés folyamán is ismertek.

A következő fejezet a tervezés során bemutatott rendszer implementálásával, és annak körülményeivel foglalkozik (3. fejezet). A fejezet elején röviden ismertetem korábbi munkánkat és annak eredményét is. Olvashatunk a megvalósítás során felhasznált nagyméretű, természetes beszédet tartalmazó korpuszok részleteiről, valamint ezeknek feldolgozásáról. Diplomatervem céljai közé tartozik, hogy a változatos dallam megvalósítására létrehozott módszerből a BME TMIT-en kifejlesztett szövegfelolvasó segítségével egy prototípus rendszert állítsak össze. A kiírásban is szereplő részfeladat megoldásának bemutatása a 3.3. és a 3.4. alfejezetekben található. A kidolgozott módszer közel valós időben képes magyar nyelvű szövegből változatos hangot előállítani.

A 4. fejezet a megvalósított, szövegfelolvasóba integrált, változatos beszéd létrehozására alkalmas rendszer értékelését tartalmazza. A rendszer tervezése során a követelmények között meghatározott tulajdonságok teljesülését vizsgálom, a megfelelő működés ezeknek automatikus ellenőrzésével biztosítható. Bemutatom, hogy mely részeket sikerült megvalósítani a tervezett rendszerből, és mi maradt ki a feladat megoldása során. Ezután annak elemzése következik, hogy milyen minőségű mondatok állíthatók elő a módszerrel. A kutatás során a korábbi fázisok is meghallgatásos tesztekkel zárultak, ezeknek körülményeit és eredményeit röviden ismertetem. A jelenlegi szubjektív teszt, amelyet a diplomaterv kiírása tartalmaz, a 4.2.2. részben ta-

lálható. A kísérlet céljának ismertetése, kivitelezésének lépései és a tesztkörnyezet bemutatása is megtörténik. Az eredmények kiértékelése a fejezet végén olvasható.

A dolgozat végén összefoglalom a kutatás célját és folyamatát. Kitérek a változatos prozódia előállításában elért eredményekre. Végül ismertetem a további kutatási lehetőségeket is, és újabb célokat tűzök ki a változatosság növelésének érdekében.

A dolgozat 1–4. fejezetének szövegében többes szám első személyt fogok használni. A Bevezetésben és az Összefoglalásban az önálló munkámat különítem el egyes szám használatával.

1. fejezet

A prozódia szerepe és modellezése a beszédszintézisben

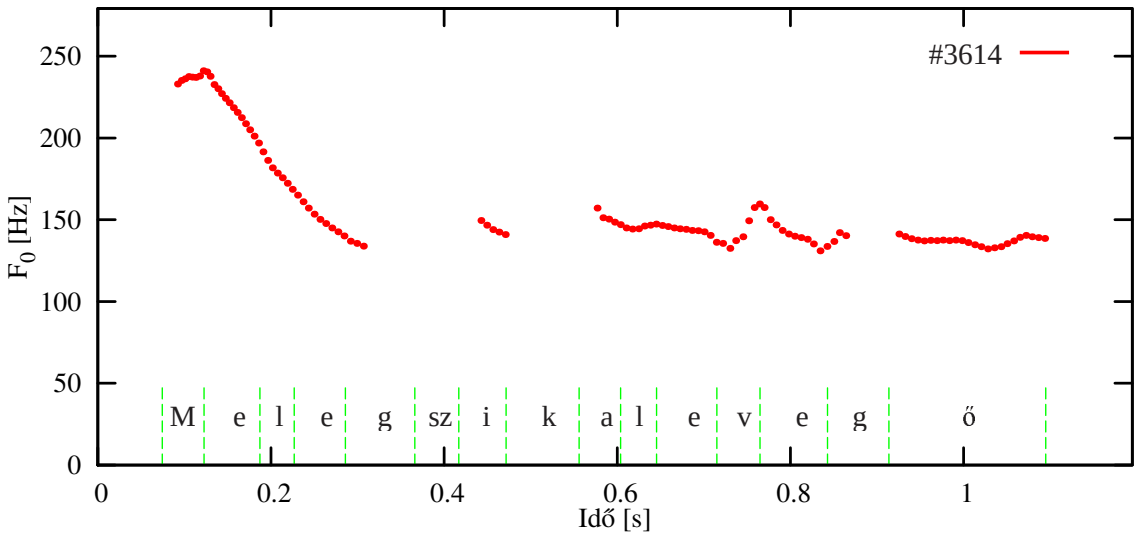
A fejezetben bemutatásra kerülnek a dolgozat megértéséhez szükséges alapfogalmak, először a prozódiáról és összetevőiről olvashatunk (1.1. alfejezet). Betekintést nyerhetünk a beszédszintetizátor-technológiák fokozatos fejlődésébe is (1.2. alfejezet). Ezután az 1.3. alfejezet röviden ismerteti a szövegfelolvasó rendszerekben alkalmazott prozódiai modelleket, majd az 1.4. alfejezetben néhányuknak részletesebb bemutatása következik. Az 1.5. alfejezet a prozódia változatosságának felderítésére irányuló kísérletekkel foglalkozik. A fejezet végén ismertetjük a jelenlegi szövegfelolvasók egyik gyengéjét is: az emberihez hasonló változatos prozódia modellezésének hiányát.

1.1. A prozódia fő összetevői szubjektív szempontok szerint

A folyamatos emberi beszéd kifejező erejét a prozódia (a dallam, a ritmus, a tempó, a hangsúlyozás, a hangerő és a hangszínezet változtatása) adja. Ezzel tudjuk érzékeltetni többek között a közölt mondat modalitását (kijelentő, kérdő stb.), lelki állapotunkat, valamint kiemelhetjük mondandónkból a fontos elemeket.

Az egyes prozódiai összetevőket szubjektív és objektív paraméterekkel is lehet jellemezni. A szubjektív paraméterek az ember által érzékelhető, érzékszerveinkkel felfogható részeket jelentik, míg objektív paramétereken a gép által mérhető adatokat értjük.

A prozódia összetevői közül a három legfontosabb szubjektív szempontok szerint a dallam, a hangsúly és a ritmus, amelyeket Olaszy és társai műve alapján ismertetünk [1]. A dallamhoz



1.1. ábra. A prozódia három fő összetevője objektív szempontok szerint a *„Melegszik a levegő."* mondatban. A piros görbe az alapfrekvencia-menetet ábrázolja, a mondat eleji F_0-emelkedés hangsúlyra utal, a zöld vonalak pedig a hangidőtartamokat mutatják.

tartozó objektív paraméter a beszéd alapfrekvenciájának[1] változása. A hangsúly többek között az intenzitás- és az alapfrekvencia-menet lokális csúcsainak helyein érzékelhető. A ritmus a beszédrészletek időtartamainak változását jelenti. A dolgozatban ezen összetevők szubjektív és objektív elnevezését is fogjuk használni, mivel azok megfeleltethetőek egymásnak.

1.1.1. Dallam

Az emberi beszédkeltés egyik legfontosabb eszköze a dallam [1, 242. oldal], amit az alapfrekvencia változtatásával hoz létre a beszélő. A beszéd dallama több szintre bontható fel. A legmagasabb (szupraszegmentális) szint határozza meg a mondatok modalitását: kijelentő, kérdő, felkiáltó, óhajtó, felszólító. A szupraszegmentális, más néven mondatszintű dallamot, amelynek kezdő- és végpontja szoros összefüggésben van egymással, az adott nyelv határozza meg. Ehhez kapcsolódva az egyes mondatok dallama alapvetően három féle lehet: emelkedő, ereszkedő vagy lebegő. A valóságban ezek kombinációi is előfordulnak. A középső szint a szó- és szótagszintű alapfrekvencia-változásokat foglalja magában. A legalacsonyabb (szegmentális) szinten a mikrointonáció jelenik meg, amely az egy hangon belüli alapfrekvencia-változást jelenti.

Az 1.1. ábrán láthatjuk egy mondat dallamát, azaz F_0-menetét. Az alapfrekvencia csak a beszéd zöngés részein értelmezett, ezért nem folytonos az F_0-görbe.

[1] A beszéd zöngés részei kvázi-periodikusak. Az alapfrekvencia (röviden F_0) a periódusidő reciproka.

1.1.2. Hangsúly

A hangsúlyozás segítségével nyomatékot helyezhetünk egy mondatrészre, szóra vagy szótagra [1, 246. oldal]. Ez három fizikai paraméterrel jellemezhető: F_0 emelés, időtartam nyújtás, intenzitás növelés. A hangsúlyozást befolyásolja a szöveg tartalma, annak értelmezése is, emiatt különböző esetekben eltérő szótagokat hangsúlyozhatunk. A magyar nyelv hangsúlyozási szabálya szerint a szavak első szótagján van nyomaték. Más nyelveknél ez nem feltétlenül van így, angolban például változó helyen van a szón belüli hangsúly.

A folyamatos beszédben nem egymástól elválasztott szavakat mondunk, hanem azokat szünet nélkül egymásba fűzzük. Azt a szócsoportot, amelyet beszédünkben egyben, szünet nélkül mondunk ki, prozódiai egységnek (más néven frázisnak) nevezzük. Írásban ezeket általában írásjelek határolják. A prozódiai egységek egy-egy hangsúlycsoportot képviselnek, és csak egy szavukon van normál[2] hangsúly, a többi szó semleges vagy hangsúlytalan lehet.

Az 1.1. ábra olyan hangsúlyt mutat a mondat elején, amely láthatóan (legalábbis részben) az alapfrekvencia emelésével valósult meg.

1.1.3. Ritmus

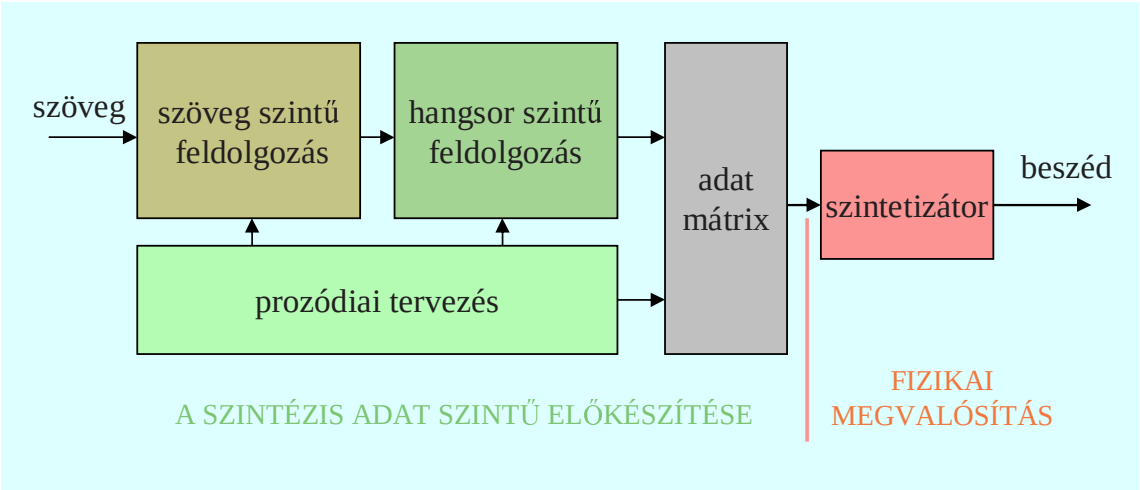
A beszéd ritmusa [1, 249. oldal] az egyes hangok hosszát, a beszédtempó változását, a beszéd ritmikáját, és a szünetek tartásának módját foglalja magában. A tempó és a ritmus többek között függ a nyelvtől, a beszélő egyéniségétől illetve érzelmi állapotától. A folyamatos beszédben az egyes hangokat hol gyorsabban, hol lassabban ejtjük, a hangidőtartamok változása ekkor 10–20% is lehet.

Az 1.1. ábrán a *„Melegszik a levegő."* mondat hangjainak időtartamai is láthatóak.

1.2. A beszédszintetizátorok generációi

A beszédszintézis nem más, mint emberi beszéd előállítása mesterséges módon, tipikusan számítógép segítségével. Amennyiben a bemenet írott szöveg, szövegfelolvasóról[3] beszélünk. Ezt a szöveget a beszédszintetizátor különböző lépéseken keresztül alakítja át emberi beszéddé, amire az 1.2. ábrán látható példa. Általános szövegfelolvasó esetén ezek a lépések a bejövő szöveg feldolgozása, előkészítése a szintézishez, valamint a beszéd létrehozása [1, 303. oldal]. Egy köztes lépés a prozódia tervezése, amellyel dolgozatunkban foglalkozunk. A prozódiai előrejelzés annyit jelent, hogy a szöveghez hozzárendeljük a dallamot, ritmust, a hangsúlyok helyeit

[2] A hangsúlyok osztályozása [1, 246. oldal]-on található.
[3] Angolul *Text-To-Speech*, röviden TTS.



1.2. 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: [1, 303. oldal].

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.

A beszédszintetizátorok különböző generációit különböztetjük meg működésük alapján, melyeket Fék és társai munkája alapján ismertetünk [2].

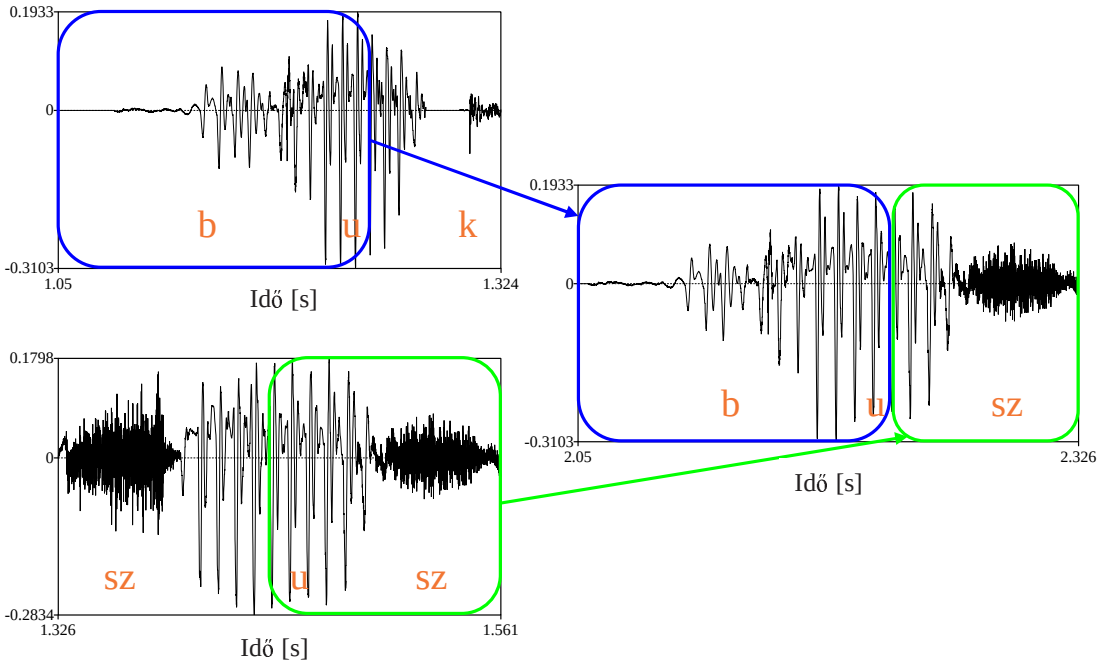
1.2.1. Formánsszintézis

A formánsszintézis volt az első olyan technológia, mellyel szöveget automatikusan érthető beszéddé lehetett alakítani. A rendszer az emberi beszéd formánsainak[4] modellezésével próbálja létrehozni a beszédhangot. Mivel a formánsszintézishez szükséges paraméterek megfelelő hangolása nehézkes, az ilyen rendszerek hangzása az érthetőség mellett meglehetősen „robotos”, ami háttérbe szorította őket. Az első magyar nyelvű szövegfelolvasó szoftver a Multivox formánsszintetizátor volt [3].

1.2.2. Elemösszefűzéses szintézis

Az elemösszefűzéses beszédszintézis során természetes beszédből kivágott hullámforma elemeket fűznek össze. A 20. század elején végzett kísérletek megmutatták, hogy a mestersé-

[4]A formáns az emberi beszédhangok jellegzetes hangszínét adó, rezonanciás úton felerősített felhangtartomány.



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

gesen előállított beszéd érthetőségéért a hangátmenetek természetessége felelős, nem maguk a fonémák [4]. 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[5] szintézishez szükséges elemek száma $38^2 = 1444$, addig triád[6] 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.

Az 1.3. 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 előírt alapfrekvencia-menet megvalósítása a legkritikusabb, ugyanis az alapfrekvencia csak körülbelül 30%-kal módosítható még elfogadható minőségben a jelenleg alkalmazott algoritmusokkal. Az ily módon létrehozott beszéd jól érthető ugyan, de messze van a természetes hangzástól.

[5] A diád (angolul *diphone*) két félhang kapcsolata, vagyis egy hangátmenet (pl. „a-b”).
[6] A triád (angolul *triphone*) a környezetfüggő hangot jelenti (pl. „a-b-o”).

1.2.3. 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/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 mennyire van közel egymáshoz.

1.2.4. Rejtett Markov modell alapú szintézis

A statisztikai alapú, rejtett Markov modelleket[7] 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), míg a tényleges szintézis a felismerés inverze, aminek eredménye a hullámforma. Ezzel a módszerrel lehetővé válik különböző beszédstílusok, érzelmek modellezése a HMM paraméterek megfelelő módosításával.

1.2.5. A beszédszintetizátorok összehasonlítása

A beszédszintetizátorok fokozatos változáson mentek keresztül az elmúlt 25 évben. A legegyszerűbb technológiáktól eljutottunk a bonyolult modellt alkalmamazó rendszerekig, amit az 1.1. táblázat összegez. A formánsszintetizátorokkal leginkább csak „robotos”-nak mondott hang hozható létre, igaz, kis erőforrás használata mellett. A diád és triád elemeket összefűző

[7] 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.

1.1. táblázat. A beszédszintetizátorok összehasonlítása.

	Előny	**Hátrány**	**Példa**
Formánsszintézis	kis erőforrásigény	„robotos” hang sok paraméter	Multivox formánsszintetizátor
Elemösszefűzés	kis erőforrásigény jól állítható prozódia	jelfeldolgozás miatt torzulás	Profivox diád/triád elemes
Elemkiválasztás	közel természetes hangzás	nagy tárhelyigény prozódia nem állítható	Kísérleti korpusz alapú rendszer
HMM alapú	beszédfelismerésben alkalmazott technológia	statisztikai ismeretek szükségesek	HTS 2.0

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 tanítóadatbázis mellett is jó minőséget tudnak szintetizálni.

A dolgozat során a BME TMIT[8]-en kifejlesztett Profivox [7] szövegfelolvasót használtuk tesztek elvégzésére. A Profivox magyar nyelvű beszédszintetizátor, amelynek legújabb változata az 1444 diád mellett 6000 CVC[9] triád-elemet is tartalmaz. A rendszer több felolvasó hanggal rendelkezik, amelyek közül egy férfi változatot alkalmaztunk.

1.3. Prozódiai modellek csoportosítása

Az 1.2. alfejezetben említettük a szövegfelolvasók két fő lépését: az első a bemeneti szöveg alapján szimbolikus információt hoz létre, majd ebből a második lépés során előáll a beszéd. A prozódia modellezése, a paraméterek megfelelő beállítása az első lépésben történik a „prozódiai tervezés” során (1.2. ábra). A szimbolikus információ szövegből történő származtatására sokféle modell ismert, melyeket röviden ismertetünk.

A modelleket két dimenzió szerint csoportosíthatjuk: az első alapján meg kell különböztetnünk a leíró jellegű és a szuperpozíciós megvalósításokat. A másik csoportosítási szempont a modellek tanítása: ember által definiált szabályok alapján, illetve gépi tanulás segítségével, adatvezérelt módon is működhetnek. Ezek a jellemzők a prozódiai modelleket más-más szempontból írják le, a gyakorlatban kombinációikat használják. A dolgozatban alkalmazott Profivox beszédszintetizátor szabály alapú, szuperpozíciós modellel rendelkezik [8].

[8]Budapesti Műszaki és Gazdaságtudományi Egyetem - Távközlési és Médiainformatikai Tanszék

[9]*Consonant-Vowel-Consonant*, vagyis mássalhangzó-magánhangzó-mássalhangzó

1.3.1. Leíró jellegű modellek

A leíró jellegű, nyelvi modellek célja, hogy az intonációt címkék segítségével írják le. Az egyik ilyen rendszerben, a ToBi[10]-ban [9] ezek a címkék a jellegzetes alapfrekvencia-változásokat jelölik: magas (*High*, H), alacsony (*Low*, L), szóhangsúly (L*), frázishangsúly (H-). A rendszer alulspecifikáltnak számít, mert minimális címkehalmazzal próbálja leírni az intonációt.

A ToBi az angol nyelv címkézésére alkalmas, más nyelvekre különböző kiterjesztéseit alkalmazzák (pl. GToBi[11] a német nyelvre [10], IViE[12] az angol nyelv bővített változatára [11]). Az alapvető probléma a leíró jellegű modellekkel, hogy a címkék szöveghez rendelése csak kézi vagy félautomatikus módszerekkel oldható meg, ami drága és időigényes.

1.3.2. Szabály alapú modellek

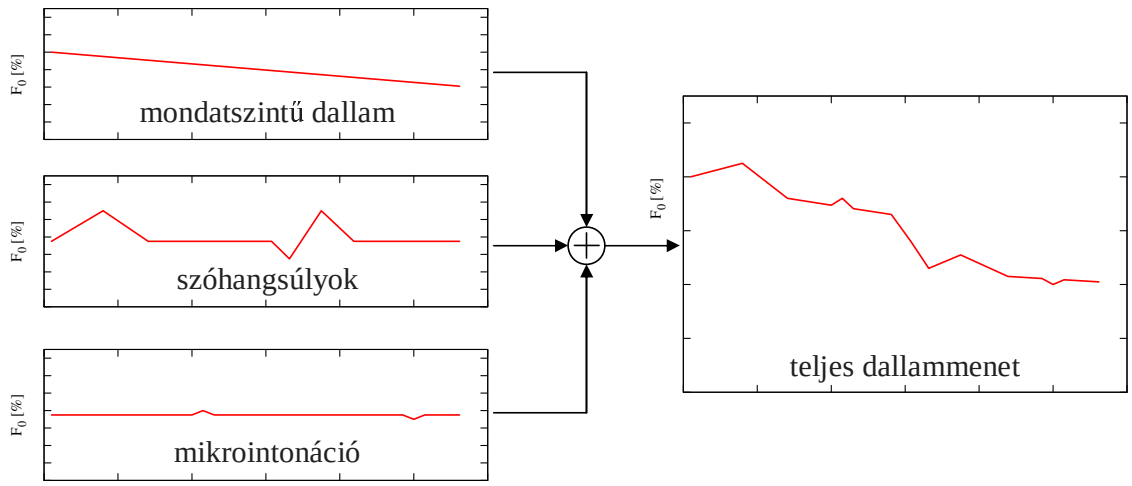
A prozódia modellezése szabályok segítségével is történhet [12, 3. fejezet, 32. oldal]. Ez esetben a szöveg egyes részeihez (mondat, szó, szótag, hang) szabályokat rendelünk, melyek a létrehozandó prozódiát definiálják (pl. hangsúly a mondat elején, alapfrekvencia-csökkentés a mondat végén, hangok megkülönböztetése a környezetük alapján).

