Beszédszintetizátor prozódiai változatosságának növelése

TDK dolgozat

Készítette:
Csapó Tamás Gábor
csapszi@sch.bme.hu

Konzulensek:

Dr. Németh Géza
nemeth@tmit.bme.hu

Dr. Fék Márk
fek@tmit.bme.hu

2007. november

Tartalomjegyzék

Ábrák jegyzéke

Táblázatok jegyzéke

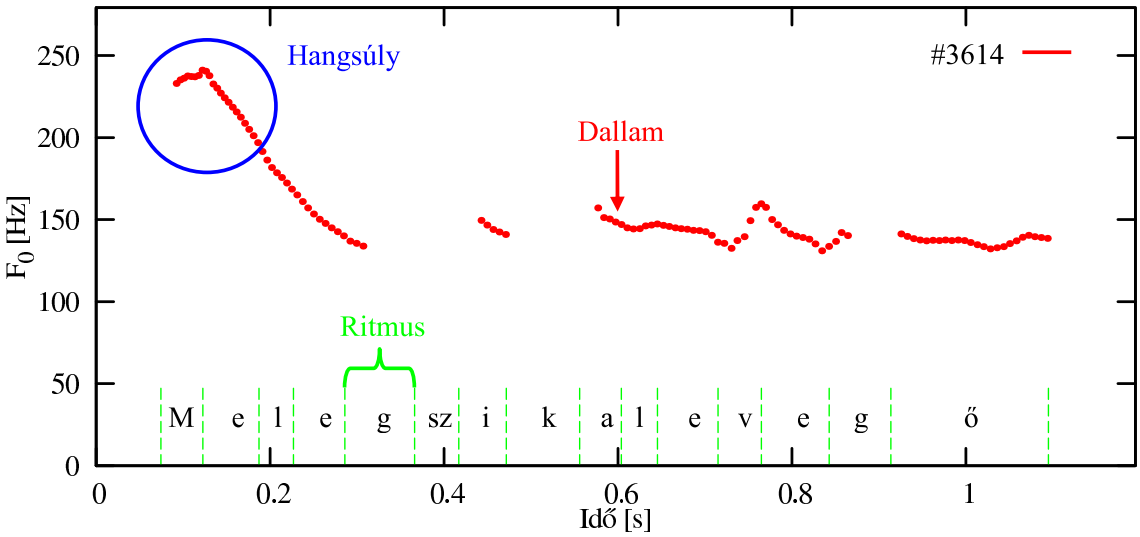
1. Bevezetés

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 esetében), továbbá ha az igénybe vett szolgáltatás telefonvonalon keresztül érhető el.

A beszédszintézis rendszerek minőségét az alapján ítélik meg, hogy az általuk keltett beszéd mennyire hasonlít az emberi beszédre. 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 év során jelentős kezdeti eredményeket értünk el a beszédszintetizátorok prozódiájának változatosabbá és természetesebbé tétele területén nagyméretű természetes beszédkorpusz felhasználásával. A jelen dolgozatban először a témához tartozó szakirodalmat tekintjük át (2. fejezet), majd a korábban kidolgozott módszer továbbfejlesztési irányait ismertetjük (3. fejezet). Egyrészt finomítottunk a távolságmértéken, amelyet a bemeneti szöveg és a beszédkorpuszban szereplő természetes mondatból származó prozódia társításához használunk, másrészt a korábbi módszerek eredményességét nagyobb beszédkorpuszon is vizsgáltuk. Kidolgoztunk egy módszert a dallam másolására kisebb egységek alapján. Végül bemutatjuk, hogyan történt a módszerünkkel előállított mondatok minőségének értékelése, és a kapott eredményeket is vizsgáljuk (4. fejezet).

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



1. ábra. A prozódia három összetevője. (Mondat: *„Melegszik a levegő”*)

2. Elméleti háttér

A következőekben bemutatásra kerülnek a dolgozat megértéséhez szükséges alapfogalmak. A 2.1. alfejezet az emberi beszéd prozódiájának legfontosabb összetevőit mutatja be. A beszédszintetizátorok három főbb generációjáról a 2.2. alfejezetben olvashatunk. A 2.3. alfejezet a prozódia megfelelő modellezésével foglalkozik. Ezek az alfejezetek egy korábbi munkánkban [1][1] már részletesen bemutatásra kerültek, így a jelenlegi dolgozatban csak a leglényegesebb gondolatok szerepelnek belőlük.

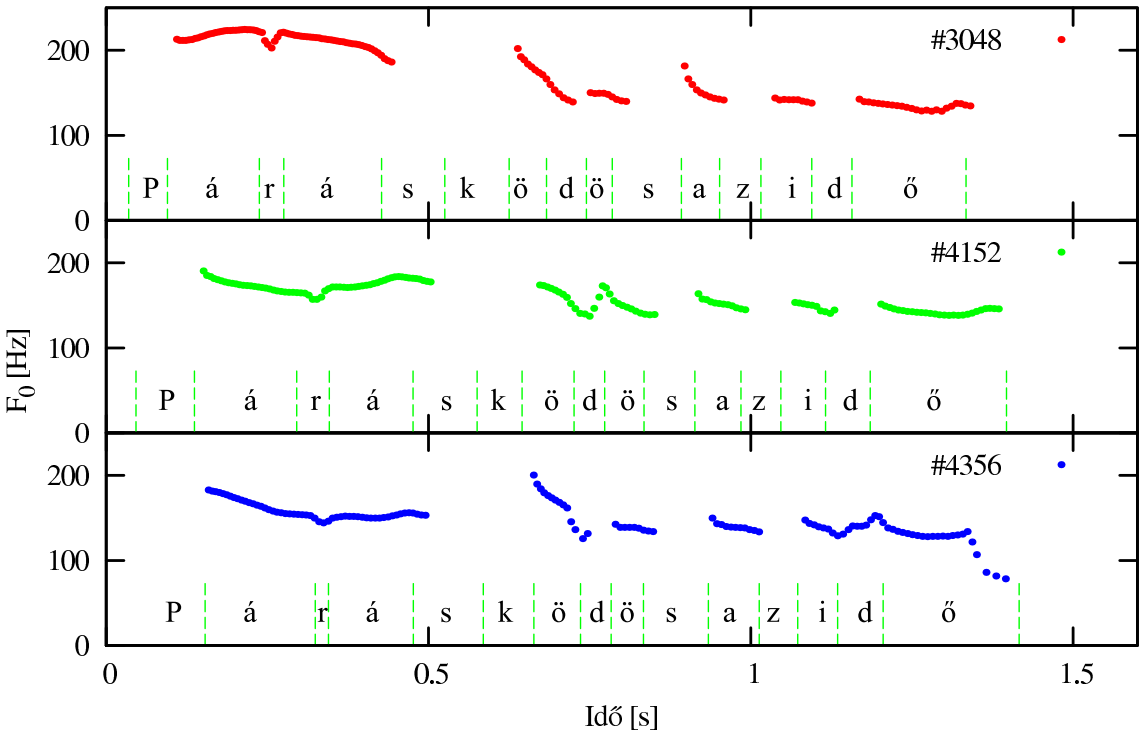
A 2. fejezet további részeiben a beszédszintézis szakirodalmában ismert korpusz alapú prozódia-generálási módszerek néhány fajtáját (2.4. alfejezet), valamint egy, a prozódia változékony részeivel foglalkozó tanulmányt (2.5. alfejezet) mutatunk be.

2.1. A prozódia három összetevője

A folyamatos emberi beszéd prozódiája három fő részből áll: dallam, hangsúly, ritmus. Ezek objektív és szubjektív paraméterekkel is jellemezhetőek. Objektív paraméteren a gép által mérhető adatokat, míg szubjektív paraméteren az emberi érzékszervekkel felfogható részeket értjük.

Dallam A dallamra [2, 242. oldal], vagyis a beszéd hangmagasságára jellemző objektív paraméter az alapfrekvencia (F_0), vagyis a zönge változása az időben. Az emberi beszéd dallama több szintre bontható fel. A legmagasabb, szupraszegmentális szint határozza meg a mondatok modalitását: kijelentő, kérdő, felkiáltó, óhajtó, felszólító. Ehhez kapcsolódóan a beszéd dallama lehet emelkedő, ereszkedő vagy lebegő. A középső szint a szó- és szótagszintű alapfrekvencia-változásokat foglalja magában. A legalacsonyabb (szegmentális) szinten

[1] A tavalyi TDK dolgozat a BME I épület könyvtárában érhető el.



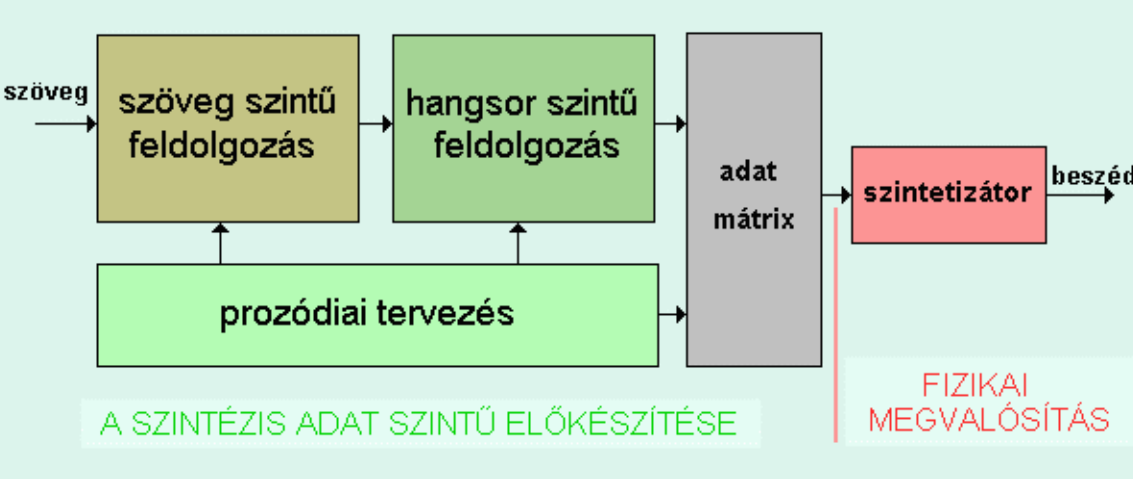
2. ábra. Prozódiai változatosság az emberi beszédben. (Mondat: *„Párás, ködös az idő”*).

a mikrointonáció jellemzői jelennek meg. Az 1. ábrán egy mondat dallamát, F_0-menetét láthatjuk. Az alapfrekvencia csak a beszéd zöngés részein értelmezett, ezért nem folytonos az F_0-görbe.

Hangsúly A hangsúlyozás [2, 246. oldal] nem más, mint nyomaték helyezése egy mondatrészre, szóra vagy szótagra. A hangsúly három fizikai paraméterrel jellemezhető: F_0 emelés, időtartam nyújtás, és intenzitás növelés. A magyar hangsúlyozási szabály szerint a szavak első szótagja a hangsúlyos. Más nyelveknél ez nem feltétlenül van így, angolban például változó helyen lehet egy-egy szavon belül a nyomaték. Az 1. ábra egy olyan hangsúlyt mutat a mondat elején, amely láthatóan az alapfrekvenia emelésével lett megvalósítva.