A szabályok ember által definiáltak, így megalkotásuk nehéz és időigényes, de hiba esetén könnyen javíthatóak. Hátrány viszont, hogy a természetes nyelvek nem írhatók le tökéletesen szabályok segítségével. A szabály alapú modellek előnye abban rejlik, hogy kiszámíthatóak, azaz mindig hasonló minőségű prozódiát tudnak létrehozni. Azért fontos ez, mert az ember nehezen tűri a változást, ha már egy bizonyos minőséget megszokott (pl. az egyik mondat szépen szól, a következő gyengébb minőségű).

1.3.3. Adatvezérelt modellek

Adatvezérelt[13] módon úgy lehet prozódiai modellt létrehozni, hogy valamilyen nagyméretű beszédkorpuszból megpróbáljuk kinyerni a természetes beszéd tulajdonságait [12, 3. fejezet, 32. oldal]. A korpusz segítségével készíthető olyan rendszer, amely írott szöveg alapján következtetni tud a beszéd akusztikai paramétereire korrelációk, összefüggések keresésével. Így a prozódiát leíró szabályok megalkotása nem kézzel, hanem automatikus módszerekkel történ-

[10] *Tones and Break indices*, magyarul hanglejtés és szünet címkék.
[11] *German Tones and Break indices*.
[12] *International Variation in English*, vagyis az angol nyelv nemzetközi változatai.
[13] Más néven gépi tanulás, angolul *Machine learning*.



1.4. ábra. Szuperpozíciós F_0 modell működése: a szupraszegmentális (felül), szószintű (középen) és szegmentális (alul) dallamot összeadva jön létre a mondat végső dallammenete.

het. Ezen adatvezérelt modellek hátránya, hogy a megfelelő adatbázis elkészítéséhez igen nagy mennyiségű adatot kell kézzel felcímkézni.

Számos gépi tanulási módszert alkalmaztak már a prozódia modellezésére. Erre egy példa Strom megvalósítása, melyben négy CART[14] segítségével sikeresen tudta modellezni a hangsúlyokat, F_0 értékeket [13]. Tao és társai neurális háló alapú prozódiai modellel állították be a szótagok F_0 értékeit, amely így természetesebb mandarin nyelvű beszédet tudott létrehozni korábbi módszerükhöz képest [14].

Az adatvezérelt, gépi tanulás alapú modellek alapfeltétele tehát egy megfelelő méretű, címkézett beszédkorpusz rendelkezésre állása.

1.3.4. Szuperpozíciós modellek

A szuperpozíciós modellek fő jellemzője, hogy a prozódia összetevőinek különböző szintű megvalósításait (pl. mondat-, szó-, hang-szint) adják össze, vagyis szuperponálják egymásra. A szintek modellezése külön-külön történik, például először meghatározva a mondatdallamot (emelkedő, egyenletes, eső), utána a szó- vagy szótagszintű hangsúlyokat (erős, neutrális, negatív), végül a mikrointonációs változásokat, ahogy ez az 1.4. ábrán is látható. Az ábrán F_0-ra bemutatott módszerhez hasonlóan végezhető el a prozódia másik összetevőjének, az intenzitásnak a modellezése, míg az időzítés meghatározása bonyolultabb.

[14] *Classification And Regression Tree*, adatok osztályozására használható regressziós fa.

1.2. táblázat. A prozódiai modellek összehasonlítása.

	Előny	**Hátrány**	**Példa**
Leíró jellegű	egyszerű leírás minimális címkehalmaz	kézi címkézés minden nyelvre más	ToBi, GToBi, IViE
Szabály alapú	javítható szabályok kiszámítható	nyelv változik, szerkezetét nehéz leírni	Profivox
Adatvezérelt	nem kell sok ismeret automatikus	nagyméretű, címkézett korpusz szükséges	neurális háló, CART
Szuperpozíciós	többlépcsős modell jól leírja a nyelvet	bonyolult működés komplex ismeret kell	Profivox, Fujisaki

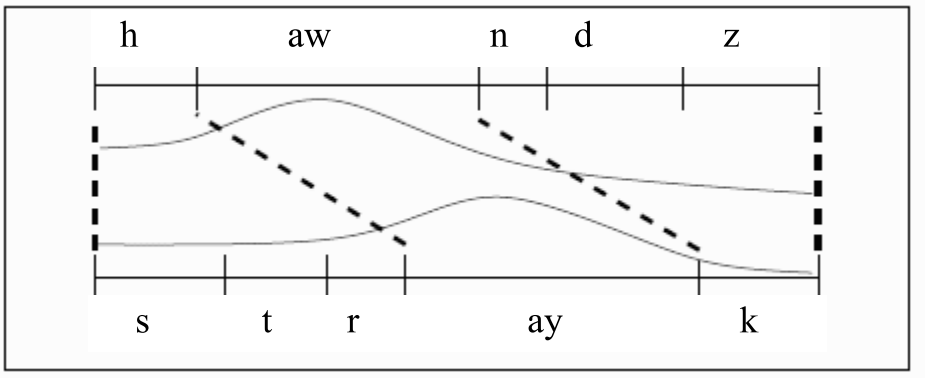
Az egyik konkrét szuperpozíciós megvalósítás Fujisaki modellje, amelyben az F_0-menet a frázisszintű dallam és a hangsúlyok összeadásával jön létre [15].

1.3.5. A prozódiai modellek összehasonlítása

Az 1.3. alfejezetben röviden bemutattuk a beszédszintetizátor rendszerekben használatos prozódiai modellek alapvető típusait. A különböző modellek összehasonlítása az 1.2. táblázatban látható. Az eddigieket összegezve elmondhatjuk, hogy a leíró jellegű modellek ugyan egy egyszerű címkehalmazzal dolgoznak, a gyakorlatban mégis nehézkesen használhatóak. A szabály alapú modellek kiszámítható minőségű prozódiát tudnak létrehozni, de mivel a természetes nyelvek nem reguláris szerkezetűek, nehezen lehet hozzájuk megfelelő szabályokat definiálni. Az adatvezérelt modellek egyértelmű előnye a modellezéshez szükséges paraméterek korpuszból kinyerhetősége, de hátrányuk is ebben van, hiszen a nagyméretű korpuszok kezelése nehéz. A szuperpozíciós modellek valamilyen más modellel együtt alkalmazva jól leírnak egy-egy nyelvet, de megalkotásukhoz komplex ismeret szükséges.

1.4. A korpusz alapú prozódiai modellek

Számos olyan módszer ismert a beszédszintézis szakirodalmában, amely a megfelelő prozódia generálásával foglalkozik, mint azt az előző alfejezetben láthattuk. A következőekben részletesebben bemutatunk néhányat ezek közül, amelyeknek közös jellemzője, hogy az adott bemeneti szöveghez tartozó prozódiát valamilyen természetes beszédből álló korpusz alapján hozzák létre. Az előző alfejezetben már találkozhattunk néhány ilyennel az adatvezérelt modellek között (1.3.3. rész), a jelenlegi alfejezetben pedig ezek kiegészítése következik nem feltétlenül gépi tanulás alapú módszerekkel. Ezekben a rendszerekben az emberihez hasonló dallam-



1.5. ábra. Meron-féle F_0 másolás. Felül: forrás szótag, alul: célszótag. A szaggatott vonalak a szótagok három részre osztását és az F_0-menet másolását jelzik. Forrás: [16].

menet létrehozása azzal garantálható, hogy a szintetizálandó mondat alapfrekvencia-menetét az adatbázisból vett kisebb-nagyobb elemek (pl. szótag, szó) segítségével határozzák meg.

1.4.1. Egyszerű modellek

Az egyszerű modellekben a prozódia meghatározása egy lépésben történik szemben a kombinált megközelítésekkel, ahol ez több szinten valósul meg. A felhasznált elemek tipikusan „fix” méretűek (pl. szótag).

Meron módszere

Meron a korábban elkészített, szabály alapú beszédszintézis rendszert egészítette ki egy olyan modullal, amely a létrehozott prozódia természetességét javítja [16]. Azért őrizte meg a szabály alapú rendszert is, mert ennek előnye a kiszámíthatóság, vagyis mindig hasonló minőségű prozódia készíthető vele. Módszerében tehát egyesíti a szabály alapú eljárások robosztusságát és a korpusz alapú megközelítések természetességét.

A korpusz létrehozásakor automatikus módszerekkel szótagokra bontják azt, és minden egyes szótaghoz három F_0 értéket tárolnak: a magánhangzó előttit, a magánhangzó közepén és utána lévő alapfrekvencia értéket. A szótagokra bontás előnye, hogy szinte tetszőleges bemeneti mondathoz lehet találni prozódiai mintákat. A korábbi módszerekben [17] ugyanis teljes tagmondat alapján történt a keresés, és így előfordult, hogy nem volt megfelelő prozódia-minta. Ha nagy beszédkorpusz áll rendelkezésre, akkor mindkét módszer hasonlóan teljesít, azonban kis adatbázis (268 mondat) használatával egyértelműen Meron konstrukciója hatásosabb. Mivel

a cikkben csak egy kísérleti rendszerről esik szó, nem ismert pontosan, hol van ez a határ az adatbázis méretében.

Az egyes természetes prozódia-darabok keresése a korpusz alapú elemkiválasztásos beszédszintetizátorokban használt módon történik, torzítási és összefűzési költségek használatával. A megtalált darabok összefűzése során jelfeldolgozási módszerek segítségével érik el, hogy a végső dallammenet „sima" legyen. Egy-egy szótag F_0-jának másolása három részletben történik, ahogy az 1.5. ábrán látható. A felső dallammenet az adatbázisbeli forrásszótaghoz, az alsó pedig a szintetizálandó célhoz tartozik, a szaggatott vonalak pedig a szótagok három részre osztását jelzik, középen a magánhangzóval. A másolás szakaszonként lineáris függvények segítségével történik, idővetemítéssel.

Mivel egy-egy írott mondatnak nagyon sokféle szóbeli megvalósítása lehet, a rendszernek döntenie kell arról, hogy milyen értelemmel szintetizálja a mondatot. Azonban a jelentés, érzelmek, és egyéb viselkedésmódok modellezése meglehetősen bonyolult lenne, így ezzel nem is foglalkoznak részletesen. Mivel nem ismert, hogy a TTS-sel a szövegnek melyik szóbeli reprezentációját kellene létrehoznia, ezért egy olyat választanak, amely a módszerrel legtermészetesebben előállítható.

Raux és Black módszere

Raux és Black modellje [18] hasonló az előző megvalósításhoz. A leglényegesebb különbség, hogy ebben az új megközelítésben a prozódia másolásához használt alapegység a szegmens, amely nem más, mint egy szótagon belüli egység. Ez megfelelő flexibilitást ad a rendszernek, és lehetővé teszi a makro- és mikroprozódia másolását is. Így elvileg lehetővé válik, hogy akár egy-egy szótag dallamát is több különböző adatbázisbeli F_0-elemből állítsák össze, azonban a módszer gyakorlatban egyben kezeli a szótagokat.

Módszerüket a Festival beszédszintetizátor rendszerben [19] implementálták, a rendszer elemkiválasztási és -összefűzési lehetőségeit kihasználva. Az adatbázis felcímkézése és szegmensekre osztása automatikusan történt, az elemek csoportosításával egyben.

Ahhoz, hogy az egyes F_0 szegmensek időzítése is megfelelő legyen, időbeli kinyújtásra, illetve összehúzásra is szükség volt. Azt is vizsgálták, hogy jobb lesz-e a szintetizált beszéd minősége, ha az egyes szegmensek között F_0-simítást alkalmaznak. A módszer kiértékelése során az derült ki, hogy az esetek többségében az F_0-szegmensekből létrehozott dallam jobb volt a korábbi, szabály alapú megvalósításnál.

Saito módszere

Saito is természetes F_0 elemekből építi fel a dallammenetet módszerében [20]. Fontosnak tartja, hogy a beszédadatbázisának felépítésekor, valamint a beszéd szintézisekor is minimális legyen a természetes F_0-menetek módosítása. A mondatszintű F_0-görbét egy lineáris-regressziós statisztikai modell segítségével készíti el. Az adatbázisbeli mondatok alapfrekvencia-egységekre tördelése nyelvi információk alapján történik meg, a hangsúlyok vizsgálatával. Ezek az egységek japán nyelv esetén tipikusan a szavak. Az alapfrekvencia-menet tárolása magánhangzónként egy F_0 értékkel történik, amelyet a hang közepén mintavételeznek, hogy elkerülhető legyen a szomszédos mássalhangzók befolyása.

A szintézis során először nyelvi információ (hangsúlytípus, frázishossz, fonémák típusa) alapján az F_0-menet vázlata kerül meghatározásra. Ezután a vázlathoz legjobban illeszkedő F_0 elemek keresése történik meg a beszédadatbázisból. A legpontosabb jelölt meghatározását a hangsúlyok alapján, illetve fonemikus egyezés vizsgálatával végzi a módszer. Az időzítés másolására csak akkor kerül sor, ha az F_0-menettel való szinkronitása biztosítható, ellenkező esetben ugyanis jelentős torzítás alakulhat ki. A szintézis harmadik lépésében a megtalált alapfrekvencia-elemek összefűzése következik. Alapvető cél az adatbázisbeli F_0 módosításának elkerülése, így csak F_0 szint eltolást alkalmaz az algoritmus.

A módszer sikerességének vizsgálatára elvégeztek egy kísérletet, melynek eredményei azt mutatják, hogy a generált dallammenet nagyon hasonlít az emberihez.

Iriondo és társai módszere

Iriondo és társai módszerének újdonsága, hogy az esetalapú következtetést[15] használja a prozódia szöveg alapján történő meghatározásához [21]. A CBR négy lépését a prozódia-modellezés problémájához igazították. A módszer betanítására egy olyan korpuszt használnak, amelyben összekapcsolják a szöveg elemzéséből származó információt a beszédkorpuszbeli prozódiai paraméterekkel. A szintézis során először a hangidőtartamokat állítják be, ezután következik az F_0-görbe meghatározása, amihez időbeli normalizálást is alkalmaznak. Így olyan polinomiális alapfrekvencia-menetet tudnak előállítani az F_0 értékeket a fonémák közepéhez rendelve, amely minimális távolságra van a korpuszbeli mintáktól.

Kísérleteik alapján a módszer jó eredményeket mutat. Megközelítésük egyelőre kezdetleges, de a terveik szerint alkalmas lesz különböző érzelmek szintetizálására is.

[15] Az esetalapú következtetés (angolul *Case-Based Reasoning*, röviden CBR) módszertan a gépi tanulás egyik változata, lényege, hogy az aktuális problémához a múltban keres hasonlóságot, és ez alapján alakítja ki az előrejelzést.

1.4.2. Kombinált, többszintű modellek

Az általános prozódiai modellek között (1.3. alfejezet) bemutatott szuperpozíciós elv a korpusz alapú modellekben is alkalmazható, azaz a szintetizálandó prozódiát először modellezhetjük külön-külön, több szinten, majd ezeket összeadva egy bonyolultabb, szuperpozíciós modellt kapunk.

Dong és Lua módszere

Dong és Lua nevéhez a példa alapú prozódia generálás módszere fűződik [22]. Prozódiai adatbázisukban a valódi beszédből vett mondatokat három részre bontják: mondatszintű prozódia, prozódia-minta frázis szerint és szótagszintű prozódia. A példa-korpusz elkészítése során felmérik az adatbázisbeli szótagok F_0 értékeit, majd hangkörnyezet (előző, jelenlegi és következő szótag) alapján csoportosítják, és egy táblázatban tárolják ezeket.

Egy adott szintetizálandó mondatot először szavakra bontanak, majd szófaj-analízist végeznek rajta. Az egy-egy szótaghoz tartozó dallammenetet és időtartamokat statisztikai módszerek segítségével határozzák meg. A szintézis során a szótagok dallammenetét a korábban elkészített táblázatból keresik ki a hangkörnyezet-információ alapján, külön kezelve a zöngés és zöngétlen részeket. A szótag időtartama is hasonló módon kerül kiszámításra. A frázisszintű prozódia-minták meghatározásához az adatbázisból keresnek hasonló mintát nyelvi szempontok (pl. a szövegből származtatott fonetikai információ) szerint. A generálandó mondat prozódia-mintájának a hozzá legjobban illeszkedő adatbázisbeli mondatot választják. A mondatszintű prozódia általános dallamformák (kijelentő, kérdő stb.) alapján kerül meghatározásra. A végső dallammenetet a mondat-, frázis-, és szótagszintű prozódia-minták kombinációjaként állítják elő. A mondat időzítését, szüneteit is a példa-korpusz alapján valósítják meg.

Van Santen és társai módszere

Van Santen és társainak megközelítésében az az újdonság, hogy a beszédkorpuszban többféle szempont szerint keresnek a bemeneti szöveghez illő prozódia-mintát [23]. Egyrészt egy olyan sorozatot keresnek az adatbázisban, amelynek fonémái a bemenethez hasonlóak.[16] Másrészt több olyan részt próbálnak találni, amelyek prozódia szintjén várhatóan illeszkedni fognak (pl. hasonló a hangsúlyszerkezetük).[17] Ez tipikusan frázis, hangsúly, és fonéma egysé-

[16] A fonemikusan hasonló sorozatokat cikkükben *phonemic unit sequence*-nek hívják. Definíciójuk szerint például a *medal* és a *neighbour* szó fonemikusan hasonló.

[17] A prozódia szintjén illeszkedő sorozatokat *prosodic unit sequence*-nek nevezik.

1.3. táblázat. A korpusz alapú modellek összehasonlítása.

Módszer	F_0 **egység mérete**	**Előny**	**Hátrány**
Meron	szótag	szabály és korpusz alapú vegyítve	sok jelfeldolgozás
Raux és Black	szegmens	flexibilitás mikro- és makroprozódia	teljesen adatvezérelt, nincs kézi beavatkozás
Saito	szó	minimális F_0-módosítás	csak F_0 szint eltolás, nem pontos dallam
Iriondo és társai	prozódiai egység	CBR módszertan alkalmazása	még csak kezdetleges módszer
Dong és Lua	szótag, frázis, mondat	statisztikai módszer alkalmazása	lassú mintakeresés
Van Santen és társai	fonéma, hangsúly, frázis	F_0 és időzítés együttes másolása, kisebb adatbázis	hierarchia szintek összeadása kérdéses

get jelent. Az utóbbi elemsorozatok kombinációjából, jelfeldolgozás segítségével hozzák létre a bemeneti szöveghez tartozó dallammenetet.

A különbség Raux és Black megvalósításához [18] képest egyrészt az, hogy „nyers” F_0 minták összefűzése helyett szuperpozíciós megközelítést alkalmaznak, többszintű prozódiaminta összeadásával, azaz egymásra szuperponálásával. Ezáltal megszűnik az elemkiválasztásos rendszerekben ismert összefűzési hiba, a létrehozott dallammenet folytonos lesz. Másrészt az időtartamok definiálását is a prozódiai illeszkedést alkalmazva végzik el, aminek az az előnye, hogy az alapfrekvenciát és időzítést együtt másolva természetesebb lesz a prozódia szerkezete.

A megvalósításban használt egyik legfontosabb ötlet tehát az F_0-menetek dekompozíciója három részre, amelyekkel más-más környezetben eltérő prozódia valósítható meg. A korábbi elemösszefűzéses rendszerek mesterségesen előírt intonációja helyett természeteset alkalmaznak, az elemkiválasztásos rendszerekhez kapcsolódó előny pedig az adatbázis méretének jelentős csökkenése.

1.4.3. A korpusz alapú modellek összehasonlítása

A korpusz alapú F_0 modellek működése hasonló az elemkiválasztásos beszédszintetizátorok működéséhez, vagyis felvett beszédből származtatott „sablonok” segítségével állítják elő a dallammenetet. Ezeknek az F_0 sablonoknak a mérete határozza meg a rendszer működését. Ha hosszú egységeket használunk, a beszéd szupraszegmentális szerkezete megmarad.

Ugyanakkor ilyen nagy egységből valószínűsíthetően kevés van egy adatbázisban, és így nem feltétlenül található illeszkedő egység egy konkrét mondathoz, és jelfeldolgozással kell kisegíteni a találat hiányát, ami a minőség romlásához vezet.

Az 1.3. táblázatban az itt bemutatott módszerek összehasonlítása látható. Alapvető különbség van köztük az F_0-egységek méretében. Meron munkájában szótagokat, esetleg több szótagot együtt használt fel, ami viszonylag jó dallamösszeállítási lehetőséget biztosít. Raux és Black módszerében egy kisebb egység, a szótagon belüli szegmens került felhasználásra. Saito szavak szintjén végezte el a prozódia másolását, Iriondo és társai pedig prozódiai egységeket alkalmaztak. A hosszú és rövid egységeket kombinálva Dong és Lua módszerükben a mondat-, frázis- és szószintű F_0-minták egymásra szuperponálásával hozta létre a dallammenetet. Van Santen és társai munkájukban szintén három féle egységre bontották a beszédkorpuszbeli mondataikat, és ezek kombinációjával definiálták az alapfrekvencia-menetet.

Meron módszerének előnye, hogy a korábbi szabály alapú eljárások robosztusságát, és a korpusz alapú modellek természetességét egyesíteni tudja, ugyanakkor a sok rövid F_0-minta összefűzésekor alkalmazott jelfeldolgozás ront a rendszer minőségén.

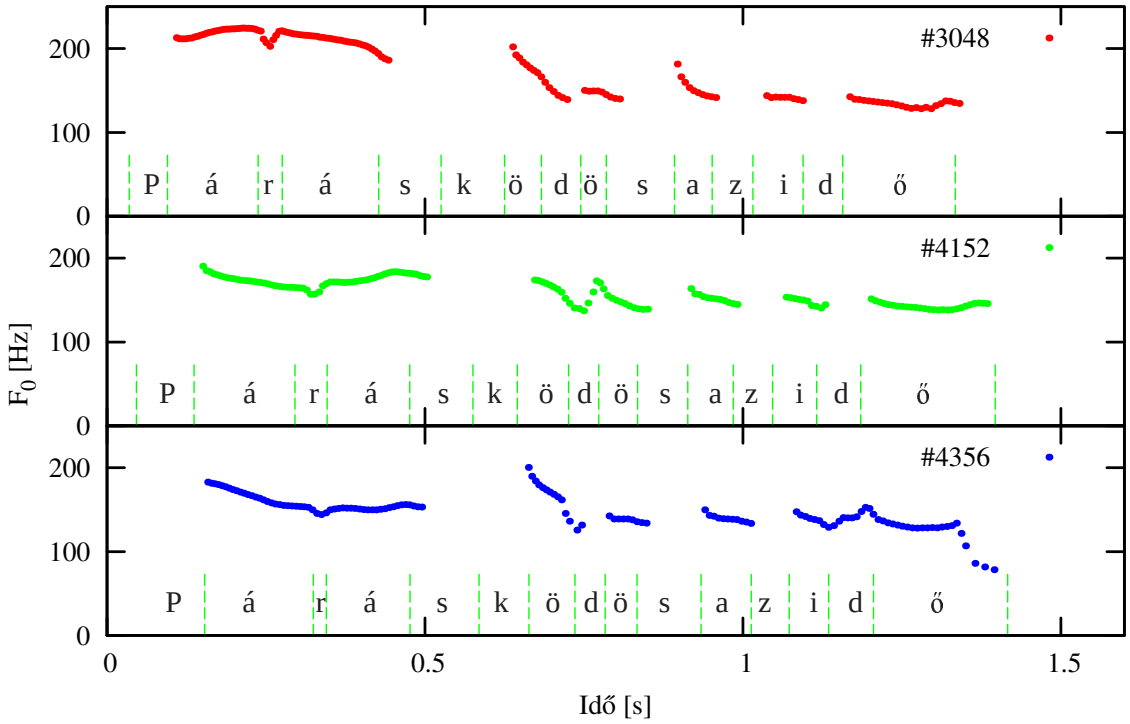
Raux és Black munkája egy teljesen adatvezérelt prozódia generálást eredményezett, amely megfelelő flexibilitásának köszönhetően jól tudja modellezni a beszéd makro- és mikroszegmentális szerkezetét, és növeli a természetességet. Az adatvezéreltségnek köszönhetően a módszer költséghatékony módja a természetes F_0-modell létrehozásának, de ezáltal a rendszer működése teljes mértékben az adatoktól függ, nehéz az esetleges hibák kézi javítása.