Ritmus A ritmus [2, 249. oldal] az egyes beszédhangok hosszát, a beszéd ritmikáját és a szünetek tartásának módját jelenti. A folyamatos beszédben az egyes hangokat hol gyorsabban, hol lassabban ejtjük, a hangidőtartamok változása ilyenkor 10-20% is lehet. Az 1. ábrán a *„Melegszik a levegő”* mondat hangjainak időtartamai is láthatóak.

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 a a 2. ábra is mutatja. Az ábrán a *„Párás, ködös az idő”* mondat három különböző kiejtési módját láthatjuk.



3. ábra. Példa egy beszédszintetizátor felépítésére. Forrás: [2, 303. oldal]).

A három változat hasonló, de mégis észrevehető különbség van közöttük az alapfrekvencia-menetben és a hangok időtartamában.

2.2. Beszédszintetizátorok

A beszédszintézis nem más, mint emberi beszéd előállítása mesterséges módon, tipikusan számítógép segítségével. Amennyiben a bemenet írott szöveg, szövegfelolvasóról, TTS[2] rendszerről beszélünk. Ezt a szöveget a beszédszintetizátor különböző lépéseken keresztül alakítja át emberi beszéddé. A 3. ábrán egy ilyen TTS-re látható példa. Először szimbolikus információt hoz létre a rendszer a bemeneti szöveg alapján, amit az ábra bal oldala „A szintézis adat szintű előkészítésé”-vel jelez. Ezen belül található a megfelelő prozódia meghatározása (vagyis a dallammenet, intenzitás és időtartamok hozzárendelése a bemeneti szöveghez, azaz „Prozódiai tervezés”), ami dolgozatunk fő témája. A szimbolikus információ alapján hozza létre a TTS a „Fizikai megvalósítás” során a kimeneten a beszédet.

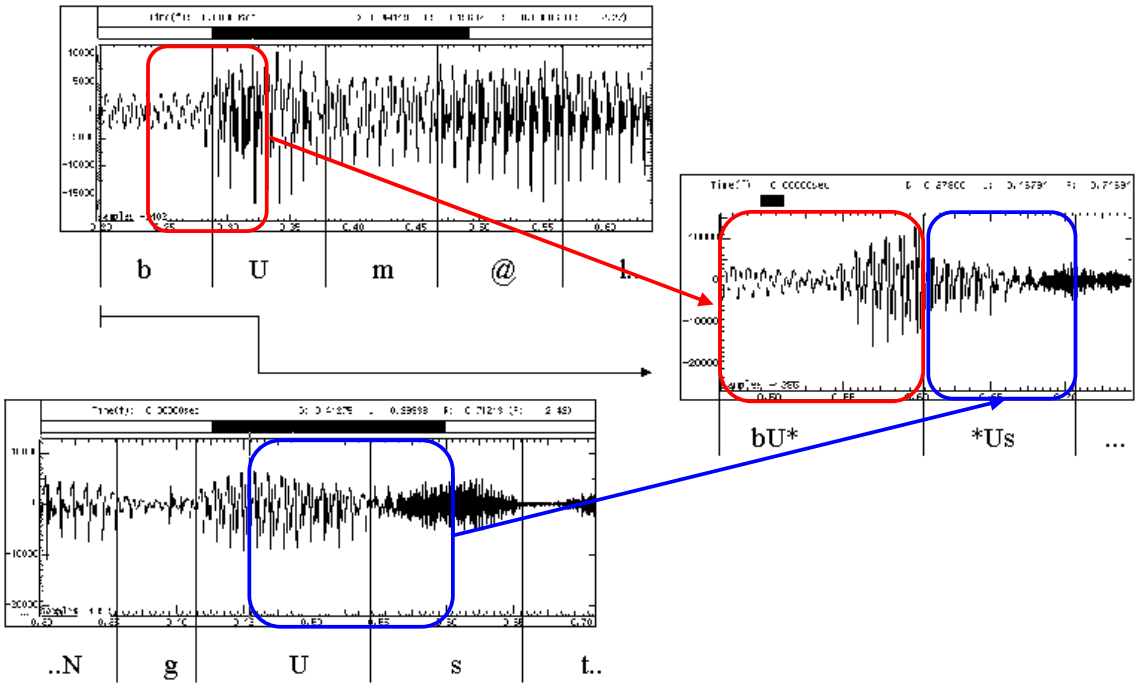
A beszédszintetizátoroknak három főbb generációját különböztetjük meg, melyeket most Fék és társai munkája alapján ismertetünk [3].

Formánsszintézis A formánsszintézis volt az első olyan technológia, mellyel szöveget automatikusan érthető beszéddé lehetett alakítani. Ezek az emberi beszéd formánsainak[3] modellezésével próbálták létrehozni a beszédhangot. Az ilyen rendszerek hangzása az érthetőség mellett meglehetősen „robotos”, ami háttérbe szorította őket.

Elemösszefűzés A 20. század elején végzett kísérletek megmutatták, hogy a beszédszintézis érthetőségéért a hangátmenetek természetessége felelős. Az elemösszefűzéses beszédszintézis során természetes beszédből kivágott hullámforma elemeket fűznek össze. Attól függően

[2] Text-To-Speech

[3] A formáns az emberi beszédhang jellegzetes színét adó, rezonanciás úton felerősített felhangtartomány.



4. ábra. Német nyelvű diád elemek összefűzése: a *„verbummeln”*-ből /bU/ és *„Languste”*-ból /Us/ összefűzésével előáll a *„Bus”* szó. Forrás: [4].

különböztetjük meg az elemösszefűzéses rendszereket, hogy mekkora elemeket fűznek össze: a diádos rendszerekben az elem a két félhang (vagyis egy hangátmenet, pl. *a-b, a-c*), a triádos rendszerekben pedig környezetfüggő hangok (pl. *a-b-a, a-c-a*) tárolása történik meg. Ezek akár vegyesen is alkalmazásra kerülhetnek egy rendszeren belül. A 4. ábra a diádok összefűzésére mutat példát: két különböző hangkörnyezetből kivágott diád elem egymás után helyezésével jön létre a „Bus” szó. Az elemek összefűzése után az előálló beszéd megfelelő prozódiájáról is gondoskodni kell jelfeldolgozási módszerek segítségével. Az ily módon létrehozott beszéd jól érthető ugyan, de még nem természetes hangzású.

A dolgozat során a BME-TMIT[4]-en kifejlesztett Profivox [5] beszédszintetizátort használtuk tesztjeink elvégzésére. A Profivox magyar nyelvű beszédszintetizátor, aminek legújabb változata az 1444 diád mellett 6000 triád-elemet is tartalmaz. A rendszer több felolvasó hanggal rendelkezik, amik közül egy férfi változatot alkalmaztunk.

Elemkiválasztás Az elemösszefűzéses technológia továbbfejlesztése a korpusz alapú, elemkiválassztásos beszédszintézis. Az újdonság itt egyrészt az, hogy nagyobb beszédadatbázis áll rendelkezésre, amiben egy-egy elem többször, többféle formában is előfordulhat, másrészt az elemek hosszabbak: szavak vagy akár szókapcsolatok is lehetnek. Az elemek összefűzése során a rendszer minél hosszabb elemeket keres az adatbázisban, amik a bemeneti szöveghez illeszkednek. Mivel a diádos/triádos rendszerekhez képest az elemek hosszabbak, így kevesebb összefűzési pont lesz a létrehozott beszédben, ami a természetesség növeléséhez vezet.

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

2.3. Prozódiai modellek

A 3. ábrán látható módon a bemeneti szövegből kimeneti hang előállítása során egy beszédszintetizátorban szükség van a prozódiai paraméterek megfelelő beállítására, a „prozódia tervezés”-re. Az efféle szimbolikus információ szövegből történő származtatására sokféle modell ismert a szakirodalomban, melyeket röviden ismertetünk.

Leíró jellegű modellek A leíró jellegű modellek célja, hogy az intonációt címkék segítségével írják le. Az egyik ilyen rendszerben, a ToBi[5]-ban [6] ezek a címkék a jellegzetes alapfrekvencia-változásokat jelölik: magas (High, H), alacsony (Low, L), szóhangsúly (L*), frázishangsúly (H-). A ToBi az angol nyelv címkézésére alkalmas, más nyelvekre különböző kiterjesztéseit alkalmazzák. Az alapvető probléma a leíró jelleű modellekkel, hogy a paraméterek származtatása csak kézi vagy félautomatikus módszerekkel oldható meg, ami drága és időigényes.

Szabály alapú modellek A prozódia modellezése szabályok segítségével is történhet [7, 3. fejezet, 32. oldal]. Ekkor a szöveg egyes részeihez (mondat, szó, szótag, hang) szabályokat rendelünk, melyek a létrehozandó dallammenetet definiálják (pl. hangsúly a mondat elején, alapfrekvencia-csökkentés a mondat végén). A szabályok ember által definiáltak, így megalkotásuk nehéz és időigényes, de ha hiba merül fel bennük, könnyen javíthatóak. A természetes nyelvek ugyanakkor nem reguláris szerkezetűek, így nem írhatóak le teljesen szabályok segítségével, mert mindig lesznek kivételek.

A szabály alapú modellek nagy előnye abban rejlik, hogy kiszámíthatóak: mindig hasonó minőségű prozódiát tudnak létrehozni. Ez azért fontos, mert az ember nehezen tűri a változást, ha egy bizonyos minőséget már megszokott.

Gépi tanulás Gépi tanulással (machine learning) [7, 3. fejezet, 32. oldal] úgy lehet prozódiai modellt létrehozni, hogy valamilyen nagyméretű beszédadatbázisból megpróbáljuk kinyerni a természetes beszéd tulajdonságait. Például ha van egy címkézett szöveggel rendelkező beszédkorpuszunk, neurális háló segítségével a rendszer következtetni tud a beszéd akusztikai paramétereire korrelációk, összefüggések keresésével. Ezen adatvezérelt modellek hátránya, hogy a megfelelő adatbázis elkészítéséhez akár több millió adatot kell kézzel felcímkézni. Előny viszont, hogy a prozódiát leíró szabályok megalkotása nem kézzel, hanem automatikus módszerekkel történhet.

Szuperpozíciós modellek A szuperpozíciós modelleknek az a fő jellemzője, hogy a prozódia összetevőinek különböző szintű megvalósításait (pl. mondat-, szó-, hangszint) adják össze, vagyis szuperponálják egymásra. A szintek modellezése külön-külön történik, pl. először meghatározva a mondatdallamot (emelkedő, egyenletes, eső), utána a szó- vagy szótagszintű hangsúlyokat, végül a mikrointonációs változásokat.

[5] Tones and Break indeces

A gyakorlatban ezeket a prozódiai modelleket általában egymással ötvözve használják. A dolgozatban alkalmazott Provivox beszédszintetizátor szabály alapú, szuperpozíciós modellel rendelkezik [8].

2.4. Prozódia másolása korpusz alapján

Számos olyan módszer ismert a beszédszintézis szakirodalmában, mely a megfelelő prozódia generálásával foglalkozik. A következőekben bemutatásra kerül néhány ezek közül, melyeknek közös jellemzője, hogy az adott bemeneti szöveghez tartozó prozódiát valamilyen természetes beszédből álló adatbázis alapján hozzák létre. Az emberihez hasonló dallammenet létrehozása azzal garantálható, hogy a szintetizálandó mondat alapfrekvencia-menetét az adatbázisból vett kisebb-nagyobb elemek segítségével határozzák meg.