Saito módszerének pozitívuma, hogy az F_0-módosítást minimálisra próbálja állítani a torzítások elkerülésének érdekében. Ezáltal viszont a dallam beállítása nem történhet meg pontosan, hiszen csak F_0 szint eltolással nem oldható meg a tökéletes dallammásolás.

Iriondo és társai a mesterséges intelligenciából ismert eset-alapú következtetést alkalmazzák, amely hatékonynak bizonyult, de módszerük még kezdetleges, továbbfejlesztése folyamatban van.

Dong és Lua sikeresen alkalmazták statisztikai alapú módszerüket a prozódia másolására, de szubjektív kísérleteket még nem végeztek ennek ellenőrzésére. A megvalósításuk hátránya, hogy ha nem található minta az adatbázisban, akkor nem megfelelő prozódiát hoz létre a módszer. További probléma, hogy a minta-keresés folyamata meglehetősen lassú, ami egyelőre nem teszi lehetővé a valós idejű beszédszintézist.

Van Santen és társainak módszere is természetesnek tűnő F_0-menetet tud létrehozni a korábbi mesterségeshez képest, és a szükséges számítási kapacitást is jelentősen lecsökkentették. Egyelőre nyitott kérdés, hogy a modellben alkalmazott különböző szintek összeadása mennyire vezet jó eredményre, mert a prozódiának nem lehet minden összetevőjét additívan modellezni.



1.6. ábra. Prozódiai változatosság az emberi beszédben: a *„Párás, ködös az idő.”* mondat három változatának összehasonlítása. A színes görbék az alapfrekvencia-menetet, a függőleges vonalak a hangidőtartamokat mutatják.

1.5. Prozódiai változatosság

A következőkben röviden bemutatjuk, hogy mit is jelent az emberi beszéd változatossága. Ezután arról lesz szó, hogy a beszédszintetizátorokban hogyan merült fel az igény ennek modellezésére, és milyen kezdeti kutatásokat végeztek a témában. Végül bemutatunk néhány vizsgálatot, melyekben az emberi beszéd variáltságát elemezték.

1.5.1. Változatosság az emberi beszédben

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 1.6. ábra mutatja. Az ábrán a *„Párás, ködös az idő.”* mondat három különböző kiejtési módját láthatjuk ugyanazon beszélőtől. A három változat hasonló, de mégis észrevehető különbség van közöttük az alapfrekvencia-menetben és a hangok időtartamában.

1.4. táblázat. A *„Párás, ködös az idő."* mondat három változatának összehasonlítása hangonkénti alapfrekvencia értékek szerint.

	Vált.	**P**	**á**	**r**	**á**	**s**	**k**	**ö**	**d**	**ö**	**s**	**a**	**z**	**i**	**d**	**ő**
F_0	#3048	-	212	213	210	-	-	165	-	149	-	157	-	127	-	129
$[Hz]$	#4152	-	167	161	176	-	-	150	-	155	-	154	-	133	-	137
	#4356	-	165	141	152	-	-	171	-	140	-	140	-	124	-	122

1.5. táblázat. A *„Párás, ködös az idő."* mondat három változatának összehasonlítása hangidőtartam értékek szerint.

	Vált.	**P**	**á**	**r**	**á**	**s**	**k**	**ö**	**d**	**ö**	**s**	**a**	**z**	**i**	**d**	**ő**
t	#3048	60	142	38	151	97	99	57	62	40	107	59	63	79	62	176
$[ms]$	#4152	90	158	50	129	99	69	80	47	61	78	71	62	68	69	211
	#4356	59	171	20	131	108	78	72	47	50	100	78	59	62	70	210

Az 1.4. és az 1.5. táblázatok ezeket a különbségeket mutatják be részletesen: előbbin az 1.6. ábrán látható mondatváltozatok hangonkénti alapfrekvencia értékeit, utóbbin pedig a hangidőtartamokat tanulmányozhatjuk. Feltűnő a különbség a *„párás"* szó F_0-menetén, a beszélő valószínűleg máshogy próbálta hangsúlyozni a mondatot az egyes esetekben.

1.5.2. Változatosság a beszédszintetizátorokban

A legtöbb beszédszintetizátor rendszer az emberrel 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 prozódia-minták 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 1.6. ábrán is látható) lecserélődik a legjobb, leggyakoribb mintára. Ez viszont az emberi fül számára, amely a változékonysághoz szokott, könnyen felismerhető és zavaró. 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.

Új irányzatok a beszédszintézisben

Carlson már 1991-ben jelezte, hogy a beszédszintézisben új irányzatok megjelenése várható [24]. Arról számol be, hogy a változékonysággal korábban csak a beszédfelismerőkben foglalkoztak (hiszen ott is nagyon fontos, hogy egy-egy szöveg különböző bemondott változatait megfelelően tudják modellezni), de a szövegfelolvasókban is fontossá fog ez válni. Pontos elképzelésekről még nem olvashatunk a cikkben, de kérdés szintjén felmerül, hogy milyen esetekben lehet a változékonysággal foglalkozni, és mikor fontosabb az emberi beszédprodukcióban felmerülő kényszerek betartása (például a tüdőnkből kiáramló levegő útja meghatározza, hogy milyen hangokat tudunk létrehozni). Fontosnak tartja a beszédstílusok alkalmazását, amivel lehetővé válna, hogy dinamikusan változtassuk az időzítést, a szünetstratégiát és az intonációt, így a beszélők közötti különbségek modellezése is megtörténhetne. Cikkének befejezésében megemlíti, hogy még sok akadálya van a változatosság megvalósításának beszédszintetizátorokban.

Véletlen-e a beszéd változatossága?

Dutoit szintén arról ír átfogó munkájában, hogy a szintetizált beszéd változatosságával eddig keveset foglalkoztak, mert ezt nehezebb megvalósítani, mint elsőre gondolnánk [25]. Ennek oka az lehet, hogy a változékonyság fogalma nem csupán véletlenszerűséget jelent, hanem azt kell megtalálni, hogy a beszélő hogyan variál. Az emberi beszédben ugyanis a véletlennek tűnő jelenségek koherensek egymással, amit figyelembe kell venni ilyen rendszer tervezésekor.

A beszéd szünethosszainak variáltsága

Zvonik és Cummins az emberi beszéd szünethosszait és ezek változatosságát vizsgálták [26]. Cikkükben azt állítják, hogy a különböző beszélők szünetstratégiájában hatalmas eltérések mutatkoznak. A folyamatos beszéd során különbség van például a professzionális bemondók és az átlagos beszélők között a szünetek elhelyezésében. Nyelvenként is vannak megfigyelhető eltérések: az olaszban rövidebb, míg a spanyolban hosszabb az átlagos szünethossz. Ha viszont több beszélő egyszerre beszél (cikkükben ezt szinkron beszédnek nevezik), akkor a folyamatos egymásra figyelés miatt a szünetek variáltsága nagymértékben lecsökken. Összességében elmondható, hogy keveset tudunk még az emberi beszéd szünetidőtartamait meghatározó tényezőkről.

Társalgások vizsgálata

Campbell is említést tesz arról, hogy az emberi beszéd változékonysága nem véletlenszerű, de a mindennapi beszélgetések rendkívül változatosak [27]. A beszédszintetizátorokban alkalmazott adatbázisoknak viszont az a tipikus jellemzőjük, hogy hangstúdióban felvett, gondosan megtervezett felolvasott beszédet tartalmaznak. Emiatt az emberek közti társalgás során használt változatos beszédből csak nagyon kevés jellemző található bennük, így kérdéses, hogy mennyire alkalmazhatóak ezek például beszédstílusok megvalósítására.

1.5.3. Kísérlet a változatosság elemzésére két párhuzamos korpuszon

Chu és társai a beszéd variáltságával foglalkoznak munkájukban [28]. 1000 mondatot kétszer felvettek, 6 hónap különbséggel, és azt vizsgálták, hogy az egyező mondatoknak mennyire hasonló illetve eltérő a prozódiája.

Kutatásuk szerint a prozódiát fel lehet bontani invariáns és változó részekre. Az invariáns jellemzők közé tartozik például a frázishatár előtti időbeli nyújtás, a hangsúlyok és F_0 emelés/csökkentés összefüggése. A prozódia változékonysága két típusú lehet. Az első esetben, amelyet JND[18]-nek is hívnak, szinte észrevehetetlenek a prozódiabeli változások, míg a második esetben észrevehetjük például a dallambeli változást, de ez nem módosítja az átviendő gondolat értelmét. Ezek alapján tehát egy-egy szövegnek sokféle prozódiai megfelelője lehet.

Chu és társai a két adatbázis párhuzamos vizsgálata során azt vették észre, hogy a mandarin beszéd ritmusszerkezete stabilnak számít, mivel a beszélő fél év elteltével is hasonló ritmusstratégiát alkalmazott. Az egyes szótagok átlagos alapfrekvenciája és időtartama között viszont jelentősebb volt a különbség.

A szerzők cikkükben bemutatnak egy beszédszintetizátor rendszert, amely megkísérli a prozódiai változatosság létrehozását. A módszer célja, hogy ne mindig csak a legjobb lehetőséget keresse meg, hanem a rossz esetek kihagyásával a maradékból véletlenszerűen válasszon. A megközelítés sikeresnek bizonyult, és használható az angol illetve mandarin nyelv szintézisére.

Újabb kísérleteik során kifejlesztettek egy rendszert, amely egy elemkiválasztásos beszédszintetizátor természetellenes prozódia észleléssel [29]. Ennek segítségével elkerülhető a korpusz alapú rendszerekben sokszor előforduló összefűzési hiba, ugyanakkor valamilyen mértékben megvalósítható a prozódia változatossága.

[18] *Just Noticeable Differences*, vagyis éppen észrevehető különbségek.

1.5.4. Magyar nyelvű kijelentő mondatok vizsgálata

A korpusz alapú beszédszintézis nyelvi, fonetikai kérdéseinek elemzése során Olaszy vizsgálatokat végzett beszédadatbázisokon, többek között kijelentő mondatok alapfrekvencia-menetének tanulmányozásával [30].

A mintamondatok mindegyikén jellemezte az alapfrekvencia változást annak töréspontjaival. Az elemzés szerint a kijelentő mondat F_0-menetében változást okoz a mondathangsúly helye, a szó hangsúlyos volta, a hangsúlyos szavak helye a mondatban, valamint a prozódiai egységek határai. Az F_0 általában a mondat első hangsúlyos szótagján a legmagasabb értékű, illetve amennyiben van mondathangsúly, akkor az a legmagasabb. A hangsúlyos szavak első szótagjában F_0 emelkedés található, majd a második szótagban visszacsökkenés tapasztalható. Minél távolabb vagyunk a mondat elejétől, annál kevésbé emelkedik ki a hangsúlyos szótagok alapfrekvenciája. A hangsúlyok közötti részeken az F_0 enyhe esést mutat, azonban ezt a tendenciát megváltoztathatják a prozódiai egységek határai, illetve a mondathangsúly. Ilyenkor nem esés, hanem szintentartás vagy enyhe emelkedés következik be.

Az F_0 szórása a kijelentő mondatoknál meglehetősen nagy. A mondatokban nem lehet jellemző F_0 karakterisztikát találni, ami a mondat belseji hangsúlyok más-más elhelyezkedéséből fakad. A mondat elejére ki lehet mondani, hogy magasabb alapfrekvenciával rendelkezik, mint a mondat vége. Az egyedüli egységes pont, ami minden kijelentő mondat esetén hasonló, a mondat végének F_0 értéke.

Látható tehát, hogy a magyar nyelvű kijelentő mondatok olyan nagy változatosságot mutatnak, hogy nem lehet őket egyszerű sémával leírni. A különbségek a hangsúlyok eltérő elhelyezése mellett abból is fakadnak, hogy a beszélő különböző helyzetekben más-más jelentést akar kifejezni, amihez a prozódia variálása nagy segítséget nyújt.

2. fejezet

Prozódiai változatosságot biztosító rendszer tervezése

A változatosság több forrás szerint is szükségessé válik a mai beszédszintetizátor rendszerekben, ahogy az előző fejezetben bemutattuk ([24], [25], [26], [27] és [28]). Eddig azonban keveset foglalkoztak a természetes beszéd ezen tulajdonságának átültetésével a mesterséges beszédre, így ideális téma kutatásaink számára.

Célunk egy olyan modul létrehozása volt, amely beszédszintetizátorhoz kapcsolható, és a prozódia tervezésének lépése során az emberi beszéd változatosságát utánozza. Ez oly módon valósítható meg, hogy egy-egy bemeneti mondathoz a rendszer több különböző prozódiájú változatot is elő tud állítani, amelyek közül szintéziskor egyet véletlenszerűen választ. Így megoldható az, hogy ugyanazon mondat máshogy szóljon többszöri szintetizálás során, azaz csökkenthető a monotonitás.

A különböző prozódiai minták keresését a korábban ismertetett korpusz alapú modellek működéséhez (1.4. alfejezet) hasonlóan terveztük megoldani. Nagyméretű beszédkorpuszt felhasználva, a bemeneti szöveghez a korpuszból mintát keresve megoldható, hogy egy-egy mondathoz több, eltérő F_0-menet- vagy időzítés-alternatívát állítson elő a módszer.

A fejezet a következő részekből áll: a 2.1. alfejezet bemutatja, hogy milyen követelményeknek kell eleget tennie a megvalósítandó rendszernek ahhoz, hogy alkalmas legyen változatos beszéd létrehozására. Ezután bemutatjuk a lehetséges megoldási alternatívákat a 2.2. alfejezetben. Közülük a kiválasztott módszer megvalósításának tervét tartalmazza a 2.3. alfejezet. Végül ismertetjük a tervezett rendszer erősségeit, gyengéit és működési korlátait (2.4. alfejezet).

2.1. táblázat. Lefedettségi arányok összehasonlítása *„Az említett magasnyomású övtől északra és délre elhelyezkedő ciklonok területén sok a felhő, többfelé esik az eső, az északi tájakon helyenként havazik."* mondat példáján, ha azt feltételezzük, hogy csak az utolsó frázist tudjuk változatosan megvalósítani.

Mért egység	**Összes**	**Talált**	**Lefedettség**
Mondat	1	0	0,00%
Frázis	3	1	33,33%
Szó	22	5	22,73%
Szótag	54	13	24,07%
Hang	122	30	34,59%

2.1. Követelmények a rendszerrel szemben

A rendszer tervezését annak meghatározásával kezdtük, hogy mit várunk el a megvalósítandó beszédszintetizátor-modultól. Nem feltétlenül szükséges minden szintetizált mondatot „változatos" prozódiával létrehozni, a monotonitás csökkentéséhez az is elegendő, ha a mesterségesen előállított beszéd egy része változatos.

2.1.1. Lefedettségi arány

Ahhoz, hogy mérhető legyen az eredeti, szabály alapú prozódiai modellhez képest a változatosság, definiáltunk egy „lefedettségi arányt", amely megmutatja, hogy a szintetizált beszédnek mekkora része készült az új módszerrel. A lefedettséget sokféleképpen lehet mérni, különböző egységekre (pl. mondat, szó) számolhatjuk a találati arányt, vagyis hogy mekkora részhez találtunk változatos prozódiát. Legegyszerűbb százalékos formában kifejezni ezt a következő módon:

$$\text{Lefedettség} = \frac{\text{Talált egységek száma}}{\text{Összes egység száma}} \tag{2.1}$$

Vegyük észre, hogy különböző egységeket alkalmazva eltérő találati arányokat kaphatunk, ahogy ez a 2.1. táblázatban látható. A vizsgált egység méretétől függően nagy eltérések lehetnek a lefedettség arányában. Nem mindegy tehát, hogy mondat, frázis, szó, szótag vagy hang szintjén vizsgáljuk a lefedettséget. A továbbiakban a lefedettségen szótagszámok alapján mért értékeket értünk.

A rendszer működése során először kisebb (kb. 50%) lefedettség elérését céloztuk meg, hiszen ezzel az lenne elérhető, hogy az összes szintetizált beszéd fele természetesebben szóljon, ami már észrevehető változás.

2.1.2. Változatok száma

A prozódiai változatosság eléréséhez az szükséges, hogy egy-egy mondatot ne mindig azonos hanggal mondjon a rendszer. A kérdés az, hogy hány különböző prozódiájú változatot célszerű egy mondathoz rendelni annak érdekében, hogy ezek közül véletlenszerűen választva ne mindig ugyanúgy szóljon a szintetizált beszéd.

Érdemes lenne megvizsgálni, hogy természetes beszédben mennyire különböztethető meg, ha egy-egy mondat eltérő realizációit halljuk, hiszen az eltérés ezek között jelentkezhet a dallamban, ritmusban, hangsúlyban külön-külön vagy egyszerre is. Azt is fontos tudni, hogy egy beszélő egy adott mondatot hány nagyjából különböző változattal mond. Ezeket a vizsgálatokat természetes beszédből felvett korpuszon lehet elvégezni, de ekkor az azonos mondatok különböző változatainak számát a korpusz mérete is befolyásolja (hiszen nem lehet „végtelen” terjedelmű korpuszt vizsgálni). Az 1.5. alfejezetben már volt erről szó, az 1.6. ábra egy korpuszból származó mondat különböző változatait mutatja.

A mesterséges beszédet megvalósító rendszerünkben a valódi beszédnél kevesebb prozódia-változat is elegendő lehet jó minőség eléréséhez. Kezdetben az lenne az ideális, ha legalább 3–4 lehetséges prozódiát tudnánk hozzárendelni egy-egy bemeneti mondathoz.

2.1.3. Futásidő

Nagyon fontos, hogy olyan sebességgel kell működnie a változatosságot megvalósító módszernek, amivel nem nő meg jelentősen a jelenlegi beszédszintézis rendszer futási ideje. Ahhoz tehát, hogy közel valós időben működjön a beszédszintetizátor, megfelelő gyorsaságú algoritmusok választása szükséges.

2.2. Megvalósítási lehetőségek

A célunk eléréséhez megfelelő beszédszintetizátor-technológiát kell választani, amely lehetővé teszi a prozódia szabad vezérlését. A változatosság többféle prozódiai modell kiegészítésével is megoldható lenne. Ez általában azt jelenti, hogy a meglévő prozódiai modellhez egy olyan modult kapcsolunk, amely véletlenszerűvé teszi a prozódia szöveghez rendelését.

A lehetséges megoldások számbevétele során kiválaszthatjuk a célunk eléréséhez legalkalmasabbat.

2.2.1. Beszédszintetizátor-technológia

Az 1.2. alfejezetben bemutattuk a beszédszintetizátorok főbb generációit. Ezek közül a formánsszintézist nehézkes paraméter-beállítása miatt nem célszerű alkalmazni, a rejtett Markov modell alapú szintézis pedig még nem eléggé kiforrott technológia.

Elemösszefűzés

Az elemösszefűzéses rendszer alkalmazásának előnye, hogy ez már régóta használatban lévő, jól működő technológia. A létrehozott beszéd tökéletesen érthető, bár nem teljesen természetes. A tervezett módszerünk szempontjából fontos, hogy a prozódia bizonyos korlátok között (pl. az elemek F_0-jának módosítása csak 30%-kal lehetséges a jelenleg alkalmazott algoritmusokkal) ugyan, de szabadon vezérelhető. Könnyen beállítható tehát a szintetizálandó beszédhez tartozó dallam, ritmus és a hangsúlyok, amelyeket a beszédszintetizátor az elemeken végzett jelfeldolgozás segítségével realizál.

Korpusz alapú elemkiválasztás

A korpusz alapú, elemkiválasztásos rendszerekkel az emberihez nagyon hasonló beszéd előállítása is megoldható. Mivel az összefűzött elemek természetes beszédből felvettek és viszonylag hosszúak (szavak, vagy akár szókapcsolatok), a jó minőségű hang előállítása biztosított. A prozódiát viszont nem lehet olyan szabadon beállítani, mint az elemösszefűzéses rendszerben. A prozódia a korpusz felvételekor eldőlt, és nem célszerű módosítani. Jelfeldolgozási módszerek segítségével megoldható lenne a prozódia beállítása, de ez valamilyen mértékben rontaná a felvett beszéd minőségét, ami elkerülendő.

Az elemkiválasztásos beszédszintetizátorok működéséből következik, hogy mivel nagyméretű beszédkorpusz áll rendelkezésre, ezekben egy-egy szó gyakrabban is előfordulhat. A szó különböző változataiból a beszédszintézis során a legjobban illeszkedőket választják ki és fűzik össze. Ahhoz, hogy változatosságot lehessen megvalósítani ebben a rendszerben (azaz egy adott bemeneti mondathoz ne mindig ugyanaz legyen a kimenet) az lenne szükséges, hogy a módszer összefűzési költségébe beavatkozzunk.

2.2.2. Alkalmazott prozódiai modell

Az 1.3. és az 1.4. alfejezetekben ismertetésre kerültek a beszédszintetizátorokban alkalmazott prozódiai modellek. Ezek közül a szabály alapú, valamint az adatvezérelt (és a korpusz alapú) alkalmas arra, hogy könnyen lehessen vele változatos mesterséges beszédet létrehozni.

Szabály alapú modell

Legegyszerűbb megoldásnak az kínálkozik, hogyha az eddig jól bevált, szabály alapú modellt egészítenénk ki. Megoldható lenne ez úgy, hogy a korábban alkalmazott fix szabályok (pl. a mondat utolsó szótagjában az alapfrekvencia legyen 65 Hz) helyett olyan szabályokat alkalmaznánk, amelyek kis perturbációkat is megengednek (pl. a mondat utolsó szótagjában az alapfrekvencia legyen 60–70 Hz közötti).

Azonban a prozódiát megfelelően leíró szabályok kézi megalkotása nehéz, így azt is nehéz meghatározni, hogy mekkora intervallumon belül lehet szabadon változtatni a paramétereket. A szabály alapú modell ilyen „véletlen” kiegészítésével tehát egyáltalán nem garantálható, hogy természeteshez hasonló beszédet lehetne létrehozni.

Adatvezérelt modell

Az előzőnél jobb megoldás, ha adatvezérelt modellt alkalmazunk. A nagyméretű beszédkorpusz felhasználása ugyanis magában foglalja azt is, hogy a korpuszban lévő véletlen változásokat át tudjuk ültetni a szintetizálandó szövegre.

Így tehát az a fontos, hogy az adatvezérelt modell ne mindig csak a legjobb lehetséges prozódiát keresse meg, hanem több lehetőség közül véletlenszerűen válasszon. Mivel az egyes eltérő prozódiájú változatok mind adatvezérelt módon, a természetes beszéd tulajdonságait másolva jönnek létre, a természetesség a módszer működéséből adódik.

2.3. A megvalósítandó beszédszintetizátor rendszer terve

A különböző lehetőségek áttekintése után tehát kiválasztottuk, hogy milyen rendszert fogunk létrehozni a prozódiai változatosság megvalósítására. Elsődleges szempont volt, hogy a módszer kb. 50%-os lefedettséget tudjon elérni, és minden lefedett mondathoz elő tudjon állítani legalább három lehetséges változatot. Erre leginkább egy elemösszefűzéses beszédszintetizátor alkalmas adatvezérelt modellel, amely kiegészíthető oly módon, hogy nemdeterminisztikusan válassza meg a prozódiát.