2.4.1. Példa alapú prozódia generálás

A sokféle megvalósítás egyike Dong és Lua nevéhez fűződik [9], melyet ők példa alapú prozódia generálásnak neveznek. Prozódiai adatbázisként valódi beszédből vett mondatokat használnak, amiknek az F_0-menetét három részre bontják: mondatszintű prozódia, prozódia-minta frázis[6] szerint, és szótagszintű prozódia.

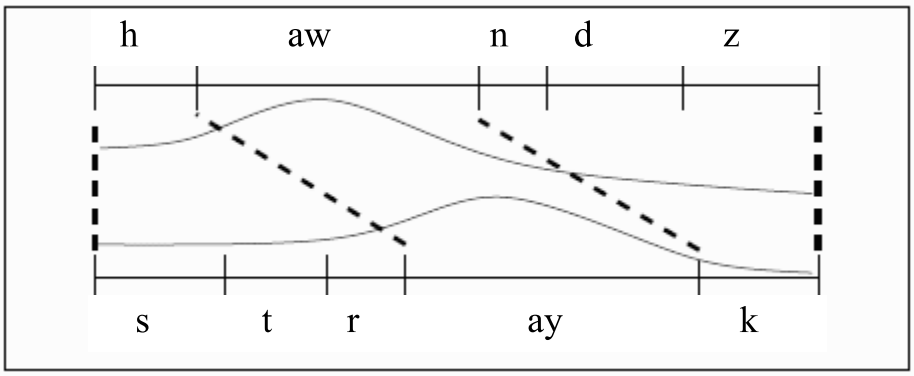
Egy adott szintetizálandó mondatot először szavakra bontanak, majd szófaj-analízist végeznek rajta. Az egy-egy szótaghoz tartozó meghatározandó dallammenetet és időtartamokat statisztikai módszerek segítségével hozzák létre. A példa-korpusz elkészítése során felmérik az adatbázisbeli szótagok F_0 értékeit, majd hangkörnyezet (előző, jelenlegi és következő szótag) alapján csoportosítják, és egy táblázatban tárolják ezeket.

A beszéd szintézisének lépései:

1. szöveg analízis
2. szótag prozódiájának meghatározása
3. frázis prozódiájának meghatározása
4. mondat prozódiájának meghatározása

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.

[6] A frázis, más néven prozódiai egység az a szócsoport, amit szóban egyben, szünet nélkül mondunk. Írásban általában vessző határolja a prozódiai egységeket.



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: [10]

2.4.2. Prozódiai elemkiválasztás

Egy másfajta próbálkozás Merontól származik [10]. A korábban elkészített, szabály alapú beszédszintetizátor rendszerüket egészítették ki egy olyan modullal, ami a létrehozott prozódia természetességét javítja. Azért őrizték meg a szabály alapú rendszert, mert ennek előnye a kiszámíthatóság, vagyis mindig hasonló minőségű prozódia készíthető vele. A szakirodalomban ismert teljesen korpusz alapú prozódia modellező módszerek is jó minőséget tudnak megvalósítani, de egyes esetekben természetellenes dallammenetet állítanak elő, ami nagyon zavaró lehet. Ebben a munkábn egyesítik a szabály alapú eljárások robosztusságát és a korpusz alapú módszerek természetességét.

A prozódia generálása tehát két részben történik: a szintetizálandó mondat szövege alapján a szabály alapú metódus dönt a létrehozandó intonáció események helyéről, ezáltal egy szimbolikus információt produkálva. A végső dallammenet meghatározása korpusz alapon történik, a természetes beszéd sajátosságait utánozva. 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, ami a módszerrel legtermészetesebben előállítható.

A beszédkorpusz létrehozásához egy új technikát alkalmaznak: az adatbázis felvételekor megkérik a beszélőt, hogy a szabály alapú TTS-hez hasonlóan beszéljen, imitálja azt. Először a bemondó meghallgat egy szintetizált mondatot, majd megpróbálja azt ugyanabban a stílusban elismételni. Ennek előnye, hogy csökken a korpuszban lévő beszédhang és a TTS-ben lévő szabályok közti különbség, és egyszerűbbé válik a természetes prozódia másolása. A hang felvétele után automatikus módszerekkel felcímkézik azt, és szótagokra bontják. Az egyes szótagokhoz három F_0 értéket tárolnak: a magánhangzó előttit, a magánhangzó közepén és utána lévő alapfrekvencia értéket. Ezen kívül tárolásra kerülnek a hangsúlyok helyzetével kapcsolatos információk, valamint a megelőző és rákövetkező szótag tulajdonságai is.

A szótagokra bontás előnye, hogy szinte tetszőleges bemeneti mondathoz lehet találni

prozódiai mintákat. A korábbi módszerekben [11] 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 használatával egyértelműen Meron konstrukciója hatásosabb. Ha egy szöveghez teljes mondatnyi prozódia-mintát talál egyben, akkor azt használja, azonban ennek hiánya esetén kisebb, különböző mondatokból származó természetes részekből is össze tudja rakni a prozódiát.

Meron konstrukciójának lépései:

1. intonáció események meghatározása szabály alapon
2. illeszkedő szótagsorozatok keresése
3. összefűzési költség számítása
4. elemösszefűzés
5. időzítés beállítása

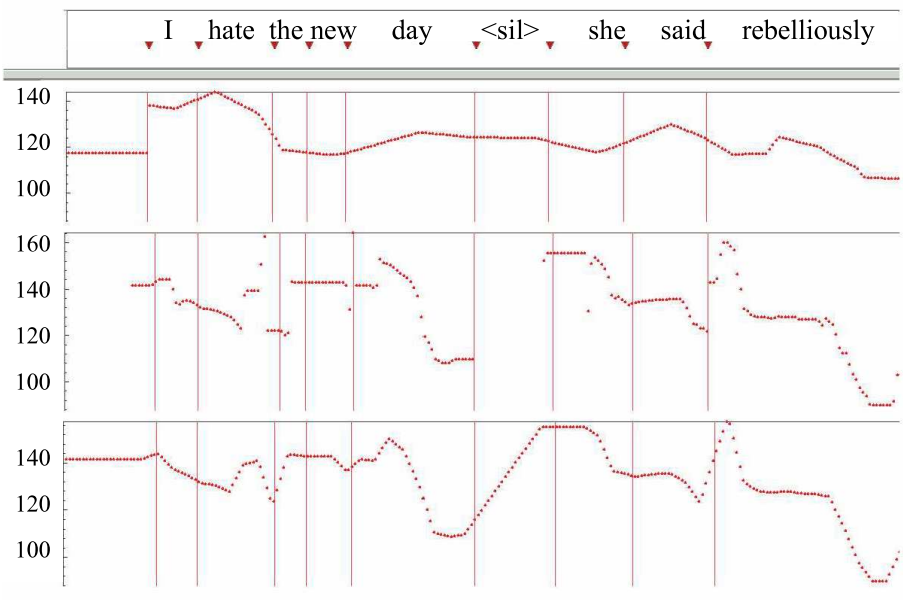
Az egyes természetes prozódia-darabok keresése a korpusz alapú elemkiválasztásos beszédszintetizátorokban használt módon történik, torzítási és összefűzési költségek használatával. A megtalált darabok összefűzése során jelfeldolgozási módszerek segítségével érik el, hogy a végső dallammenet „sima” legyen. Egy-egy szótag F_0-jának másolása három fix pont alapján történik, ahogy az 5. ábrán látható. A felső dallammenet az adatbázisbeli forrásszótaghoz, az alsó pedig a szintetizálandó célhez tartozik, a szaggatott vonalak pedig a szótagok három részre osztását jelzik, középen a magánhangzóval. A másolás szakaszonként lineáris függvények segítségével történik, időbeli széthúzással.

A prozódia másik összetevőjének, az időzítésnek másolása nem ennyire egyszerű a cikk szerzője szerint. Megvalósításukban csak egyes könnyen kezelhető esetekben végzik el az időtartamok módosítását minták alapján. Ha a szintetizálandó szótaghoz találnak teljesen egyező szótagot a prozódia-adatbázisban, akkor a szótag fonémáinak időtartamait normalizálás nélkül felhasználják a célszótagban. Ennek előnye, hogy ha teljes egyezés van a bemeneti mondat és egy adatbázisbeli mondat között, akkor az időzítés beállítása is jó lesz.

2.4.3. Minták keresése beszéd-adatbázisban

Raux és Black modellje [12] hasonló az előző megvalósításhoz (2.4.2. rész). A leglényegesebb különbség, hogy ebben az új megközelítésben a prozódia másolásához használt alapegység a szegmens, ami nem más, mint egy szótagon belüli egység. Ez megfelelő flexibilitást ad a rendszernek, és lehetővé teszi a makro- és mikroprozódia másolását is.

Az elemkiválasztásos rendszerek sikere általában azon alapszik, hogy megkerülik a prozódia modellezését azzal, hogy természetes beszédből kivágott mintákat fűznek össze. Így természetesen hangzó beszédet lehet létrehozni, de a prozódia megfelelő vezérlése nem lehetséges. A vezérlés hiánya még szembetűnőbb olyan esetben, amikor például érzelmes beszédet akarunk szintetizálni. Persze lehetne minden feladatra különböző beszédadatbázist konstruálni, azonban ez nagyon idő- és erőforrásigényes munka, illetve a későbbiekben nehéz módosítani



6. ábra. F_0-generálás Raux és Black művében. Felül: szabály alapú, középen: F_0-szegmens másolás, alul: F_0-szegmens másolás simítással. Forrás: [12].

az adatbázist. Tehát a prozódia megfelelő modellezésével jelentősen csökkenthető az az adatmennyiség, ami különböző típusú szintetizált hangok létrehozásához szükséges.

Ebben a módszerben egy a korábbiakhoz képest kisebb egység, a szótagon belüli szegmens került felhasználásra. Ezzel elvileg lehetővá válik, hogy akár egy-egy szótag dallamát is több adatbázisbeli F_0-elemből állítsák össze, azonban a módszer gyakorlatban egyben kezeli a szótagokat. A rendszer flexibilitása naggyá válik ezzel, ami mindenféle intonáció esemény (pl. hangsúly, szünet) modellezését lehetővé teszi.

Módszerüket a Festival beszédszintetizátor rendszerben [13] implementálták, a rendszer elemkiválasztási és -összefűzési lehetőségeit kihasználva. Az adatbázis felcímkézése és szegmensekre osztása automatikusan történt, az elemek csoportosításával egyben. A bemeneti szöveg szintézise a következő lépésekben történik:

1. szöveg analízis
2. F_0 címkék meghatározása
3. F_0 szegmenscsoport keresése
4. legjobb elem kiválasztása a csoportból
5. összefűzési költségek kiszámítása
6. hang szintetizálása LPC[7] módosítással

[7]Linear Predictive Coding

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. Ahogy a 6. ábrán is látható, a simítás hatása a mondat dallammenetén jól észrevehető, ezt azonban az emberi fül kevésbé hallja, mert a szakadások amúgy is a szótagok határain vannak. A módszer kiértékelése során az derült ki, hogy az esetek többségében az F_0-szegmensekből létrehozott dallam jobb volt a korábbi, szabály alapú megvalósításnál. Főleg hosszabb szövegek szintézise esetén zavaró a szabály alapú modellel létrehozott dallam, mert az meglehetősen monoton hangot hoz létre.