2.3.1. Prozódia-minta adatbázis létrehozása

Az adatvezérelt modellhez legelőször egy nagyméretű beszédkorpusz szükséges, amelyet a rendszer működéséhez fel kell dolgozni. A „nyers" beszédet (hangot és a szöveges átírást) át kell alakítani olyan formátumba, amely alapján könnyen megvalósítható a prozódia adatvezérelt módon történő szöveghez rendelése. A korpusz feldolgozásának tervezett lépései a 2.1. ábrán láthatóak. A beszédkorpuszt mondatonként vizsgálva először fel kell bontani kisebb egységekre, ezután a szótagokat kell megszámlálni a későbbi felhasználás céljából. A természetes beszéd F_0- és időtartam-értékei, valamint hangsúlyszerkezete alapján létrehozható egy „prozódia-minta adatbázis", amelynek segítségével a prozódia megfelelő beállítása lehetővé válik.

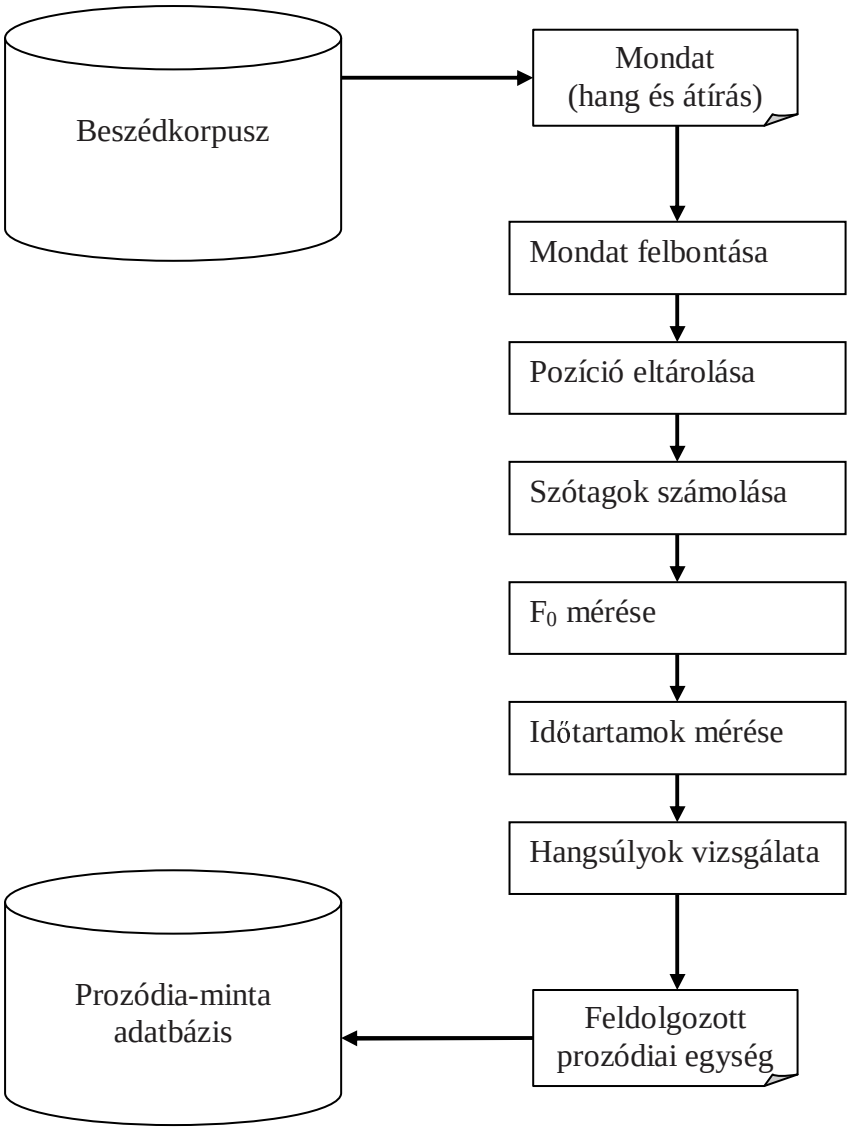
2.3.2. A rendszer működése

Az 1.2. ábrán láthattuk, hogy egy szövegfelolvasó először szimbolikus információt hoz létre a bemeneti szövegből. Ez a szimbolikus információ tartalmazza a szintézis során megvalósítandó prozódiát is. Mivel módszerünk egy TTS kiegészítése, ebből a szimbolikus információból indul ki. A tervezett működést a 2.2. ábra mutatja be részletesen.

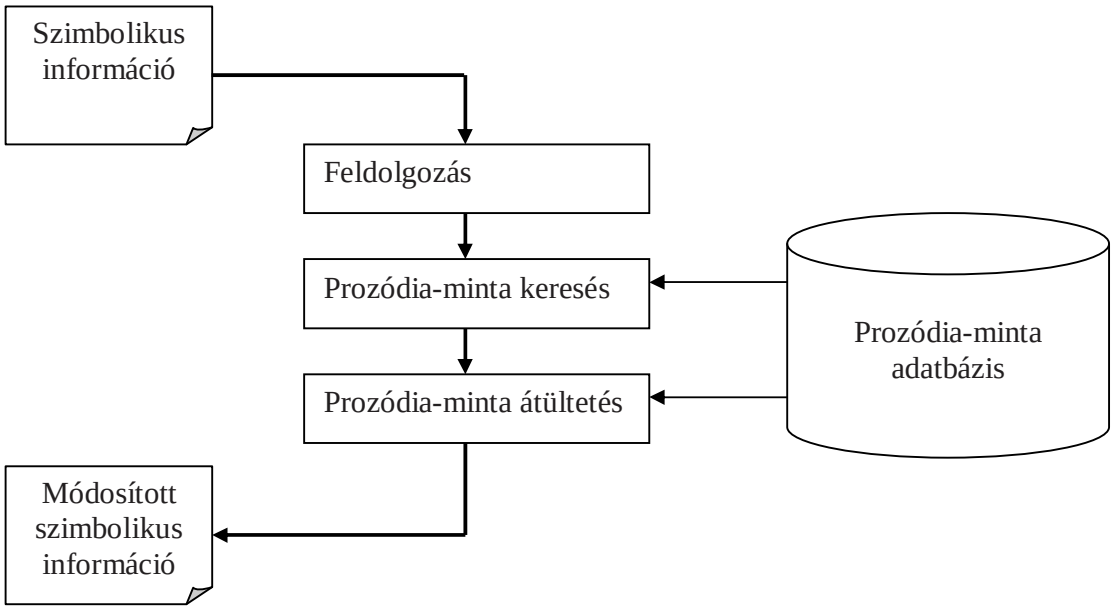
A beszédszintetizátor által az előfeldolgozás során létrehozott szimbolikus információból kiindulva különböző lépéseken keresztül történik meg a prozódia szöveghez rendelése. Ezt úgy kívántuk megoldani, hogy a prozódia-minták felhasználásával, valamilyen hasonlósági mérték alapján lehessen a dallamot, ritmust és hangsúlyokat az adatbázisból a szöveghez másolni. A hasonlósági mértéknek a mondatok szótagszerkezetét terveztük felhasználni. Ez nem más, mint a mondatokban lévő szavak szótagszámai (pl. a *„Hazánkban szerdán folytatódik a párás, fülledt idő."* mondat szótagszerkezete: $3 + 2 + 4 + 1 + 2 + 2 + 2$). A módszer tehát a bemenethez hasonlót keres a prozódia-minták közül szótagszám alapján, majd ezt a bemeneti szimbolikus információhoz társítja. A módosított szimbolikus információ további feldolgozása pedig a beszédszintetizátorban folytatódik.

2.3.3. Illesztés szövegfelolvasóhoz

A megvalósítandó, változatosságért felelős modult egy már meglévő szövegfelolvasóhoz kell illeszteni. Erre a legkézenfekvőbb megoldás, ha a beszédszintetizátor működése során létrejött szimbolikus információn keresztül módosítjuk a prozódiát. A szimbolikus információ egy adatmátrix, amelyben minden hangra beállíthatóak bizonyos paraméterek (pl. F_0, hangidőtartam).



2.1. ábra. Beszédkorpuszból prozódia-minta adatbázis létrehozásának terve. A korpuszt mondatonként vizsgálva, a megfelelő paramétereket (mondaton belüli pozíció, szótagszerkezet, F_0, időtartam és hangsúly) eltárolva előállítható az adatbázis, amelyet a prozódia szöveghez rendelése során lehet felhasználni.



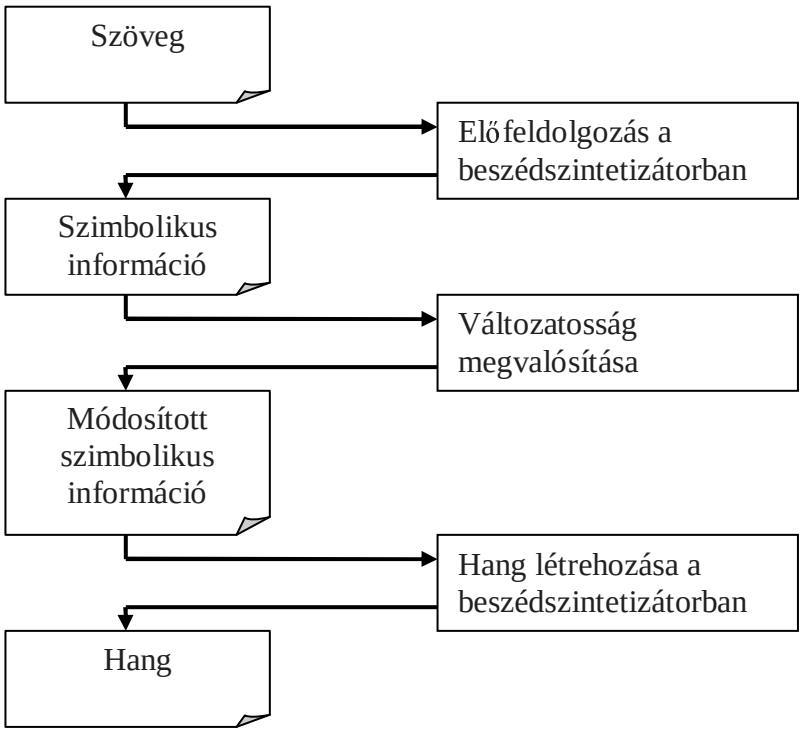
2.2. ábra. A változatosságot megvalósító rendszer terve. A szimbolikus információ alapján, a prozódia-minta adatbázisból hasonló mondat keresésével, és a megtalált mondat paramétereinek átültetésével valósul meg a prozódia megalkotása.

A 2.3. ábrán a TTS-hez illesztés tervét láthatjuk. A beszédszintetizátor hozza létre a bemeneti szöveg alapján a szimbolikus információt, amelyet módszerünk a változatosság igényeinek megfelelően módosít. Végül a beszédszintetizátor a módosított szimbolikus információt alakítja beszéddé, amely várhatóan kevésbé lesz monoton a korábbi megvalósításoknál.

2.4. A tervezett módszer előnyei, hátrányai és korlátai

A sokféle lehetőség közül kiválasztott beszédszintetizátor-modul számos előnnyel és hátrányokkal is rendelkezik. Mivel elemösszefűzéses beszédszintetizátorban alkalmazzuk, a létrehozott beszéd jó minőségű lesz. A módszer elvileg kisebb változtatásokkal használható lesz elemkiválasztásos TTS-ben is.

Az alkalmazandó prozódiai modell adatvezérelt jellegű, amelynek segítségével a természetes beszéd prozódiája másolható. A másoláshoz először a bemenethez hasonló szöveget kell keresni a prozódia-minta adatbázisban. Ez a szótagszerkezet mint hasonlósági mérték szerint történik, mivel az a feltételezésünk, hogy hasonló szótagszerkezet egyben hasonló dallamot, illetve hangsúlyszerkezetet jelent. Az esetek többségében ez fennáll, de sajnos előfordulhat az is,



2.3. ábra. Beszédszintetizátorhoz illesztés terve. A beszédszintetizátor a bemeneti szöveg alapján szimbolikus információt hoz létre. Ezt módszerünk feldolgozza és módosítja, amelyet végül a TTS beszéddé tud alakítani.

hogy a hangsúlyok helyét rosszul választja meg a módszer. Adatvezérelt modellről lévén szó, a kézi beavatkozás, vagyis az ilyen esetek javítása nehezen megoldható.

A beszédkorpusz mérete nagyban befolyásolja a módszer működését és hatékonyságát. Megfelelő témájú, nagy méretű korpusz felhasználásával jó lefedettség érhető el, és várhatóan előállítható lesz elég különböző prozódiájú változat egy-egy mondathoz. Az ilyen korpuszok létrehozása azonban drága és időigényes, amennyiben új témakörre van szükség.

Összességében elmondhatjuk, hogy a tervezett rendszer csökkenteni fogja a mesterségesen létrehozott beszéd monotonitását bizonyos feltételek mellett.

3. fejezet

Prozódiai változatosságot biztosító rendszer megvalósítása

A 2. fejezetben bemutattuk a prozódiai változatosság megvalósításának lehetőségeit, melyek közül egyet kiválasztottunk megvalósításra. A jelenlegi fejezetben az implementáció részletes bemutatása következik.

Először egy korábbi kutatásról írunk (3.1. alfejezet), majd a munkánk során vizsgált beszédkorpuszok szerkezetét és feldolgozását mutatjuk be (3.2. alfejezet). Ezután a 3.3. alfejezetben ismertetjük a változatos beszéd létrehozására alkalmas módszerünket, melyet a tanszéken fejlesztett beszédszintetizátor rendszerbe illesztettünk (3.4. alfejezet).

3.1. Megvalósíthatósági teszt

Annak eldöntésére, hogy a kiválasztott módszer használható-e a gyakorlatban, korábbi kutatásunk során már végeztünk megvalósíthatósági tesztet [31]. Ennek célja az volt, hogy megtudjuk, elvégezhető-e a természetes beszéd dallamának „másolása”. Manuális és félautomatikus módszerekkel létrehoztunk néhány szintetizált mondatot, amelyek az emberi beszédet próbálták utánozni. Az ily módon előállított mondatokat egy teszt keretében összehasonlítottuk a Profivoxban alkalmazott szabály alapú prozódiai modellel. A teszt eredményéből az derült ki, hogy a dallammásolás megvalósítható, így tehát érdemes automatikus módszert létrehozni, mely szövegfelolvasóba illesztett modulként csökkenti a monotonitást.

3.2. Felhasznált beszédkorpuszok

Korábban már többször említésre került, hogy az adatvezérelt, korpusz alapú prozódiai modellek működéséhez beszédkorpuszra van szükség. Az ilyen korpuszok természetes beszédből felvett mondatokat tartalmaznak különböző átírásokkal.

3.2.1. Beszédkorpuszok bemutatása

A dolgozatban használt beszédkorpuszokat a BME TMIT bocsátotta rendelkezésünkre, mivel ezek korábbi kutatásokhoz készültek. A korpuszok szerkezete Fék és társai munkája alapján kerül bemutatásra [2]. A természetes beszédből felvett hanganyagok professzionális bemondótól származnak, és magyar nyelvű kijelentő, valamint felszólító mondatokból állnak.

Egy-egy mondathoz a következő részek tartoznak:

- hullámforma
- szöveges átírás
- fonetikus átírás a hang-, szó- és szünethatárokkal
- zöngeperiódus-határok

A korpuszok létrehozásakor a kiindulási egységek a mondatokhoz tartozó, természetes beszédet tartalmazó hullámforma fájlok voltak, ezek címkézése automatikusab történt. Minden mondathoz tartozik szöveges átírás, amely a felolvasott szöveget tartalmazza. Ebből lehetett létrehozni a magyar nyelv megfelelő hasonulási szabályainak segítségével a fonetikus átírást, amelyben a kiejtett fonémák jelennek meg. A korpuszbeli mondatok hang-, szó- és szünethatárainak jelölését egy beszédfelismerő program végezte el automatikusan [32]. A zöngeperiódushatárok, vagyis a mondat F_0-menetének meghatározása a Praat fonetikai beszéd-analizátor programban implementált alapfrekvencia-detektálás alapján történt [33].

A következőkben az egyes gyűjteményeket mutatjuk be röviden. A 3.1. táblázat ezen korpuszokból jelenít meg egy-egy példamondatot.

Időjárás

Az „Időjárás” korpusz 5232 időjárás-előrejelzés témájú mondatot tartalmaz. A hanganyag összeválogatása olyan célból történt, hogy azt egy időjárás-előrejelzés témakörben alkalmazható elemkiválasztásos beszédszintetizátorban használni lehessen. Emiatt a mondatok hosszúsága nem az általános beszédre jellemző, sok esetben a mondathossz meghaladja az ötven szótagot is.

3.1. táblázat. Példamondatok a felhasznált beszédkorpuszokból.

Korpusz	Példamondat
Időjárás	*„Az elkövetkező napokban is igen meleg igazi júliusi időre számíthatunk, és egészen csütörtökig nem nagyon várható csapadék, de főleg a hajnali órákban néhol lehet zápor, zivatar."*
Harang	*„Ezen a héten minden délben a deszki Magyarok Nagyasszonya plébánia templom harangja szól a Kossuth rádióban."*
Árlista	*„Tájékoztatjuk, hogy az árlisták kettőezerhét január tizenötödikétől érvényesek!"*
Prompt	*„Szíves türelmét és megértését köszönjük!"*
Vonat	*„Az első vágányon tolatást végeznek."*

Harang

Ezen korpuszhoz a hanganyagot a BME TMIT bocsátotta rendelkezésünkre, azonban a felcímkézést nekünk kellett elvégeznünk. A beszédkorpusz témáját tekintve harangokról szól. A Kossuth rádióban minden héten változtatják a déli harangszót, és ezzel kapcsolatosan egy rövid összefoglalót olvasnak fel az adott harang történetéről. Négy harangismertetést, összesen 50 mondatot használtunk fel a korpusz létrehozásához.

Első feladatként az összefoglalókat mondatokra bontottuk, majd a beszédet visszaalakítottuk szöveges formába, vagyis leírtuk a hallott szöveget. Így a fonetikus átírás már elvégezhető volt, amely alapján a BME TMIT-en kifejlesztett beszédfelismerőt [32] kényszerített módban alkalmazva létrehoztuk a mondatok elemhatárait leíró címke fájlokat. A zöngeperiódus-határok jelölése is a korábbiakhoz hasonlóan automatikus módszerrel történt.

Árlista

Az „Árlista" beszédkorpusz jellemzője, hogy olyan mondatokat tartalmaz, amelyekben sok számjegy felolvasása fordul elő. Összesen 3436 viszonylag rövid mondatból áll ez a gyűjtemény.

Prompt

A 3749 mondat között sok ismétlődő található a „Prompt" korpuszban, ami alkalmassá teszi ezt arra, hogy az emberi beszéd változatosságát vizsgáljuk benne. Egy-egy szövegrészlethez több különböző szóbeli realizáció tartozik ugyanattól a bemondótól, így megkereshetőek a különbségek és hasonlóságok a dallamban, ritmusban és a hangsúlyozásban.

3.2. táblázat. A beszédkorpuszok méretének összehasonlítása (mondatok és frázisok szerint).

Korpusz	Mondatok száma	Frázisok száma	Átlagos frázisszám
Időjárás	5232	12827	2,45
Harang	50	72	1,44
Árlista	3436	7841	2,28
Prompt	3749	9208	2,46
Vonat	515	912	1,77

Vonat

A „Vonat” beszédkorpusz egy vasúti utastájékoztató rendszer mondatait tartalmazza. 515 mondattal ez a második legkisebb felhasznált korpusz.

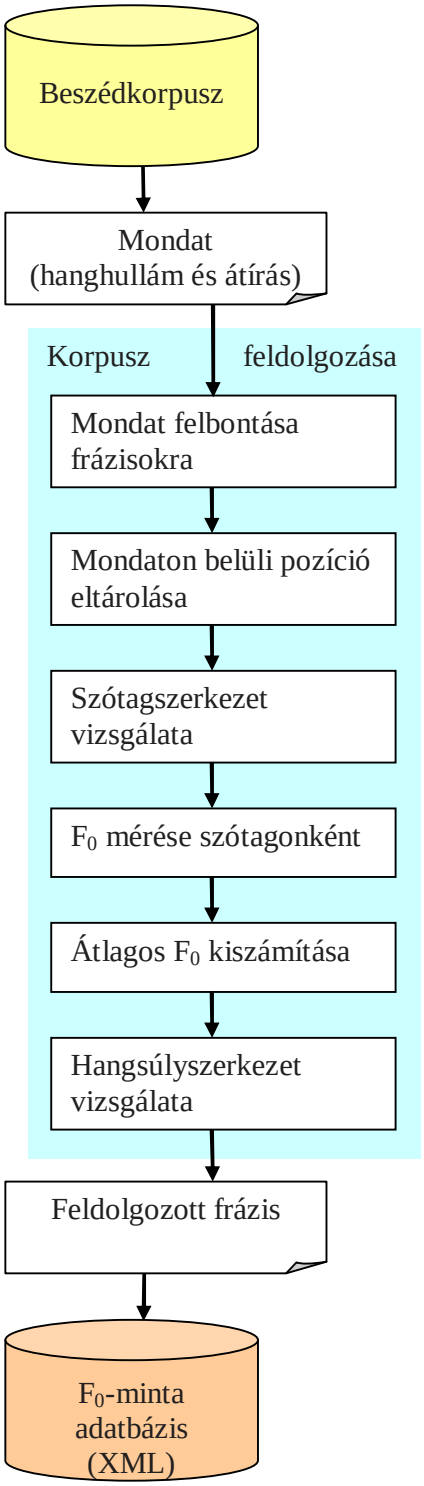
3.2.2. F_0-minta adatbázis létrehozása

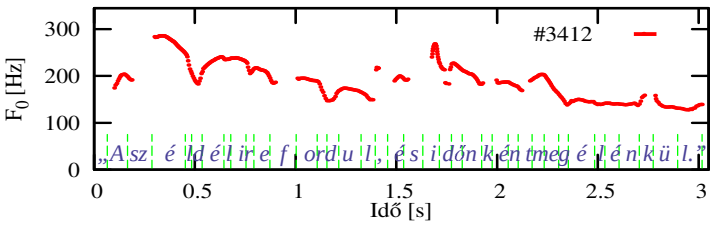
A beszédkorpuszokból először olyan adatbázisokat hoztunk létre, amelyek a módszerünk működéséhez szükséges címkéket tartalmazzák az egyes mondatokhoz, ahogy ennek tervét a 2.3.1. részben bemutattuk. A jelenlegi megvalósítás során először csak a mondatok alapfrekvenciáját vizsgáltuk, ami így a prozódián belül a dallam és a hangsúlyozás beállítását teszi lehetővé.

A megtervezett rendszer alapján (2.1. ábra) hajtottuk végre a korpuszok feldolgozását (3.1. ábra). Az ábrán végigkövethetjük ezen automatikus, manuális beavatkozás nélkül végrehajtható folyamat lépéseit egy konkrét példa kíséretében, amely az „Időjárás” gyűjteményből származik.

A módszer megfelelő működéséhez először a beszédkorpuszbeli mondatokat felbontottuk kisebb egységekre, frázisonként kezeltük őket a továbbiakban. Az egyes korpuszok mondatainak és frázisainak számát a 3.2. táblázat tartalmazza. Észrevehetjük, hogy az „Időjárás” és a „Prompt” gyűjteményben a mondatok számához képest meglehetősen sok a prozódiai egységek száma. Ezek a korpuszok hosszabb mondatokat tartalmaznak, egy mondat átlagosan 2,45 illetve 2,46 frázisból áll. A 3.2. ábrán látható az „Időjárás” beszédkorpusz mondatonkénti és frázisonkénti szótagszámainak hisztogramja. Azt érdemes megfigyelni, hogy a teljes mondatok hosszúak (nem ritka az ötven szótag feletti), míg a frázisok rövidebbek, így könnyebben kezelhetőek.

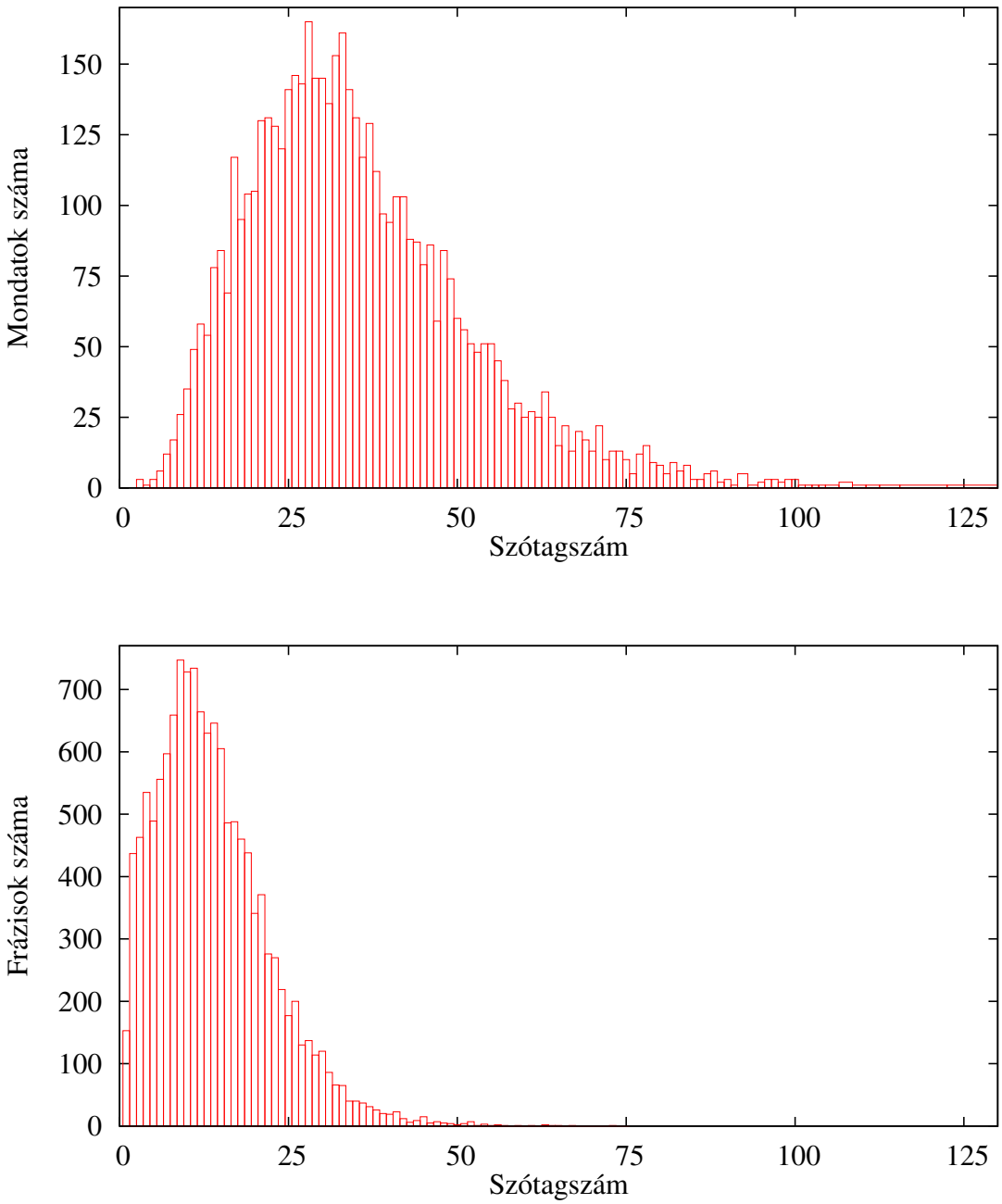
A mondatok felbontása után a mondaton belüli pozíció eltárolása következett. Minden frázishoz jelöltük, hogy első, középső, vagy utolsó helyen található az eredeti mondaton belül.



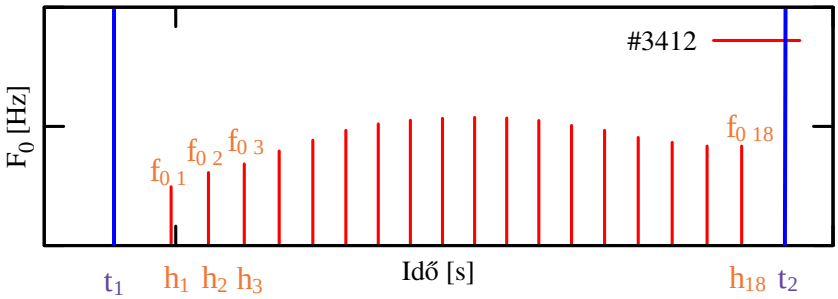


„A szél délire fordul,”	„és időnként megélénkül.”
Első	Utolsó
1+1+3+2	1+3+4
195, 255, 234, 229, 205, 178, 214 Hz	195, 222, 207, 189, 166, 145, 142, 135 Hz
215 Hz	175 Hz
-, W, N, N	-, W, N

3.1. ábra. Beszédkorpuszból F_0-minta adatbázis létrehozása. Bal oldalon a folyamat, jobb oldalon egy példa látható.



3.2. ábra. Mondatok és frázisok szótagszámának gyakorisága az „Időjárás” korpuszban. Az 5232 mondat átlagosan 2,45 frázisból áll (összesen 12 827 frázisra bontottuk fel a korpuszt). A mondatok meglehetősen hosszúak (nem ritka az ötven szótag feletti), míg a frázisok rövidebbek, így könnyebben kezelhetőek.



3.3. ábra. Szótag átlagos alapfrekvenciájának kiszámítása.

Erre azért volt szükség, mert a frázisok mondaton belüli pozíciója jelentős mértékben befolyásolja a dallamot.

Ezután a szótagszerkezeteket vizsgáltuk: minden frázishoz eltároltuk a szavak számát, a frázis összes szótagjának számát, valamint a szavankénti szótagszámokat (ez utóbbi tulajdonképpen a „szótagszerkezet”). A szótagok határainak meghatározását egyszerűsített módon végeztük: a szövegben egy magánhangzó és az előtte álló mássalhangzók tartoztak egy csoportba.

Az egyes szótagokhoz eltároltuk a hozzájuk tartozó alapfrekvencia-értéket. Az F_0-menet folytonos a beszéd zöngés részein, így minden szótaghoz egy átlagos F_0-értéket mentettünk el. Ennek menete a 3.3. ábrán követhető. Az i. szótag kezdetét t_i-vel, a j. zöngeperiódus-határt pedig h_j-vel jelöljük. Az alapfrekvencia-értékek (f_{0_j}) a következő módon, a zöngeperiódus-hosszak reciprokaként számolhatóak:

$$f_{0_j} = \frac{1}{h_{j+1} - h_j} \tag{3.1}$$

Az i. szótag átlagos alapfrekvenciáját a 3.2. egyenlet alapján számítjuk (feltéve, hogy h_k a szótag határán belüli első, h_n az utolsó zöngeperiódus-határ):

$$\text{szótag}\overline{F_{0_i}} = \frac{\sum_{j=k}^{n} f_{0_j}}{n - k + 1} \tag{3.2}$$

A szótagok mellett a teljes frázisra jellemző átlagos értéket is hasonlóan számoljuk (feltéve, hogy h_1 a frázis határán belüli első, h_m az utolsó zöngeperiódus-határ):

$$\text{frázis}\overline{F_0} = \frac{\sum_{j=1}^{m} f_{0_j}}{m} \tag{3.3}$$

Az értékek pontos kiszámítása és eltárolása rendkívül fontos, hiszen ezek alapján történik a későbbiekben a dallam másolása. A korpusz-feldolgozó programot ezért úgy állítottuk be, hogy jelezze a természetellenesen magas F_0 értékeket, amelyek a zöngeperiódus-határok nem

tökéletes detekciójából származnak. A jelzett helyeken a tárolt alapfrekvencia-érték manuálisan korrigálható.

A hangsúlyszerkezet adatbázisban tárolása is szükséges volt a későbbi F_0-minta kereséshez. A hangsúlycímkék meghatározása a korpuszbeli szöveges átírások alapján, a Profivox beszédszintetizátorban alkalmazott hangsúlydetekciós algoritmus segítségével történt [34]. A rendszer öt féle hangsúlytípust különböztet meg:

- **[F]**: fókusz, mondatközpont
- **[E]**: nyomaték, értelmi hangsúly
- **[W]**: normál szóhangsúly
- **[N]**: semleges szó
- **[-]**: negatív hangsúly

A legerősebb hangsúlytípus a mondatközpont, amely jelentős alapfrekvencia-emelést és időbeli nyújtást jelent, a leggyengébb pedig a negatív hangsúly, amely tulajdonképpen az F_0-menet lokális csökkenése.

A beszédkorpuszból F_0-minta adatbázis létrehozását egy C# nyelven [35] írt program végezte el. Az adatbázis fájlban tárolása XML [36] adatstruktúra segítségével történt, mert ez a formátum szabványos, platformfüggetlen és széles körben elterjedt. Az egy-egy frázishoz eltárolt tulajdonságokra az F.1.1. függelékben található példa. A teljes adatbázis feldolgozásának ideje a legnagyobb korpusz („Időjárás”) esetén is 10 perc alatt volt egy Pentium IV 2600 Mhz-es számítógépen.

3.2.3. Kísérlet hangidőtartam-minta adatbázis létrehozására

Ahhoz, hogy a mesterséges beszéd előállítása során a dallam mellett az időzítést is korpusz alapján lehessen létrehozni, hangidőtartam-minta adatbázisra is szükség van. Ennek segítségével megvalósítható lenne a hangidőtartamok természetes beszéd alapján történő beállítása.

A felhasznált korpuszokban a fonetikus átírások alapján vizsgáltuk a hangidőtartamokat. Ezeknek adatbázisban tárolása szavanként történt. Az egyes szavakhoz tároltuk a fájl nevét, amelyben megtalálható, a szó hangjait a Profivox rendszer jelölése szerint, valamint az egyes hangokhoz tartozó időtartamokat. Mivel a korpuszok létrehozásakor az időtartamok felmérése automatikusan történt, a hanghatárok meglehetősen pontatlanok. Ekkora beszédkorpuszok esetén azonban manuális felmérés nagyon költséges és időigényes lenne.

A hangidőtartam-minta adatbázis fájlban tárolása is XML adatstruktúra szerint történt, az F.1.2. függelék tartalmaz egy példát erre. Jelenlegi megvalósításunkban ezt az adatbázist még nem használtuk fel, mert elsődleges célunk a dallam vizsgálata volt.

3.2.4. Változatosság elemzése a gyakorlatban

A vizsgált beszédkorpuszok között több olyan is van, amelyben ismétlődő mondatok, illetve mondatrészletek fordulnak elő. Ezeket elemezve alaposabban is vizsgálható az emberi beszéd változatossága. A jelen munka során csak kezdetleges jelleggel, néhány példán keresztül tettük ezt meg azok dallamát megvizsgálva.

A „Prompt" korpuszban gyakori a teljes mondatok ismételt előfordulása. Ezek közül kiválasztottunk egyet (*„A díjcsomagokban az alábbi szolgáltatások havidíjmentesen vehetők igénybe."*), amelynek különböző változatait vizsgáltuk. A 3.4. ábra felső részén ezen mondat eleje látható, az egyes változatok szótagjainak átlagos F_0-értékeit összehasonlítva. Észrevehetjük, hogy a 13 változat közül sok olyan van, amely alig mutat eltérést.

Az „Időjárás" beszédkorpuszban csak rövidebb frázisok fordulnak elő többször. A 3.4. ábra alsó részén a *„zivatarok alakulhatnak ki"* prozódiai egység különböző előfordulásainak szótagonkénti átlagos F_0-értékeit tanulmányozhatjuk. Az öt változat egyes szótagjai között általában 5–10, de nem ritkán 20 Hz-nyi különbség is észlelhető.

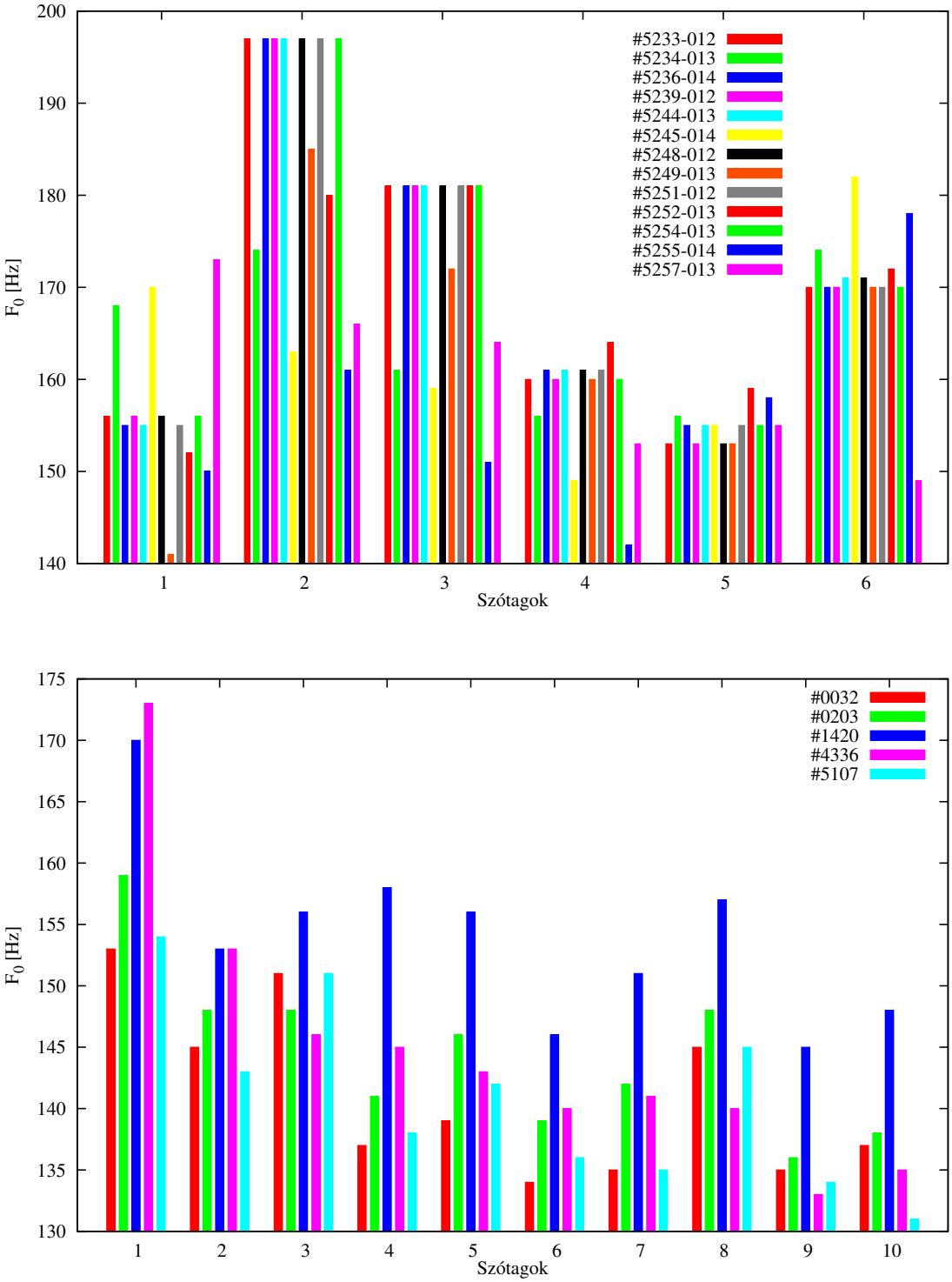
A két grafikont összehasonlítva az állapítható meg, hogy a „Prompt" korpuszban ugyan sok változata előfordul a vizsgált mondatnak, de ezek dallama között nagyon kicsi a különbség. Ezzel szemben az „Időjárás" beszédkorpuszban csak kisebb ismétlődő részeket találtunk, az elemzett esetben ezek között nagy az eltérés. Érdekes lenne megvizsgálni, hogy a különbség általánosan is érvényes-e a két korpusz között.

3.3. A megvalósított rendszer működése

Módszerünket a megtervezett rendszer (2.3.2. rész) alapján implementáltuk. Most bemutatjuk, hogyan történt a dallam korpusz alapján történő szöveghez rendelése a Profivox beszédszintetizátor segítségével.

3.3.1. Alapfrekvencia beállítása a Profivoxban

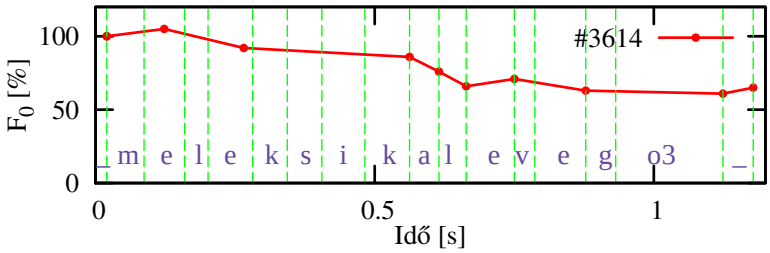
A konkrét implementáció során, a Profivox rendszerben a prozódia beállításához használt szimbolikus információ az úgynevezett intonációs mátrix. Ezen mátrix a szintetizálandó szöveg minden egyes hangjához (és szüneteihez) tartalmaz egy sort, amelyben megadhatók az aktuális paraméterek. Többek között megtalálható itt a hanghoz tartozó fonéma, az alapfrekvencia magassága egy alapértékhez képest százalékban kifejezve, illetve az, hogy az adott hangban hol kell elérni az előbbi magasságot (a hang hosszához képest, százalékban kifejezve). Az in-



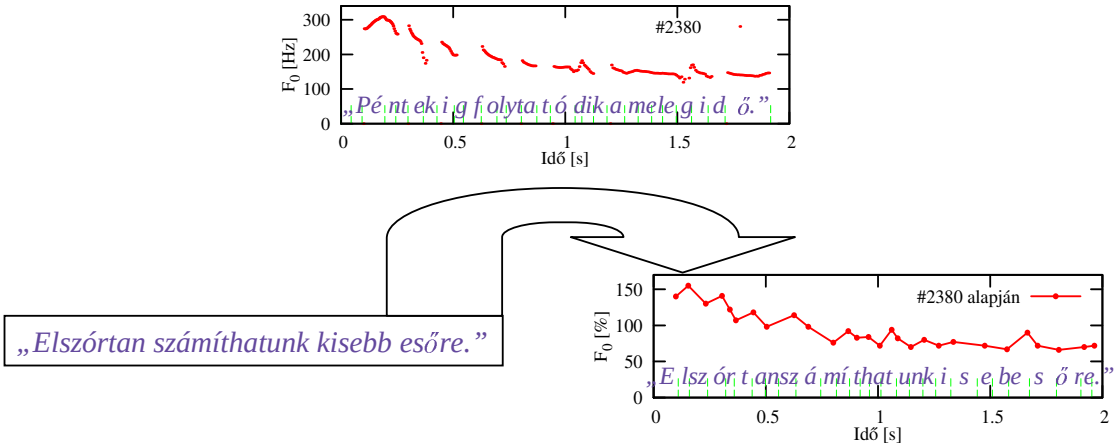
3.4. ábra. Változatosság elemzése a „Prompt” (felül) és „Időjárás” (alul) korpusz egy-egy példájában, a szótagonkénti átlagos F_0 vizsgálatával. A „Prompt”-beli változatok nagymértékben hasonlítanak egymáshoz, az „Időjárás”-beli mondatdarabok nagyobb eltérést mutatnak.

3.3. táblázat. Profivox intonációs mátrix. Minden sorban egy hanghoz tartozó paraméterek találhatóak, amivel az alapfrekvencia, időtartam és intenzitás megadható.

< 1>	<100>	<100>	< 0>	<100>	<_>	<0x70010000>	<20>
<19>	<100>	< 0>	< 90>	<100>	<m>	<0x00b0c010>	<66>
<10>	<105>	< 50>	< 90>	<100>	<e>	<0x00000010>	<73>
<32>	< 0>	<100>	< 90>	<100>	<l>	<0x00000010>	<41>
<10>	< 92>	< 80>	< 90>	< 85>	<e>	<0x00000010>	<80>
<16>	< 0>	<100>	< 81>	<100>	<k>	<0x00000010>	<61>
<27>	< 0>	<100>	< 81>	<100>	<s>	<0x00000010>	<61>
< 7>	< 0>	<100>	< 90>	<100>	<i>	<0x00000014>	<77>
<16>	< 86>	<100>	<100>	<100>	<k>	<0x00000012>	<79>
< 3>	< 76>	<100>	< 72>	< 60>	<a>	<0x00504306>	<52>
<32>	< 66>	<100>	<100>	< 60>	<l>	<0x02b04060>	<48>
<10>	< 71>	<100>	< 90>	< 80>	<e>	<0x00000060>	<86>
<24>	< 0>	<100>	<100>	<100>	<v>	<0x00000060>	<36>
<10>	< 63>	<100>	<100>	< 70>	<e>	<0x00000060>	<91>
<15>	< 0>	<100>	<100>	<100>	<g>	<0x00000060>	<53>
<38>	< 61>	<100>	< 95>	< 60>	<o3>	<0x00000066>	<192>
< 1>	< 65>	<100>	<100>	<100>	<_>	<0x70010000>	<54>



3.5. ábra. Profivox intonációs mátrix által definiált dallammenet: hangonként egy töréspont adható meg (piros pontok), amelyekkel az F_0-görbét tudjuk meghatározni (piros töröttvonal). A függőleges zöld szaggatott vonalak a tervezett hanghatárokat jelzik.



3.6. ábra. A dallammásolás módszere: a bemeneti mondathoz hozzárendeljük egy természetes beszédből felvett mondat dallamát.

tonációs mátrixban megadható a hang tervezett időtartama és intenzitása is. A 3.3. táblázat és a 3.5. ábra egy intonációs mátrixra, és az abban definiált dallammenetre mutat példát.

A dallamot tehát egy olyan töröttvonallal tudjuk megadni, amely legfeljebb hangonként egy töréspontot tartalmaz. Az alapfrekvencia ugyan csak a beszéd zöngés hangjain értelmezett, de az intonációs mátrixban zöngétlen hangokra is adhatunk meg értéket, ha pontosabbá akarjuk tenni az alapfrekvencia-görbe leírását.