Raux és Black munkája tehát egy teljesen adatvezérelt prozódia generálást eredményezett, ami megfelelő flexibilitásának köszönhetően jól tudja modellezni a beszéd makro- és mikro-szegmentális szerkezetét, és növeli a természetességet. Az adatvezéreltségnek köszönhetően a módszer költséghatékony módja a természetes F_0-modell létrehozásának, hiszen a legtöbb lépés automatikusan történik, emberi erőforrás felhasználása nélkül.

2.4.4. Prozódia többszintű elem-sorozatokkal

Van Santen és társainak megközelítésében [14] az az újdonság, hogy a beszédkorpuszban többféle szempont szerint keresnek a bemeneti szöveghez illő prozódia-mintát. Egyrészt egy olyan sorozatot keresnek az adatbázisban, amelynek fonémái a bemenethez hasonlóak, ezt *„phonemic unit sequence”*-nek hívják. Másrészt több olyan részt próbálnak találni, amik prozódia szintjén várhatóan illeszkedni fognak (pl. hasonló a hangsúlyszerkezetük), ezeket *„prosodic unit sequences”*-nek nevezik. Ez tipikusan frázis, hangsúly, és fonéma egységet jelent. Az utóbbi elemsorozatok kombinációjából, jelfeldolgozás segítségével hozzák létre a bemeneti szöveghez tartozó dallammenetet.

A szerzők szerint módszerük egyik előnye a hagyományos beszédszintézishez képest, hogy így mesterséges helyett természetes F_0-menet hozható létre. A másik előny az, hogy a számítási kapacitás négyzetesről lineárisra csökkent elemkiválasztásos rendszerekhez képest, mert külön kezelik a fonemikus és prozódiai egyezést. A különbég Raux és Black megvalósításához [12] képest egyrészt az, hogy „nyers” F_0 minták összefűzése helyett szuperpozíciós megközelítést alkalmaznak, több szintű prozódia-minta összeadásával, egymásra szuperponálásával. Ezáltal megszűnik az elemkiválasztásos rendszerekben ismert összefűzési hiba, a létrehozott dallammenet folytonos lesz. Másrészt az időtartamok definiálását is a *„prosodic unit sequences”*-ekből származtatva végzik el, aminek az az előnye, hogy az alapfrekvenciát és időzítést együtt másolva természetesebb lesz a prozódia szerkezete.

A cikkben részletesen bemutatásra kerül két módszer is az itt leírtak alapján. Többek között azt is megtudhatjuk, hogy a fonemikus egyenlőség, vagyis a *„phonemic unit sequence”* kiválasztása hogyan történik: a keresés az egyes fonémák kiejtésbeli hasonlóságán alapul. Például a „medal” és a „neighbour” szó fonemikusan egyezőnek számít.

A végső dallammenet meghatározása szuperpozíciós alapon, a következő négy lépésben történik:

1. hasonló fonémasorozat keresése
2. hasonló hangsúlysorozat keresése

3. hasonló frázissorozat keresése

4. az előbbi három dallam-összetevő összeadása, egymásra szuperponálása

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.

2.4.5. Természetes F_0 elemek statisztikai manipulációja

Saito is természetes F_0 elemekből építi fel a dallammenetet módszerében [15]. 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 adatázis elkészítése a természetes beszédből vett mondatok automatikus felbontásával kezdődik, majd az alapfrekvencia-értékek illetve a szünetek helye is meghatározásra kerül automatikus módszerrel, és kézi korrekcióval. Az alapfrekvencia-menet tárolása magánhangzónként egy F_0 értékkel történik, amit a hang közepén mintavételeznek, hogy elkerülhető legyen a szomszédos mássalhangzók befolyása.

A szintézis három fő lépésből áll:

1. F_0 vázlat meghatározása

2. F_0 elemek keresése

3. F_0 elemek összefűzése

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 másolása csak akkor történik meg, 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 segítségével elvégeztek egy kísérletet, hogy összehasonlítsák az így generált beszéd természetességét a korábbi megvalósításokkal. Az eredmények azt mutatják, hogy a generált dallammenet nagyon hasonlít ahhoz, mintha egy ember beszélne, és az esetek 78%-ában ezt preferálták a korábbi módszerekhez képest a kísérletben résztvevők.

1. táblázat. Prozódia-másolási megoldások összehasonlítása.

Módszer szerzője	Dong, Lua	Meron	Raux, Black	van Santen et al.	Saito
F_0 egység	mondat, frázis, szótag	szótag	szegmens	frázis, hangsúly, fonéma	szó

2.4.6. Prozódia-másolási megoldások összefoglalása

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. táblázatban az itt bemutatott módszerek F_0-egységei láthatóak. Meron [10] munkájában szótagokat, esetleg több szótagot együtt használt fel, ami viszonylag jó dallamösszeállítási lehetőséget biztosít. Raux és Black [12] módszerében egy még kisebb egység, a szótagon belüli szegmens került felhasználásra. Saito [15] pedig szavak szintjén végezte el a prozódia másolását.

Arra is lehetőség van, hogy a hosszú és rövid egységeket kombináljuk. Dong és Lua [9] módszerükben a mondat-, frázis- és szószintű F_0 minták egymásra szuperponálásával hozta létre a dallammenetet. Van Santen és társai [14] munkájukban szintén három féle egységekre bontották a beszédkorpuszbeli mondataikat, és ezek kombinációjával definiálták az alalpfrekvencia-menetet.