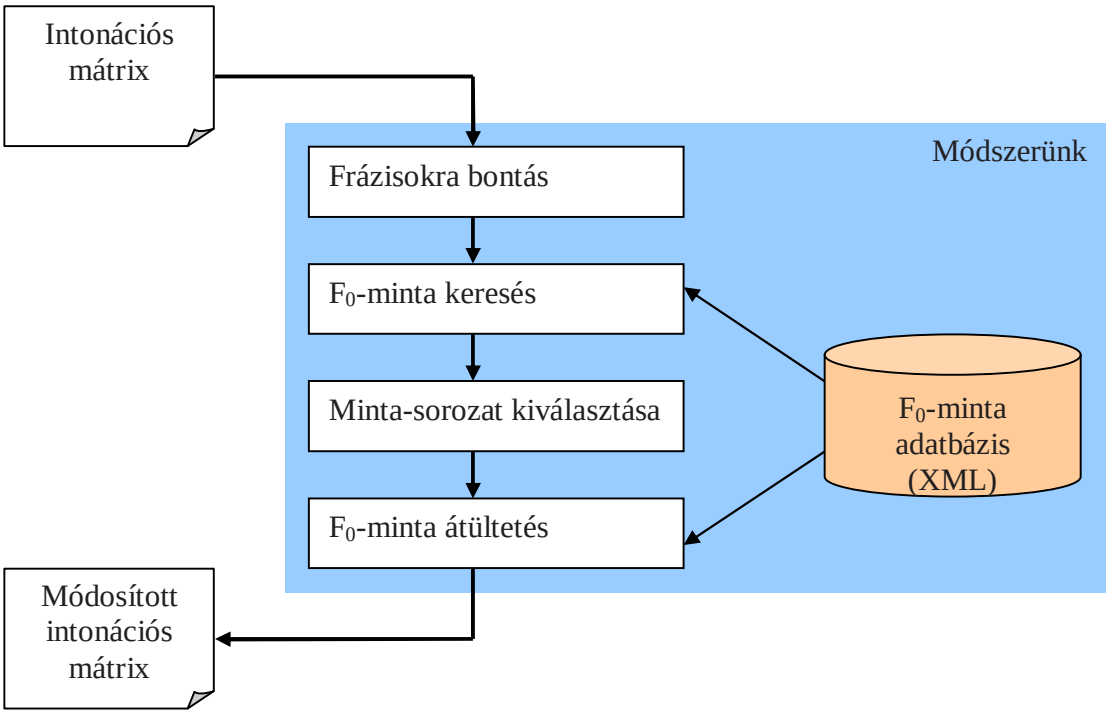
3.3.2. A dallammásolás módszere

Korábbi munkánk során több kísérletet tettünk arra, hogy a szintetizálandó szöveghez természetes beszéd alapján rendeljünk dallamot [31]. Ezen kísérletekben a bemenet szövegéhez manuálisan hozzárendeltük egy természetes beszédből felvett másik mondat dallammenetét. A 3.6. ábrán erre látható példa, a #2380-as mondat dallammenetét szótagonként hozzáillesztettük a bemeneti mondat szótagjaihoz. Ily módon sikeresen tudtuk „másolni” a természetes beszéd dallamát.

3.3.3. Változatos dallam minták alapján

A manuális dallammásolás sikere után automatikus módszert dolgoztunk ki az alapfrekvencia szöveghez rendelésére. Ennek működését a 3.7. ábrán tekintjük át.

A bemeneti mondat Profivox intonációs mátrixként áll rendelkezésre, amelyet először kisebb egységekre, frázisokra bontunk, hiszen a mintákat tartalmazó adatbázis is ilyen egysé-



3.7. ábra. A változatosságért felelős rendszer megvalósítása. A bemeneti intonációs mátrixhoz új dallamot határoz meg módszerünk az F_0-minta adatbázis alapján.

gekből áll. A prozódiai egységekre bontás az írásjelek segítségével történik, ezek képezik a frázisok határait.

Az egyes frázisokhoz az F_0-minta adatbázisból keresünk dallamot. A megfelelő minta kiválasztása többféle hasonlósági mérték szerint történhet.

Egyező szótagszerkezet

A kezdeti megvalósítás során a frázisok szótagszerkezetének egyezése alapján történt a keresés. Például a *„Hazánkban folytatódik a fülledt idő.”* frázishoz csak a $3+4+1+2$ szótagszerkezetűek megfelelőek az adatbázisból. Mivel a magyar nyelvben általában a szavak első szótagja a hangsúlyos, ezért a hasonló szótagszerkezet egyben hasonló hangsúlyszerkezetet (ezáltal hasonló dallamot) is jelent, vagyis várhatóan jó helyre kerülnek a hangsúlyok a szintetizálandó mondatban is. Ez a hasonlósági mérték tehát megfelelő a dallam másolására.

Hasonló szótagszerkezet

A pontosan egyező szótagszerkezet problémája, hogy szűkíti a lefedettség arányát azáltal, hogy csak a bemenethez szerkezetileg teljesen megegyező frázisokat használhatjuk fel az F_0-minta adatbázisból. Megvalósításunkban tehát enyhítettünk ezen a követelményen: hasonlósági mértéknek a nem teljesen egyező, de hasonló szótagszerkezetet alkalmaztuk. Ez azt jelenti, hogy a rövid (három szótagnál rövidebb) szavak esetében pontos szerkezetbeli egyezést követelünk meg, a hosszabbak esetén viszont lehet $\pm$ 1 szótag eltérés. Így a *„Hazánkban folytatódik a fülledt idő."* frázishoz megfelelő $3 + 3 + 1 + 2$ vagy a $3 + 5 + 1 + 2$ szerkezetű is.

Természetesen ilyen esetben a dallam másolása kicsit bonyolultabb, egyes F_0 értékeket el kell hagyni, vagy interpolálni kell. Mivel a hangsúly a legtöbb esetben a szó elején, főleg az első szótagon van, a szó végi interpoláció várhatóan nem lesz zavaró.

Hangsúly és pozíció

A szótagszerkezet mellett figyelembe vehető a frázis mondaton belüli pozíciója, és a hangsúlyszerkezet is. Ekkor várhatóan jobb lesz a kapott dallam, de kevesebb lehetőség közül választhatunk.

Egy-egy frázishoz nagy valószínűséggel több szerkezetileg hasonló minta is előfordul az adatbázisban. Választani kell tehát közülük egyet, ami módszerünkben véletlenszerűen történik. A véletlen választással biztosítható, hogy ha többször egymás után ugyanazt a mondatot, vagy akár hasonló szerkezetűt szintetizálunk, ezek különböző dallamúak lesznek. Így tehát hosszabb mesterségesen előállított beszédben is csökken a monotonitás.

Az F_0-minta adatbázisból kiválasztott sorozatot a bemenethez rendeli a módszer. Ennek során beállítjuk az intonációs mátrixban lévő F_0-értékeket az adatbázisbeli minták alapján. Minden szótag magánhangzójához megadjuk az adatbázisban tárolt átlagos F_0-értéket.

A Profivox rendszer ANSI C nyelven [37] íródott, így a kiegészítő modult is ezen a programozási nyelven készítettük el. A főbb függvények a 3.7. ábra egyes lépéseit valósítják meg:

- **countImf**: frázisokra bontás
- **searchPhrase**: F_0-minta keresés
- **selectF0Contour**: minta-sorozat kiválasztása
- **copyF0**: F_0-minta átültetés

A függvények pontos definíciója az F.2.1. függelékben található meg.

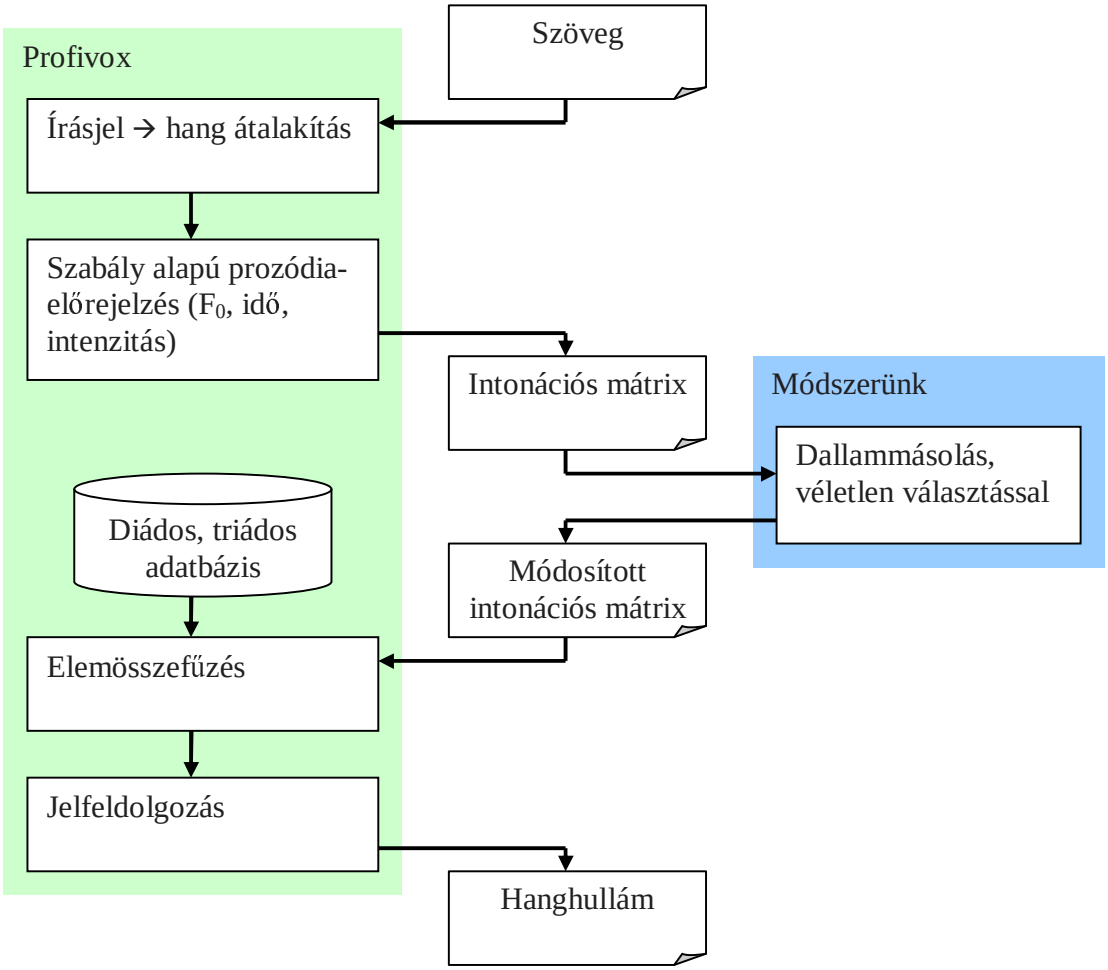
3.4. Illesztés a Profivox szövegfelolvasóba

Feladatunk az volt, hogy a dallammásolás módszerét a BME TMIT-en fejlesztett Profivox diád/triád alapú szövegfelolvasó rendszerbe integráljuk. A 3.8. ábra részletesen bemutatja ennek menetét.

A Profivox rendszer először a bemeneti szövegen előfeldolgozást végez, majd szabályok segítségével az egyes mondatokhoz rendeli a prozódiát. Ennek a folyamatnak az eredménye egy intonációs mátrix, amelyből módszerünk kiindul. A 3.3.3. részben bemutatott módon megtörténik a változatos dallam beállítása, módszerünk kimenete egy F_0-értékekben módosított intonációs mátrix. Ezt visszakapva a Profivox rendszer elvégzi az elemösszefűzést, majd létrehozza a beszédet.

A kapcsolatot a beszédszintetizátor és a kiegészítő modul között a következő ANSI C nyelven írt függvények valósítják meg, melyeknek pontos definícióját az F.2.2. függelék tartalmazza:

- **init**: működési paraméterek beolvasása, erőforrások lefoglalása
- **prosody_variation**: változatos prozódia megvalósítása
- **destroy**: erőforrások felszabadítása



3.8. ábra. Profivoxhoz illesztés megvalósítása. A Profivox rendszer a bemenetből először fonetikus átírást hoz létre. Ezután szabály alapon meghatározza a prozódiát, amelyet egy intonációs mátrixban tárol. Ebben módosítjuk módszerünkkel az F_0-t, a változtatott mátrixot pedig a Profivox beszéddé alakítja.

4. fejezet

A megvalósított rendszer értékelése

A rendszer megvalósítása után megvizsgáltuk, hogy milyen minőségű mondatokat lehet előállítani módszerünkkel. Mivel kutatásunk több lépésből állt, ismertetjük a korábbi meghallgatásos teszteket is.

A fejezetben először elemezzük, hogy milyen mértékben sikerült megvalósítani a változatosságot (4.1. alfejezet). A 4.2. alfejezetben a korábbi és jelenlegi meghallgatásos tesztek körülményeiről és eredményeiről olvashatunk.

4.1. Követelmények vizsgálata

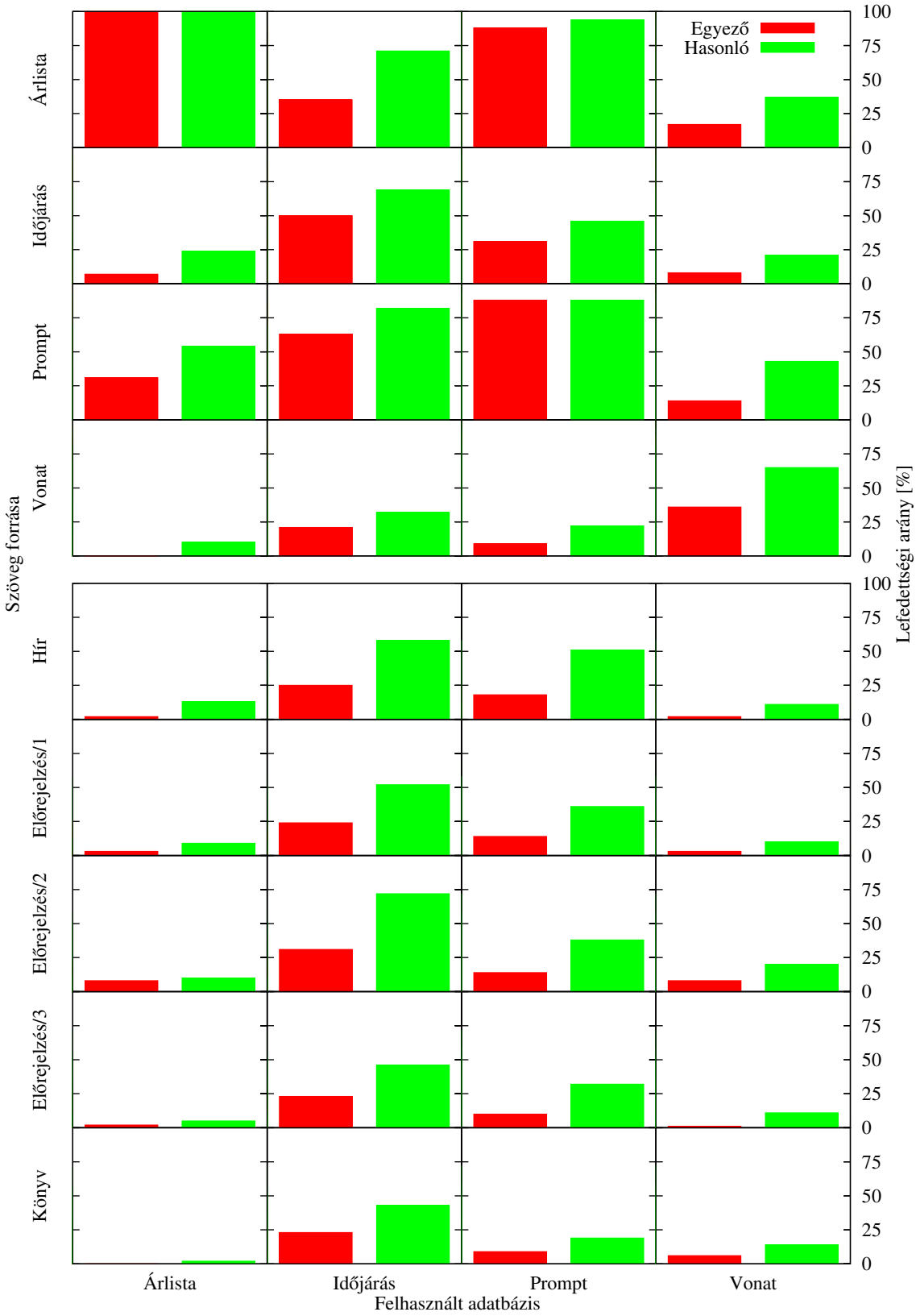
A rendszer tervezésekor számos követelményt gyűjtöttünk össze, amelyek a megfelelő működést tudják biztosítani (2.1. alfejezet). A következőkben elemezzük az elért lefedettséget, a találatok számát és módszerünk futásidejét.

4.1.1. Lefedettségi arányok vizsgálata

Először azt vizsgáltuk, hogy milyen lefedettségi arányokat lehet elérni különböző témájú bemeneti szövegekhez. A lefedettségi arányokat a szótagok száma alapján mértük a 2.1.1. részben leírtak szerint. Két különböző hasonlósági mértéket vizsgáltunk (amelyeket a 3.3.3. részben bemutattunk), és összehasonlítottuk az ezekkel elért lefedettségeket.

Korpuszbeli mondatok

A felhasznált beszédkorpuszok mindegyikéből (a „Harang” gyűjtemény kivételével) kiválasztottunk 20–30 mondatot. A „Harang” korpuszt azért nem vettük figyelembe a vizsgálat



4.1. ábra. Lefedettségi arányok vizsgálata különböző szövegeken többféle adatbázissal.

során, mert 50 mondatával túl kicsi a többi hanganyaghoz képest.

A kiválasztott mondatokat kihagyva, a korpuszok maradék részéből F_0-minta adatbázisokat hoztunk létre a 3.2.2. részben leírtak szerint. A kapott adatbázisok segítségével a változatos dallam megvalósítására kifejlesztett módszerünket alkalmazva (3.3. alfejezet) szintetizáltuk a mondatokat. Hasonlósági mértéknek az egyik esetben az „egyező", a másik esetben a „hasonló" szótagszerkezetet használtuk úgy kiegészítve, hogy lehetőleg a pozíciók és a hangsúlyszerkezet is egyezzen a bemeneti szövegben és az adatbázisbeli mintában.

A négy kiválasztott mondatcsoportot a négy adatbázis felhasználásával szintetizálva összesen 16 változatot kaptunk. Az elért lefedettségi arányok összehasonlítása a 4.1. ábra felső részén látható. A piros színű oszlopok ábrázolják azokat az eseteket, amikor „egyező" szótagszerkezet volt a hasonlósági mérték, a zöld szín pedig a „hasonló" szótagszerkezetre utal. Az vehető észre, hogy utóbbi sokszor több, mint 25%-os lefedettség-növekedést ért el az előbbihez képest. Ha az a célunk, hogy a mesterséges beszéd létrehozása során minél nagyobb részhez lehessen természetes, változatos dallamot rendelni, akkor a „hasonló" szótagszerkezetet érdemes alkalmazni. Az „egyező" szerkezet ugyanakkor pontosabb dallammásolást eredményezhet.

Megfigyelhetjük, hogy a szöveg származása nagyban befolyásolja azt, hogy milyen lefedettségi arány érhető el. Amennyiben a bemeneti mondatok szerkezete közel van az adatbázisban tároltakéhoz (pl. az „Árlista" korpuszból származó szöveg az „Árlista" adatbázissal), magasak az elért értékek. Ezzel szemben ha az adatbázis távol van a szintetizálandó mondatok szerkezetétől (pl. az „Időjárás" korpuszból származó szöveg az „Árlista" és „Vonat" adatbázissal), akkor még a 25%-os lefedettséget sem lehet elérni.

Összességében elmondhatjuk, hogy az „Időjárás" és a „Prompt" adatbázis segítségével megfelelő lefedettséget kaptunk a legtöbb esetben, hiszen a 2.1.1 részben az 50% elérését céloztuk meg. Ezzel szemben az „Árlista" adatbázis a speciális szerkezetű mondatai miatt, míg a „Vonat" adatbázis kis mérete miatt kevésbé alkalmas F_0-mintának.

Hosszabb összefüggő szövegből származó mondatok

Az elérhető lefedettségi arányt tovább vizsgáltuk a korpuszoktól független szövegeken, hiszen a valós alkalmazás során is várhatóan ilyeneket kell szintetizálni. Ehhez hosszabb összefüggő szövegeket kerestünk, amelyek lefedettségét az előző részhez hasonlóan a négy F_0-minta adatbázis felhasználásával mértük. A következő szövegeket gyűjtöttük össze:

- **Hír**: `http://index.hu` oldalról származó, 2008. április 15-i hír (18 mondat)
- **Előrejelzés/1**: `http://www.met.hu` oldalról származó, 2008. március 28-i időjárás-előrejelzés (25 mondat)

- **Előrejelzés/2**: `http://www.met.hu` oldalról származó, 2008. április 14-i időjárás-előrejelzés (22 mondat)
- **Előrejelzés/3**: `http://www.met.hu` oldalról származó, 2008. április 26-i időjárás-előrejelzés (23 mondat)
- **Könyv**: Gárdonyi Géza: „Egri csillagok" című könyv részlete (19 mondat)

Az öt szöveget a négy adatbázis felhasználásával szintetizálva összesen 20 csoportot hoztunk létre. A lefedettségi arányok összehasonlítása a 4.1. ábra alsó részén látható. Rögtön szembetűnik, hogy majdnem az összes esetben alacsonyabb a lefedettség, mint a korpuszokból származó mondatok vizsgálatakor (az ábra felső része). Ezt azzal lehet magyarázni, hogy a szövegek szerkezete jelentős mértékben eltér az adatbázisokétól, ami így kisebb találati arányt eredményez.

Az „Időjárás" adatbázis mind az „egyező", mind a „hasonló" szótagszerkezet alkalmazásával lényegesen jobb lefedettséget ért el a többi háromhoz képest. Ez a legalkalmasabb tehát a változatos dallamminták létrehozására, a továbbiakban ezért csak az „Időjárás" adatbázist alkalmaztuk.

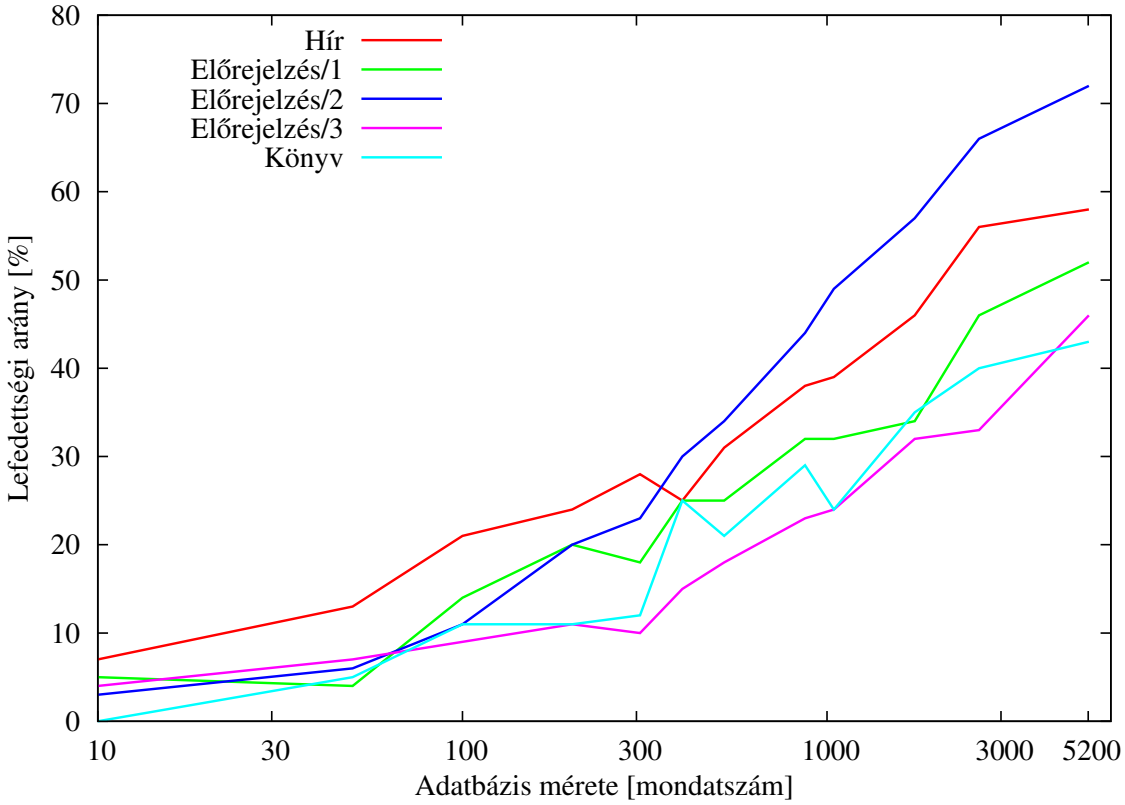
4.1.2. Lefedettség függése az adatbázis méretétől

Felmerülhet a kérdés, hogy a felhasznált F_0-minta adatbázis mérete milyen mértékben befolyásolja az elérhető lefedettséget. Ehhez az „Időjárás" adatbázist vizsgáltuk a „hasonló" szótagszerkezet hasonlósági mérték alkalmazásával. Az előző részben felhasznált, hosszabb összefüggő szövegből származó mondatokon végeztünk elemzést az F_0-minta adatbázis méretének fokozatos növelésével.

A vizsgálat eredményét a 4.2. ábra mutatja. Az elért lefedettségi arányt ábrázoltuk az adatbázis méretének függvényében, amelyet a mondatok számával adtunk meg. Az adatbázis méretét exponenciálisan kell növelni a lefedettség megfelelő növekedéséhez. 100 mondat esetén csak 10% körüli, 1000 mondat esetén már 30%-os, míg a teljes, 5232 mondatos adatbázis esetén 50%-os lefedettségi arány is elérhető. A teljes korpusz feldolgozásával megoldható tehát a követelmények között kitűzött cél elérése.

4.1.3. Változatok számának vizsgálata

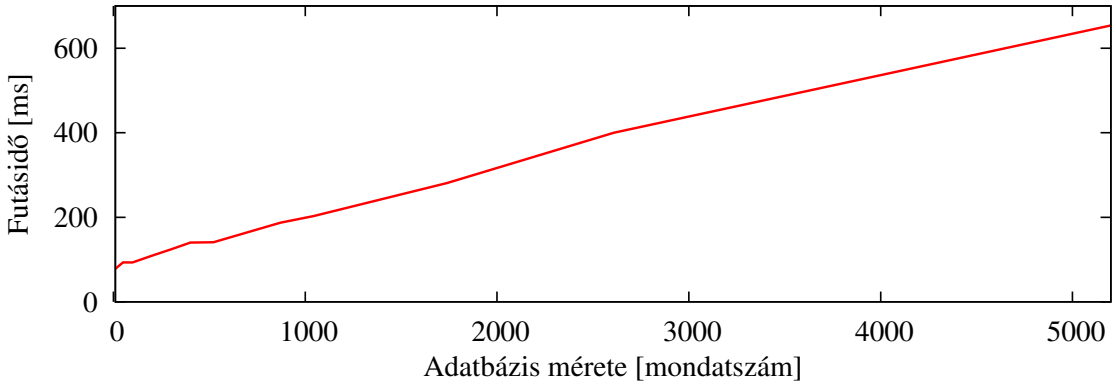
A 2.1. alfejezetben utaltunk arra, hogy a változatos beszéd létrehozásához az lenne az ideális, ha legalább 3–4 dallammintát találna a módszer a bemeneti szöveg egy-egy részéhez. Az adatbázisok szerkezetéből adódóan abban, amelyben rövidebb elemek sokszor előfordulnak,



4.2. ábra. Lefedettségi arány függése az adatbázis méretétől. Az „Időjárás” gyűjteményből egyre növekvő mennyiségű mondatot kiválasztva hoztuk létre az F_0-minta adatbázist, és közben vizsgáltuk a lefedettséget különböző mondatcsoportokon. Az adatbázis méretét logaritmikus skála szerint ábrázoltuk.

4.1. táblázat. Legalább három változat előfordulásának aránya a lefedett frázisok között az „Időjárás” adatbázis felhasználásával. A százalékos értékek azt mutatják, hogy a bemeneti szövegek összes lefedett frázisának mekkora részéhez van legalább három találat.

Szöveg forrása	**Egyező**	**Hasonló**
Árlista	54%	57%
Időjárás	73%	81%
Prompt	36%	62%
Vonat	64%	67%
Hír	67%	83%
Előrejelzés/1	62%	64%
Előrejelzés/2	52%	68%
Előrejelzés/3	48%	60%
Könyv	50%	67%



4.3. ábra. Futásidő függése az adatbázis méretétől: az idő lineárisan növekszik. A mérést az „Előrejelzés/1” szövegen végeztük.

magasabb lesz ez a szám. Ugyanakkor ha egy adatbázis sokféle szerkezetű mintát tartalmaz, akkor ezekből várhatóan kevesebb egyforma található.

Méréseket végeztünk a változatok számának megbecsülésére az „egyező” és „hasonló” szótagszerkezet hasonlósági mértéket összehasonlítva. Ugyanazokat a bemeneti szövegeket használtuk fel, mint a korábbi elemzésekben. A 4.1. táblázat ennek egy részletét jeleníti meg, amikor az „Időjárás” adatbázis F_0-mintáit alkalmaztuk. Azoknak az eseteknek az arányát jelöltük, amelyekben a bemeneti szöveg lefedett részéhez (amit a 4.1.1. részben vizsgáltunk) legalább három különböző F_0-minta található az adatbázisban.

Ez a „változatossági arány” minden esetben 50% fölött van, és a bemeneti szöveg szerkezetétől függően elérheti a 80%-ot is. A természetes emberi beszéd változatossági aránya ilyen szempontból 100% lenne, aminek megközelítése nem egyszerű feladat.

4.1.4. Futásidő vizsgálata

Módszerünk megvalósítása során arra törekedtünk, hogy ne nőjön meg jelentős mértékben a beszéd szintéziséhez szükséges idő. Természetesen a futásidő-növekedés elkerülhetetlen a korábbi szövegfelolvasóhoz képest, hiszen a programnak be kell olvasnia az F_0-minta adatbázist a merevlemezről, és megfelelő mintákat kell keresnie, amiket a bemeneti szöveghez rendel.

A 4.3. ábrán láthatjuk, hogy a futásidő milyen mértékben függ a felhasznált F_0-minta adatbázis méretétől. A mérések során az „Időjárás” adatbázist vizsgáltuk az „Előrejelzés/1” bemeneti szöveg 25 mondatával egy Pentium IV 2600 Mhz-es számítógépen. A futásidő lineáris növekedése az adatbázis beolvasásából fakad, hiszen egyre nagyobb adatmennyiséget kell be-

tölteni. A teljes, 5232 mondatból álló adatbázis felhasználásával is csak 656 ms volt a futásidő, ami így mondatonként átlagosan 26 ms-ot jelent, és nem zavaró a valós szintézis során.

4.2. Meghallgatásos tesztek

Kutatásunk során folyamatosan végeztünk meghallgatásos teszteket, amelyekkel ellenőrizni tudtuk módszerünk eredményességét. Röviden bemutatjuk, hogy kutatásunk mely lépéseken keresztül jutott a jelenlegi állapotba. A legújabb meghallgatásos teszt körülményeit részletesen ismertetjük.

4.2.1. Korábbi tesztek

Korábbi tesztjeink során először csak kisebb adatmennyiséget (200 mondat) vizsgáltunk, majd a teljes „Időjárás" korpuszt felhasználtuk. Az első kísérletekben a dallam másolásának megvalósíthatóságát elemeztük, majd fokozatosan haladtunk végső célunk felé.

Természetes mondatok módosítása

Első lépésként természetes mondatok dallammenetét módosítottuk a Praat fonetikai beszédanalizátor program segítségével [33]. A vizsgálatokat az „Időjárás" korpusz 200 mondatos részhalmazán végeztük. Az emberi beszédből felvett mondatok F_0-menetét kisebb-nagyobb mértékben változtattuk. Ezen kívül a természetes mondatokat a szövegük alapján újraszintetizáltuk, majd a szintetizált változatok dallammenetén is módosítottunk. Az eredeti és módosított hanganyagokat párokba rendeztük.

Legelső tesztünk elvégzésére 2006. októberében került sor, 26 tesztelő hallgatta meg a 28 összeállított mondatpárt. A tesztelőknek azt kellett eldönteniük, hogy hallanak-e különbséget a mondatpár két változata között, és ha igen, melyiket tartják jobbnak. Az eredményekből az derült ki, hogy az F_0-menet módosítását a legtöbb esetben észrevették a tesztet elvégzők. Az eredeti és módosított változatokat hasonló minőségűre értékelték, vagyis a megváltoztatott dallamú mondatok sem voltak rosszabbak az eredetinél. A kiértékelésből azt a következtetést vontuk le, hogy a dallam ilyen jellegű módosításával érdemes foglalkozni a változatos prozódia érdekében.

Dallammásolás manuálisan

A 3.3.2. részen már említettük, hogy a dallam másolását először manuális módszerrel oldottuk meg [31]. A szintetizált beszédet úgy hoztunk létre, hogy a mondatok dallamát természetes beszédből származó mondat alapján rendeltük a szöveghez. Itt már a teljes „Időjárás" korpuszt felhasználtuk. A manuális dallammásolással előállított mondatokat a Profivox szabály alapú mondataival párokba rendeztük, majd ezeket ismét meghallgatásos vizsgálatokban értékeltük.

A tesztet egy másik kutatással közösen végeztük 2008. március hónapban, 48 mondatpár szerepelt a vizsgálatban. Összesen 194 tesztelő vett részt benne, egy-egy mondatpárt átlagosan 21 tesztelő hallgatott meg. A feladat ismét egy-egy mondat két változatának összehasonlítása volt természetesség szerint. A tesztet elvégzők az esetek 49%-ában a dallammásolt változatot tartották jobbnak, 26%-ban egyformának hallották a két változatot, míg a Profivox szabály alapú dallammenetet csak 25%-ban preferálták. A szintetizált beszéd dallamát tehát létre lehetett hozni természetes mondatok alapján.

Dallammásolás automatikusan

A manuális dallammásolást automatikus módszerrel próbáltuk felváltani [38]. Ehhez az volt szükséges, hogy mind a bemeneti szöveget, mind az „Időjárás" korpuszbeli mondatokat frázisokra bontsuk. A beszédkorpuszt felhasználva automatikusan tudtunk a bemeneti szöveghez dallammintát rendelni. Az adatvezérelt prozódiai modell működéséből fakadó hibák kiküszöbölése csak részben történt meg, ezért a meghallgatásos tesztekhez olyan mondatokat választottunk ki, amelyek nem tartalmaztak hangsúlyhibát. Az volt a feltételezésünk, hogy ezen hibák a későbbiekben az algoritmusok finomhangolásával javíthatóak. Az időjárás-előrejelzés témaköréből 10 mondatot kiválasztva szintetizáltuk azokat, különböző dallammenetekkel. Mondatpárokat állítottunk elő, amelyekben ezeket a változatokat egymással, illetve a Profivox szabály alapú prozódiájával hasonlítottuk össze.

13 tesztelő végezte el a 37 mondatpár meghallgatását 2008. novemberében. Azt kellett megjelölniük, hogy a mondatpárok melyik változatát tartják természetesebbnek. Egy-egy mondatnak a különböző változatai között észrevehető különbség volt dallamban. A tesztelők az automatikus dallammásolással létrehozott mondatok legtöbbjét kellemes hangzásúnak értékelték. Sikerült egy-egy mondatnak több különböző, de egyforma minőségű változatát előállítanunk, ami a változatos beszéd létrehozásának elengedhetetlen kelléke.

4.2.2. Prozódiai változatosság tesztelése

Kutatásunkat tovább folytatva a dallammásolás automatikus módszerét a Profivox szövegfelolvasóba illesztettük. Ezáltal lehetővé vált hosszabb szövegek szintetizálása és vizsgálata is. A legújabb teszt célja az volt, hogy hosszabb mesterséges beszéden vizsgáljuk a változatos dallam létrehozásának eredményességét. A mondatok létrehozásához a teljes „Időjárás" F_0-minta adatbázist alkalmaztuk, a mintakeresés során a hasonlósági mérték a „hasonló" szótagszerkezet volt. A korábban ismertetett összefüggő szövegek egy részét használtuk fel ezen kísérlet elvégzéséhez. A tesztben vizsgált szövegeket az F.3.1. függelék tartalmazza, a hanganyagok pedig a CD-mellékletben találhatóak.

A szövegeket először a Profivox prozódiai modelljének módosított változatával szintetizáltuk [39], amely az eredeti szabály alapú modellhez képest jobb dallamösszeállítási képességgel rendelkezik, és alkalmas hírfelolvasásra is. Ezután a prozódiai változatosságot biztosító rendszerünk segítségével is előállítottuk a mondatokat. Az öt felhasznált szöveg két-két változatából bekezdéspárokat hoztunk létre.

Tesztkörnyezet

A létrehozott bekezdések tesztelését a BME TMIT-en kifejlesztett webes tesztelő rendszerben végeztük. Az öt bekezdéspár mindegyike kb. másfél perc hosszúságú volt, így a teljes meghallgatandó beszéd ideje 8 percre nyúlt.

A tesztelőknek a `http://speechlab.tmit.bme.hu/csapo08/` oldalt meglátogatva először egy rövid ismertetőt kellett elolvasniuk a teszt menetéről, majd néhány információt kértünk be róluk (becenév, életkor, nem, hangfelszerelés). Ezután megkezdődhetett a hanganyag meghallgatása.

A tesztet elvégzőknek minden bekezdéspár meghallgatása után két kérdésre kellett öt lehetőség közül választ adniuk. A kérdések a következőek voltak:

- „Melyiknek változatosabb a dallammenete (vagy esetleg hanglejtése)?"
- „Melyiket hallgatnád szívesebben?"

A lehetséges válaszok:

1. egyértelműen az első
2. inkább az első
3. egyforma
4. inkább a második

4.2. táblázat. A meghallgatásos teszt eredménye.

Bekezdés	**1. kérdés**						**2. kérdés**					
	=1	**=2**	**=3**	**=4**	**=5**	**Átlag**	**=1**	**=2**	**=3**	**=4**	**=5**	**Átlag**
Könyv	5	8	3	4	1	2,43	7	7	2	2	3	2,38
Hír	2	6	3	7	3	3,14	5	5	3	6	2	2,76
Előrejelzés/1	4	3	4	7	3	3,10	5	8	5	2	1	2,33
Előrejelzés/2	2	8	3	5	3	2,95	6	9	2	3	1	2,24
Előrejelzés/3	2	8	5	4	2	2,81	2	8	3	6	2	2,90

5. egyértelműen a második

A tesztelők egy-egy bekezdéspárt többször is meghallgathattak, hogy döntésüket könnyebben meg tudják hozni. A hangsorok lejátszása véletlen sorrendben történt. A szintetizált beszéd meghallgatása után a tesztelők megjegyzést is írhattak észrevételeikről.

Tesztelők

A bekezdéspárok meghallgatását 2008. májusában 21 tesztelő végezte el. A tesztelők mindannyian ép hallású, magyar anyanyelvű emberek voltak, 15–55 év közötti korral. Egy részük tanszéki munkatárs, a témához értő volt. A többiek nagy része egyetemi hallgatók köréből került ki, illetve egy látássérült is elvégezte a meghallgatást. A legtöbben átlagos minőségű hangfelszereléssel, csendes környezetben hallgatták meg a mondatokat. A rendszer rögzítette a teszt elkezdésének és befejezésének időpontját, így ellenőrizni tudtuk, hogy mindenki végighallgatta a mondatokat. A teszt átlagos meghallgatási ideje 10 perc volt.

Teszt-eredmények

Az elvégzett teszt eredményeit a 4.2. táblázat tartalmazza. Összesítve láthatjuk, hogy a tesztelők a két-két kérdésre milyen válaszokat adtak. A táblázat fejléce a lehetséges értékeket mutatja: az „=1” és „=2” oszlopok azt jelölik, hogy hányan preferálták a Profivox hírfelolvasó profillal készült változatot. Az „=3” a közömbös válaszok számára utal, míg az „=4” és „=5” azt mutatja, hogy hány tesztelőnek tetszett jobban a változatos dallamú bekezdés. Az „Átlag” oszlopban pedig ezen értékek átlaga látható.

Azt vehetjük észre, hogy a tesztelők semelyik esetben sem tartották lényegesen jobbnak a bekezdéspár valamely tagját, mert kicsi az „egyértelműen az első/második” válaszok aránya. A „Könyv” bekezdésnél mindkét kérdésre adott válasz átlaga 2,4 körül van, tehát a szabály alapú dallamot többen preferálták. A „Hír” és „Előrejelzés/1” esetén jobban észrevehető volt

a változatosság a többi mondathoz képest, hiszen itt magasabb az első kérdésre adott válaszok átlaga. Érdekes módon az „Előrejelzés/3” bekezdést kevésbé tartották változatosnak a tesztelők, mégis szívesebben hallgatták.

A válaszok nagy szórása következhet abból, hogy a meghallgatandó hangok meglehetősen hosszúak voltak, és nagyon kellett koncentrálni a megfelelő döntés meghozatalához. Egy másik ok lehet, hogy saját meghallgatásaink során a változatos dallamú változatok ugyan jobbnak tűntek, de az adatvezérelt prozódia-hozzárendelés miatt hibák is előfordultak. Több esetben a hangsúlyok nem megfelelő helyre kerültek. Előfordult az is, hogy a mondat végén nem csökkent megfelelő mértékben a dallam, ami rendkívül zavaró az emberi fül számára. Valószínűsíthetően ezen hibák miatt nem részesült egyértelmű előnyben módszerünk.

A tesztelők megjegyzései közül (amelyeket az F.3.2. függelék tartalmaz) érdemes kiemelni, hogy sokan túl hosszúnak tartották az egyes bekezdéseket, és így nehéz volt számukra összehasonlítani a két változatot. Mivel a jelenlegi teszt hosszabb szövegen próbálta vizsgálni az elért változatosságot, ezért ez elkerülhetetlen volt. A legzavaróbb dallambeli hibának azt tartották, amikor a mondatvégi intonáció nem volt megfelelő. Ezen kívül javasolták, hogy a hangsúlyok megfelelő elhelyezésére jobban figyeljünk oda. A hibák kiküszöbölése ezért egyértelmű célunk a továbbiakban.

Összefoglalás

A fejezetben összefoglalom a kutatás előzményeit, folyamatát és végső állapotát a saját munkámat elkülönítve.

A dolgozat elején elemeztem a diplomaterv-kiírást, meghatároztam munkám céljait és azok megvalósítási módját. Az 1. fejezetben először áttekintettem a beszédszintézis szakirodalmát. Ismertettem az emberi beszéd prozódiájának tulajdonságait, valamint a ma használatos szövegfelolvasók típusait és ezeknek működési elvét. Kitértem az alkalmazott prozódiai modellekre, ezek közül részletesebben ismertettem az adatvezérelt megvalósításokat, melyek nagyméretű beszédkorpusz alapján rendelik a bemeneti szöveghez a prozódiát. A fejezet végén a kutatáshoz szorosan kapcsolódó, prozódiai változatossággal foglalkozó szakirodalmat mutattam be.

A 2. fejezet elején ismertettem a jelenlegi szövegfelolvasó rendszerek egyik gyengéjét: az emberihez hasonló, változatos beszéd modellezésének hiányát. Olyan tulajdonságokat gyűjtöttem össze, amikkel egy változatos prozódiát megvalósító rendszernek feltétlenül rendelkeznie kell. Több alternatíva bemutatásával vizsgáltam ennek megoldási lehetőségét. Ezek közül egyet kiválasztva megterveztem egy rendszert, amely szövegfelolvasóhoz kapcsolva képes elérni a kívánt célt, vagyis a beszédszintetizátorok által előállított hang monotonitását csökkenti.

A tervezés után a megvalósítás részletes bemutatása következett a 3. fejezetben. Fejlesztői nézőpontból is ismertettem a létrehozott rendszert. Áttekintettem a beszédkorpuszok feldolgozásában végzett munkámat: 5 nagyméretű hanganyag-gyűjteményt vizsgáltam. F_0-minta és hangidőtartam-minta adatbázist hoztam létre a korpuszokból, majd az emberi beszéd változatosságát is elemeztem a természetes mondatokon. Bemutattam, hogyan történik a Profivox magyar nyelvű szövegfelolvasóban az alapfrekvencia beállítása a beszédszintézis során. Ismertettem a korábban megvalósított dallammásolási módszert, illetve ennek alkalmazását egy változatos beszédet létrehozni tudó programban. A megalkotott módszert a Profivox beszédszintetizátor rendszerbe illesztettem, hogy azt szélesebb körben használni lehessen.

A megvalósított rendszert többféle szempontból elemeztem (4. fejezet). Automatikus teszteket dolgoztam ki a lefedettségi arány és a változatok számának vizsgálatára. A kutatás so-

rán folyamatosan végzett meghallgatásos teszteket bemutattam, majd részletesen ismertettem a módszer értékelésére szolgáló szubjektív kísérletet.

Elért eredmények

Az irodalomkutatás során részletesen megismertem a mai szövegfelolvasók működési elvét. Közel 40 angol, német és magyar nyelvű forrást tanulmányozva alkalmam nyílt a különböző rendszerek összehasonlító elemzésére is.

A változatos beszéd létrehozására tervezett rendszert sikeresen megvalósítottam. Az elkészült beszédszintetizátor segítségével olyan beszéd állítható elő közel valós időben, amely kevésbé monoton a korábbi megvalósításokhoz képest. A módszer 5232 mondatból álló prozódiaminta adatbázist használ a szintetizálandó szövegek dallamának meghatározásához.

A módszer működését meghallgatásos vizsgálatok segítségével ellenőriztem. A kutatás során megvalósult szubjektív tesztek mindegyikében 10–20 tesztelő értékelte azokat a hangokat, amelyeket az adott fázisban létre tudtunk hozni. A korábbi tesztek eredményei azt mutatták, hogy a tesztet meghallgatók a dallam minták alapján történő másolásával létrehozott hangot preferálták a szabály alapú megvalósítással szemben. A legutóbbi teszt eredményeiből az derült ki, hogy módszerünkkel sok esetben változatos beszéd hozható létre, de ez még nem alkalmas éles rendszerben történő felhasználásra, mert sok hiba előfordul benne, amelyek az adatvezéreltségből fakadnak.

További kutatási lehetőségek

A kidolgozott módszer segítségével természetesebbé tehető a szövegfelolvasók által létrehozott prozódia. Ez az előny számos gyakorlati alkalmazásban használható, mint például SMS-, e-levél-, könyv-felolvasó, vagy telefonos tudakozó. A változatosabb prozódia főleg hosszú szövegek felolvasása esetén előnyös, hiszen ekkor zavaró leginkább a beszédszintetizátor monotonitása. A megvalósított rendszert tehát fel kell készíteni az éles használatra, hogy szélesebb körben is lehessen terjeszteni.

Érdekes lenne részletesebben megvizsgálni a felhasznált beszédkorpuszokat változatosság szempontjából, hiszen a jelen dolgozatban erre csak egy-egy mondat példáján volt mód. Pontosítani kellene a változatosság definícióját, hogy objektíven mérni lehessen azt.

A módszert más nyelvekben is lehetne alkalmazni. A finn és lengyel nyelv a magyarhoz hasonló fix hangsúlyt használ, ami alkalmassá teszi ezeket a módszer használatára. A magyar

hangsúlyozási szabály szerint mindig a szó első szótagján van a nyomaték. Más, változó hangsúlyt használó nyelvekre (pl. angol, német) a dallam minták alapján történő meghatározása bonyolultabb, de szintén lehetséges. Ehhez az alkalmazott hasonlósági mértéken kell változtatni úgy, hogy az a hangsúlyok szón belüli pozícióját is tartalmazza.

Azt az irányt is érdemes megvizsgálni, mi lenne, ha a prozódia többi komponensét (elsősorban a hangidőtartamokat) is mintákat tartalmazó adatbázis alapján hoznánk létre. Ennek megvalósításához az adatbázis már elkészült, de ki kell dolgozni egy módszert, amely a dallammal szinkronban tudja másolni a hangok időtartamait. Ezen részfeladat megoldásához a már elkészített rendszerek működésének érdemes utánanézni a szakirodalomban.

Fontos megjegyezni újra, hogy a tesztelők rossz minőségűnek ítélték azokat a mondatokat, amelyeknek végén szokatlanul magas volt a dallammenet vagy hangsúlyhibát tartalmaztak. A továbbiakban tehát javítani kell ezt is.

A változatosság koncepciója korpusz alapú, elemkiválasztásos beszédszintetizátorban is alkalmazható. Ennek megvalósításához részletesen meg kell ismerni az ilyen típusú szövegfelolvasók működési elvét. A hosszabb természetes beszéddarabok összefűzésével egyértelműen jobb minőségű szintetizált beszéd elérhető el.

Köszönetnyilvánítás

Ezúton mondok köszönetet konzulenseimnek, Dr. Németh Gézának és Dr. Fék Márknak a munkám során nyújtott segítségükért, hasznos tanácsaikért és észrevételeikért. 2 éven át tartó együttműködésünknek számos eredménye lett ezen diplomaterven kívül. Áldozatos munkájuk, folyamatos irányadásuk nélkül nem jöhetett volna létre a dolgozat. Köszönöm nekik, hogy munkájukkal megalapozták tudományos szemléletemet.

Köszönettel tartozom Zainkó Csabának a „Harang" hanganyag rendelkezésemre bocsátásáért. Kiss Gézának a fejlesztés folyamán a Profivoxba illesztés során nyújtott magyarázatait köszönöm. Bartalis István Mátyás a meghallgatásos vizsgálat kivitelezésével támogatta a megvalósított rendszer értékelését. Bőhm Tamás a vizsgálatok eredményeinek elemzésében segített sokat. A munka a BME TMIT Beszédtechnológiai Laboratóriumában készült, minden munkatárs segítségét köszönöm, aki hozzájárult ezen dolgozat létrejöttéhez.

A szubjektív teszteket elvégzőknek köszönöm, hogy meghallgatták és értékelték a hanganyagokat, valamint hasznos megjegyzéseikkel a további kutatási célok meghatározásában segítettek.

Külön köszönöm családomnak: szüleimnek, nagyszüleimnek, bátyámnak és barátnőmnek, hogy egyetemi éveim alatt folyamatosan támogattak, megteremtették számomra a diploma megszerzésének lehetőségét. Kitartásukkal és szorgalmukkal mindig jó példát mutattak.

Irodalomjegyzék

[1] 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.

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

[3] Multivox'4 Kis erőforrás igényű Magyar nyelvű szabad terjesztésű szövegfelolvasó szoftver. `http://fonetika.nytud.hu/mvox4/kezikonyv.pdf`, 2002.

[4] Wikipedia: Sprachsynthese, Diphonesynthese. `http://de.wikipedia.org/wiki/Sprachsynthese#Diphonesynthese`.

[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] Gábor Olaszy – Géza Németh – Péter Olaszi: Automatic prosody generation – a model for Hungarian. In *Eurospeech 2001* (konferenciaanyag), 1. köt. 2001, 525–528. p.

[9] Kim Silverman – Mary Beckman – John Pitrelli – Mori Ostendorf – Colin Wightman – Patti Price – Janet Pierrehumbert – Julia Hirschberg: ToBi: A standard for labelling English prosody. In *ICSLP 92* (konferenciaanyag), 2. köt. 1992, 867–870. p.

[10] Stefan Baumann – Martine Grice – Ralf Benzmüller: GToBi - a phonological system for the transcription of German intonation. In Stanislaw Puppel – Demenko Grazyna (szerk.): *Prosody 2000: Speech Recognition and Synthesis*. Krakow, Poland, 2000, Adam Mickiewicz University, Faculty of Modern Languages and Literature, 21–28. p.

[11] Maria E. Graba – Brechtje Post – William F. Nolan: Modelling Intonational Variation in English: The IViE system. In Stanislaw Puppel – Demenko Grazyna (szerk.): *Prosody 2000: Speech Recognition and Synthesis*. Krakow, Poland, 2000, Adam Mickiewicz University, Faculty of Modern Languages and Literature, 51–57. p.

[12] Németh Géza – Olaszy Gábor: Beszédinformációs rendszerek tantárgy előadás anyaga. BME TMIT, `http://speechlab.tmit.bme.hu/postnuke/modules.php?op=modload&name=Downloads&file=index&req=viewsdownload&sid=23`, 2005.

[13] Volker Strom: From text to prosody without ToBi. In *Interspeech 2002* (konferenciaanyag). 2002, 2081–2084. p.

[14] Jianhua Tao – Lianhong Cai – Herbert Tropf: An optimized neural network based prosody model of Chinese speech synthesis system. In *The 17th IEEE Region 10 International Conference on Computers, Communications, Control and Power Engineering* (konferenciaanyag), 1. köt. 2002, 477–480. p.

[15] Hiroya Fujisaki: Dynamic characteristics of voice fundamental frequency in speech and singing. In P. MacNeilage (szerk.): *The Production of Speech*. Berlin, 1983, Springer-Verlag, 39–47. p.

[16] Joram Meron: Prosodic unit selection using an imitation speech database. In *SSW4-2001* (konferenciaanyag). 2001, 113. p.

[17] Xuedong Huang – Alex Acero – Hsiao-Wuen Hon – Yun-Cheng Ju – Jingsong Liu – Scott Meridith – Mike Plumpe: Recent improvements on Microsoft's trainable text-to-speech system – Whistler. In *ICASSP 97* (konferenciaanyag). 1997, 959–962. p.

[18] Antoine Raux – Alan W. Black: A unit selection approach to F0 modeling and its application to emphasis. In *ASRU 2003* (konferenciaanyag). 2003, 700–705. p.

[19] The Festival speech synthesis system. `http://www.cstr.ed.ac.uk/projects/festival/`.

[20] Takashi Saito: Generating F0 contours by statistical manipulation of natural F0 shapes. E89-D. évf. (2006. március) 3. sz., *IEICE - Transactions on Information and Systems*, 1100–1106. p.

[21] Ignasi Iriondo – Joan Claudi Socoro – Francesc Alias: Prosody modelling of Spanish for expressive speech synthesis. In *ICASSP 07* (konferenciaanyag), 4. köt. 2007, 821–824. p.

[22] Minghui Dong – Kim-Teng Lua: An example-based approach for prosody generation in Chinese speech synthesis. In *ICSLP 2000* (konferenciaanyag). 2000, 303–307. p.

[23] Jan van Santen – Alexander Kain – Esther Klabbers – Taniya Mishra: Synthesis of prosody using multi-level unit sequences. 46. évf. (2005) 3-4. sz., *Speech Communication*, 365–375. p.

[24] Rolf Carlson: Synthesis: Modelling variability and constraints. In *Eurospeech 1991* (konferenciaanyag). 1991, 1043–1048. p.

[25] Thierry Dutoit: High-quality Text-to-Speech synthesis : an overview. 17. évf. (1997) 1. sz., *Journal of Electrical and Electronics Engineering, Australia: Special Issue on Speech Recognition and Synthesis*, 25–37. p.

[26] Elena Zvonik – Fred Cummins: Pause duration and variability in read texts. In *ICSLP 2002* (konferenciaanyag). 2002, 1109–1112. p.

[27] Nick Campbell: Developments in corpus-based speech synthesis: Approaching natural conversational speech. E88-D. évf. (2005. március) 3. sz., *IEICE - Transactions on Information and Systems*, 376–383. p.

[28] Min Chu – Yong Zhao – Eric Chang: Modeling stylized invariance and local variability of prosody in text-to-speech synhtesis. 48. évf. (2006), *Speech Communication*, 716–726. p.

[29] Dacheng Lin – Yong Zhao – Frank K. Soong – Min Chu – Jieyu Zhao: Iterative unit selection with unnatural prosody detection. In *Interspeech 2007* (konferenciaanyag). 2007, 2909–2912. p.

[30] Olaszy Gábor: A korpusz alapú beszédszintézis nyelvi, fonetikai kérdései. LXI. évf. (2006) 3. sz., *Híradástechnika*, 43–50. p.

[31] Géza Németh – Márk Fék – Tamás Gábor Csapó: Increasing prosodic variability of text-to-speech synthesizers. In *Interspeech 2007* (konferenciaanyag). 2007, 474–477. p.

[32] Péter Mihajlik – Tibor Révész – Péter Tatai: Phonetic transcription in Automatic Speech Recognition. 49. évf. (2002) 3-4. sz., *Acta Linguistica Hungarica*, 407–425. p.

[33] Paul Boersma – David Weenink: Praat: doing phonetics by computer, (version 4.6.34) [computer program]. `http://www.praat.org/`, 2006.

[34] Anne Tamm – Kálmán Abari – Gábor Olaszy: Accent assignment algorithm in Hungarian, based on syntactic analysis. In *Interspeech 2007* (konferenciaanyag). 2007, 466–469. p.

[35] The C# language. `http://msdn.microsoft.com/en-us/vcsharp/default.aspx`.

[36] Extensible Markup Language (XML). `http://www.w3.org/XML/`.

[37] Brian W. Kernighan – Dennis M. Ritchie: *A C Programozási Nyelv*. Budapest, 2003, Műszaki Könyvkiadó.

[38] Csapó Tamás Gábor – Németh Géza – Fék Márk: Szövegfelolvasó természetességének növelése. , *Híradástechnika*. Megjelenés alatt.

[39] Olaszy Gábor: Prozódiai szerkezetek jellemzése a hírfelolvasásban, a mesemondásban, a novella és a reklámok felolvasásában. In Gósy Mária (szerk.): *Beszédkutatás 2005*. Budapest, 2005, MTA Nyelvtudományi Intézet, Kempelen Farkas Beszédkutató Laboratórium, 21–50. p.

Az internetes források ellenőrzésének utolsó dátuma: 2008. május 16.

Függelék

F.1. Adatbázisok

F.1.1. F_0-minta adatbázis XML-ben

```
<?xml version="1.0" standalone="yes"?>
<Idojaras>
  <Phrase>
    <Filename>3412</Filename>
    <nWords>4</nWords>
    <nSyllables>7</nSyllables>
    <Syllables>1 1 3 2</Syllables>
    <Pitches>195 255 234 229 205 178 214</Pitches>
    <Position>F</Position>
    <Stress>- W N N</Stress>
    <MeanPitch>215</MeanPitch>
  </Phrase>
  <Phrase>
    <Filename>3412</Filename>
    <nWords>3</nWords>
    <nSyllables>8</nSyllables>
    <Syllables>1 3 4</Syllables>
    <Pitches>195 222 207 189 166 145 142 135</Pitches>
    <Position>L</Position>
    <Stress>- W N</Stress>
    <MeanPitch>175</MeanPitch>
  </Phrase>
</Idojaras>
```

F.1.2. Hangidőtartam-minta adatbázis XML-ben

```
<?xml version="1.0"?>
<ArrayOfWord xmlns:xsi="http://www.w3.org/2001/XMLSchema-
 instance" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <Word>
   <Filename>2948</Filename>
    <Text>A:ramlA:S:al</Text>
    <durations>
      <int>156</int>
      <int>33</int>
      <int>89</int>
      <int>61</int>
      <int>39</int>
      <int>100</int>
      <int>128</int>
      <int>71</int>
      <int>50</int>
    </durations>
  </Word>
  <Word>
   <Filename>2999</Filename>
    <Text>napoS</Text>
    <durations>
      <int>61</int>
      <int>80</int>
      <int>100</int>
      <int>51</int>
      <int>70</int>
    </durations>
  </Word>
</ArrayOfWord>
```

F.2. Forráskód

F.2.1. Módszerünkben használt C függvények definíciói

```
/**
 * Intonációs mátrix felbontása:
 * - a bemeneti mondat frázisokra bontása
 */
short countImf(mv5_im_s *matrix,
 unsigned short matrix_size, imf_phrase *phrases);

/**
 * Minták keresése:
 * - F0-minta adatbázisból
 * - szótagszerkezet, pozíció, hangsúlyszerkezet
 *   alapján
 */
short searchPhrase(corpus_phrase *corp_phr,
 short corp_phr_count, unsigned short nWords,
 unsigned short nSyllables, unsigned short *syllables,
 char position, char *stress, corpus_phrase *result);

/**
 * Minta-sorozat kiválasztása:
 * - véletlen választás a lehetőségek közül
 */
short selectF0Contour(corpus_phrase **phrases,
 unsigned short *phrases_count, unsigned short count,
 corpus_phrase *results);

/**
 * Dallam átültetése:
 * - kiválasztott minta-sorozat dallamának
 *   bemenethez illesztése
 * - F0-normalizálás
```

```
 */
short copyF0(unsigned short *pitches,
 unsigned short nSyllables, unsigned short first_syl,
 mv5_im_s *matrix, unsigned short matrix_size);
```

F.2.2. Profivoxba illesztéshez használt C függvények definíciói

```
/**
 * Inicializálás:
 * - működési paraméterek beolvasása
 * - adatbázis beolvasása
 * - memóriafoglalás
 */
int init(ProsVarSession *session);

/**
 * Változatos prozódia megvalósítása:
 * - intonációs mátrix felbontása
 * - minták keresése
 * - minta-sorozat kiválasztása
 * - dallam átültetése
 */
int prosody_variation(mv5_im_s *matrix_in,
 unsigned short matrix_size, mv5_im_s **pmatrix_out,
 ProsVarSession session);

/**
 * Erőforrás-felszabadítás:
 * - lefoglalt memória felszabadítása
 */
int destroy(ProsVarSession session);
```

F.3. Meghallgatásos teszt

F.3.1. Tesztelt mondatok

Időjárás-előrejelzés 1.

Szombat estig a mediterrán ciklon délebbre helyeződik, a Kárpát-medence időjárását egy nyugatról közeledő érintőleges frontrendszer alakítja. Eleinte többnyire nedves, szombaton azonban már egyre szárazabb levegő áramlik térségünk fölé, holnap főként nyugaton már több órára kisüt a nap. Csütörtökön eleinte többnyire erősen felhős volt az ég, és északkeleten gyenge havazás is előfordult, majd főként a középső és keleti országrészben átmenetileg felszakadozott a felhőzet, és 4-6 órára a nap is kisütött. Délután a Dunántúlon fordult elő szórványosan eső, záporeső. A hőmérséklet csúcsértéke 10 és 16 fok között alakult, csak északkeleten volt néhol hidegebb. Éjszaka mindenütt megvastagodott a felhőzet, és elszórtan esett az eső.

Időjárás-előrejelzés 2.

Szombaton dél felől erősen megnövekedett a felhőzet, de az ország északi felén még 2-4 órára kisütött a nap. Északnyugaton és a főváros térségében szórványosan, másutt mindenütt esett az eső. Az északnyugati szelet többfelé erős, északnyugaton és a Nyírségben helyenként viharos lökések kísérték. A hőmérséklet csúcsértéke többnyire 15 és 20, nyugaton 11 és 15 fok között változott. Éjszaka nyugat felől fokozatosan megszűnt a csapadék, de keleten többfelé volt még eső, zápor, zivatar. A ma reggelig lehullott csapadék mennyisége általában csapadéknyom és 10, délen és keleten 10 és 20 milliméter között alakult, de Pécs környékéről 25, 26 millimétert jelentettek.

Időjárás-előrejelzés 3.

A főváros külterületén csütörtökön a hőmérséklet középértéke 12,1 fok volt, ez 0,3 fokkal alacsonyabb, mint a sokévi átlag. Ma 12 órakor Budapesten a hőmérséklet 17 fok volt, a tengerszintre átszámított légnyomás 1024 hektopaszkál, gyengén süllyed. A Dunántúlon erősen megnövekszik a felhőzet, és szombaton is többször lesz felhős az ég, de emellett ma is, holnap is néhány órára kisüt a nap. Az ország többi részén túlnyomóan napos idő várható, csak időnként lesz több a felhő. Elsősorban nyugaton valószínű elszórtan eső, zápor. Az északi szél a Dunántúlon megélénkül, holnap ott egy-egy erős széllökés is lehet. A legalacsonyabb éjszakai hőmérséklet 3 és 8, a legmagasabb nappali hőmérséklet szombaton 16 és 21 fok között alakul.

Hír: BKV

Kevesebb utast számol össze a BKV, mint amennyi a valóságban a járművein utazik, legalábbis erre engednek következtetni azok a levelek és egy videófelvétel, amelyek az Index szerkesztőségébe is eljutottak. A BKV állítja, bár a mérések sohasem pontosak, alaptalanok az észrevételek, az utasok félreértették, amikor nem számolták meg őket. A BKV kétféle módszerrel számolja rendszeresen utasait, de áprilisban alkalmi felmérés is történt a Gazdasági és Közlekedési Minisztérium megbízására. A buszokon és trolikon a légrugóba épített műszerrel, a kötöttpályás járműveken (villamos, hév, metró) utasszámláló dolgozókkal számolják az utasokat. Áprilisban a GKM megbízásából bizonyos agglomerációs (Budapest közigazgatási határát átlépő) járatokon alkalmi jelleggel élőmunkás módszerrel számolták az utasokat.

Könyv: Egri csillagok

A bombák, kalácsok és koszorúk sercegve fogtak lángot. Nagy, szikrázó ívekben száz meg száz tüzes szivárvány. De az ostromlók elszántan törtetnek, kapaszkodnak, erőlködnek, tolakodnak fel a falakra. Az ostromlétrák gyorsan kapcsolódnak. A létrákon mókusokként szöknek fölfelé a janicsárok, az aszabok és a gyalogsággá vált lovasság. Csattog fenn a csákány a létrák kapcsán.

F.3.2. A tesztelők megjegyzései

- „Sok mondat van egy mintában. Nehéz párhuzamot vonni. Nagyon kicsi a különbség a két minta között, nehéz döntést hozni.”
- „Kicsit talán hosszúak az egyes bekezdések, mire a másodikat is meghallgatja az ember, már nem emlékszik az elsőre. Ez persze ismételt hallgatással kiküszöbölhető.”
- „Volt olyan rész, ahol a mondat végi intonáció nem volt megfelelő, ezért nem lehetett eldönteni, hogy mondat vége van-e, vagy vessző. A következő mondatból derült ki csak, hogy ez már nem annak a mondatnak a folytatása. Ez nem a standard Profivoxban volt.”
- „A változóbb hanglejtésű az mintha mindig zajosabb, ekhósabb lenne. Ez teszi az egyébként kellemes hanglejtésű élményt kellemetlenné.”
- „Túl hosszúak a bejátszások, jobb lett volna mondatonként összehasonlítani, így is a media player csúszkáját húzogattam, hogy mondatról mondatra hasonlítsam össze a bejátszásokat, tehát az eredmény ugyan az, csak a tesztelő dolgát nehezíti meg.”

- „Nekem kissé hosszúak az összehasonlítandó szakaszok. Bár úgy többször kell kattingatni, értékelni, de két-három mondatos szakaszokra bontva jobban összehasonlíthatók volnának, és bizonyára előfordulna annyiban néhány eltérő rész. Nagyon sokat dobna az új változaton, ha az időtartam-struktúrát is átvenné a természetes mondatból valahogyan. És persze a hangsúlyok figyelembevétele is fontos lépés lesz.”
- „Többnyire nehéz volt eldönteni, melyik a változatosabb. Talán a két rövidebb mondatnál egyértelműbb lesz az eredmény.”
- „A tesztet a JAWS for Windows 6.20 képernyőolvasó programmal végeztem, amellyel nem találtam meg az ismételt meghallgatást lehetővé tevő funkció aktiválási lehetőségét a tesztszövegek elhangzását követően. Ezért döntéseimet egyszeri meghallgatás alapján hoztam meg. Lehet, hogy választásaim finomodtak volna, ha többször is végig tudtam volna hallgatni a felolvasásokat.”
- „A felvételek iszonyatosan hosszúak.”

F.4. A CD-melléklet tartalma

F.4.1. Adatbázisok

`Árlista_F0.xml`	„Árlista” korpuszból létrehozott F_0-minta adatbázis
`Árlista_idő.xml`	„Árlista” korpuszból létrehozott hangidőtartam-minta adatbázis
`Harang_F0.xml`	„Harang” korpuszból létrehozott F_0-minta adatbázis
`Harang_idő.xml`	„Harang” korpuszból létrehozott hangidőtartam-minta adatbázis
`Időjárás_F0.xml`	„Időjárás” korpuszból létrehozott F_0-minta adatbázis
`Időjárás_idő.xml`	„Időjárás” korpuszból létrehozott hangidőtartam-minta adatbázis
`Prompt_F0.xml`	„Prompt” korpuszból létrehozott F_0-minta adatbázis
`Prompt_idő.xml`	„Prompt” korpuszból létrehozott hangidőtartam-minta adatbázis
`Vonat_F0.xml`	„Vonat” korpuszból létrehozott F_0-minta adatbázis
`Vonat_idő.xml`	„Vonat” korpuszból létrehozott hangidőtartam-minta adatbázis

F.4.2. A vizsgálatokban felhasznált anyagok

Lefedettségi vizsgálatokhoz tartozó szövegek

`Árlista_01-20.txt`	„Árlista” korpuszból származó mondatok
`Időjárás_01-20.txt`	„Időjárás” korpuszból származó mondatok
`Prompt_01-30.txt`	„Prompt” korpuszból származó mondatok
`Vonat_01-25.txt`	„Vonat” korpuszból származó mondatok

`Előrejelzés1_01-22.txt`	„Előrejelzés/1” szöveg mondatai
`Előrejelzés2_01-19.txt`	„Előrejelzés/2” szöveg mondatai
`Előrejelzés3_01-20.txt`	„Előrejelzés/3” szöveg mondatai
`Hír_01-15.txt`	„Hír” szöveg mondatai
`Könyv_01-16.txt`	„Könyv” szöveg mondatai

Meghallgatásos teszt anyaga

`Előrejelzés1.txt`	„Előrejelzés/1” bekezdés szövege
`Előrejelzés1.mp3`	„Előrejelzés/1” bekezdéspár hanganyaga (először a szabály alapú hírfelolvasó modullal készült változat, utána a változatos dallamú)
`Előrejelzés2.txt`	„Előrejelzés/2” bekezdés szövege
`Előrejelzés2.mp3`	„Előrejelzés/2” bekezdéspár hanganyaga (először a szabály alapú hírfelolvasó modullal készült változat, utána a változatos dallamú)
`Előrejelzés3.txt`	„Előrejelzés/3” bekezdés szövege
`Előrejelzés3.mp3`	„Előrejelzés/3” bekezdéspár hanganyaga (először a szabály alapú hírfelolvasó modullal készült változat, utána a változatos dallamú)
`Hír.txt`	„Hír” bekezdés szövege
`Hír.mp3`	„Hír” bekezdéspár hanganyaga (először a szabály alapú hírfelolvasó modullal készült változat, utána a változatos dallamú)
`Könyv.txt`	„Könyv” bekezdés szövege
`Könyv.mp3`	„Könyv” bekezdépár hanganyaga (először a szabály alapú hírfelolvasó modullal készült változat, utána a változatos dallamú)
`teszt.html`	A meghallgatásos teszt nyitóoldala
`teszt.mp3`	A nyitóoldalon alkalmazott teszthang