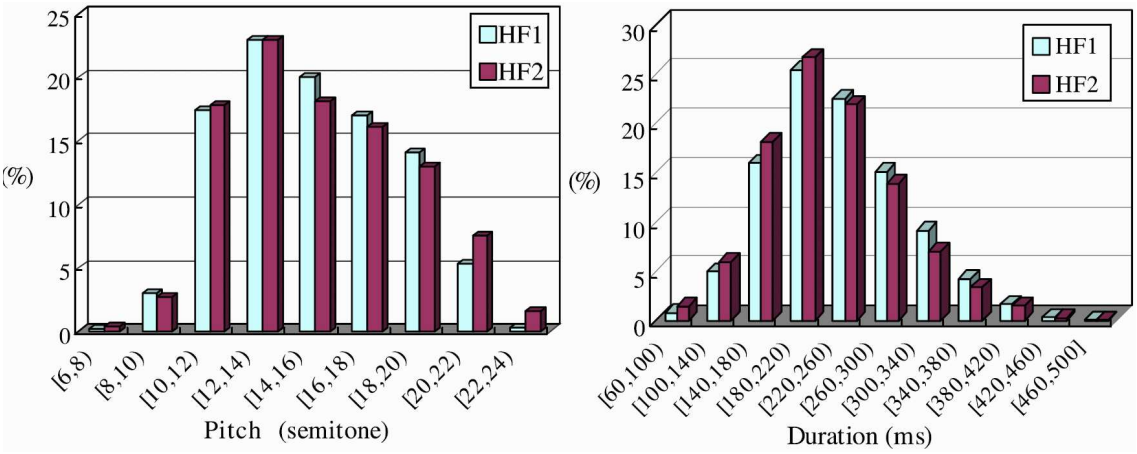
2.5. Prozódiai változatosság elemzése

A korpusz alapú prozódia-generálási módszerek áttekintése után most bemutatásra kerül egy olyan munka, amely a prozódiai változatosság elemzésével és megvalósításával foglalkozik beszédszintitizátorokban.

2.5.1. A prozódia invariáns és változó részei

Min Chu és társai a beszéd variáltságával foglalkoznak munkájukban [16]. 1000 mondatot kétszer felvettek, 6 hónap különbséggel, és azt vizsgálják, hogy az egyező mondatoknak mennyire hasonló illetve eltérő a prozódiája.

A legtöbb beszédszintetizátor rendszer determinisztikusan állítja elő a prozódiát. Ez sokszor ismétlődő, monoton dallamminták túlzott előfordulásához vezet, ami zavaró a szintetizált beszédben. A prozódiát fel lehet bontani invariáns és változó részekre. Az invariáns



7. ábra. Átlagos alapfrekvencia és időtartam Min Chu és társai két adatbázisában. HF1 a korábban, HF2 a később felvett adatbázist jelenti. Az átlagos alapfrekvencia logaritmikus skálázású relatív értékként van kifejezve. Forrás: [16].

rész egy átlagos érték a részletek nélkül, míg a változékony rész a prozódia szabad vezérlését, a részletességet jelenti.

A prozódia minták ismétlődése azért fordulhat elő a TTS rendszerekben, mert például egy elemkiválasztásos szintetizátor mindig a legjobb prozódiát próbálja egy-egy mondathoz rendelni. Így az emberi beszéd változatossága lecserélődik a legjobb, leggyakoribb mintára. Ez viszont az emberi fül számára, ami a változékonysághoz szokott, könnyen felismerhető. Beszédünk stílusát sokszor szándékosan is variáljuk, ha különböző dolgokat akarunk kifejezni. Sokszor pont 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.

A prozódia invariáns jellemzői közé például a frázishatár előtti időbeli nyújtás, a hangsúlyok és F_0 emelés/csökkentés összefüggése tartozik. A prozódia változékonysága két típusú lehet. Az első esetben, amit JND[8]-nek is hívnak, észrevehetetlenek a prozódiabeli változások, míg a második esetben észrevehetjük például a dallambeli változást, de ez nem módosítja az átviendő gondolat értelmét. Tehát egy-egy szövegnek sokféle prozódiai megfelelője lehet.

Min Chu és társai munkájukban tehát $2 \cdot 1000$ mondatból álló beszéddallam-adatbázist használtak, ahogy ezt korábban említettük. Az adatbázisok felvétele után a címkézés automatikusan történt, így meghatározásra kerültek például a szótag- és szünethatárok. A két adatbázis párhuzamos vizsgálatával az vehető 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 jelentősebb különbség van, ami a 7. ábrán is látható. Azt vehetjük észre a hisztogramokon, hogy annak ellenére, hogy a két adatbázis ugyanazon bemondó azonos mondatait tartalmazza, eltérések mutatkoznak. Természetesen ez a különbség eltérő lehet beszélőtől függően.

[8]Just Noticeable Differences

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

Újabb kísérleteik során kifejlesztésre került egy módszer, mely egy elemkiválasztásos TTS, természetellenes prozódia észleléssel [17]. Ennek segítségével elkerülhető a korpusz alapú rendszerekben sokszor előforduló működési hiba (ami a különböző egységek összefűzésekor jelentkezik), ugyanakkor valamilyen mértékben megvalósítható a prozódia változatossága.

3. Prozódiai változatosság növelése a gyakorlatban

A témához tartozó szakirodalom ismertetése után most bemutatjuk, hogyan próbáltuk meg korpusz alapú prozódia-generálás segítségével növelni a szintetizált beszéd változatosságát. Célunk hasonló volt a 2.5. alfejezetben bemutatott módszerhez, vagyis hogy a TTS a beszéd szintézise során ne mindig a legjobb prozódia-mintát generálja, hanem egy elfogadható tartományon belül változatosabb dallamot tudjon létrehozni.

A 3.1. alfejezetban bemutatjuk a munkánk során használt beszéddallam-adatbázis szerkezetét. Ezután ismertetésre kerül, hogyan lehet a Profivox szövegfelolvasóban beállítani a szintetizált mondat dallammenetét (3.2. alfejezet). A 3.3. és a 3.4. alfejezetben különböző dallammásolási lehetőséget mutatunk be. Végül az utolsó részben (3.5. alfejezet) megvizsgáljuk, hogyan lehet módszerünket egy beszédszintetizátor rendszerben alkalmazni.

3.1. Felhasznált beszéddallam-adatbázis

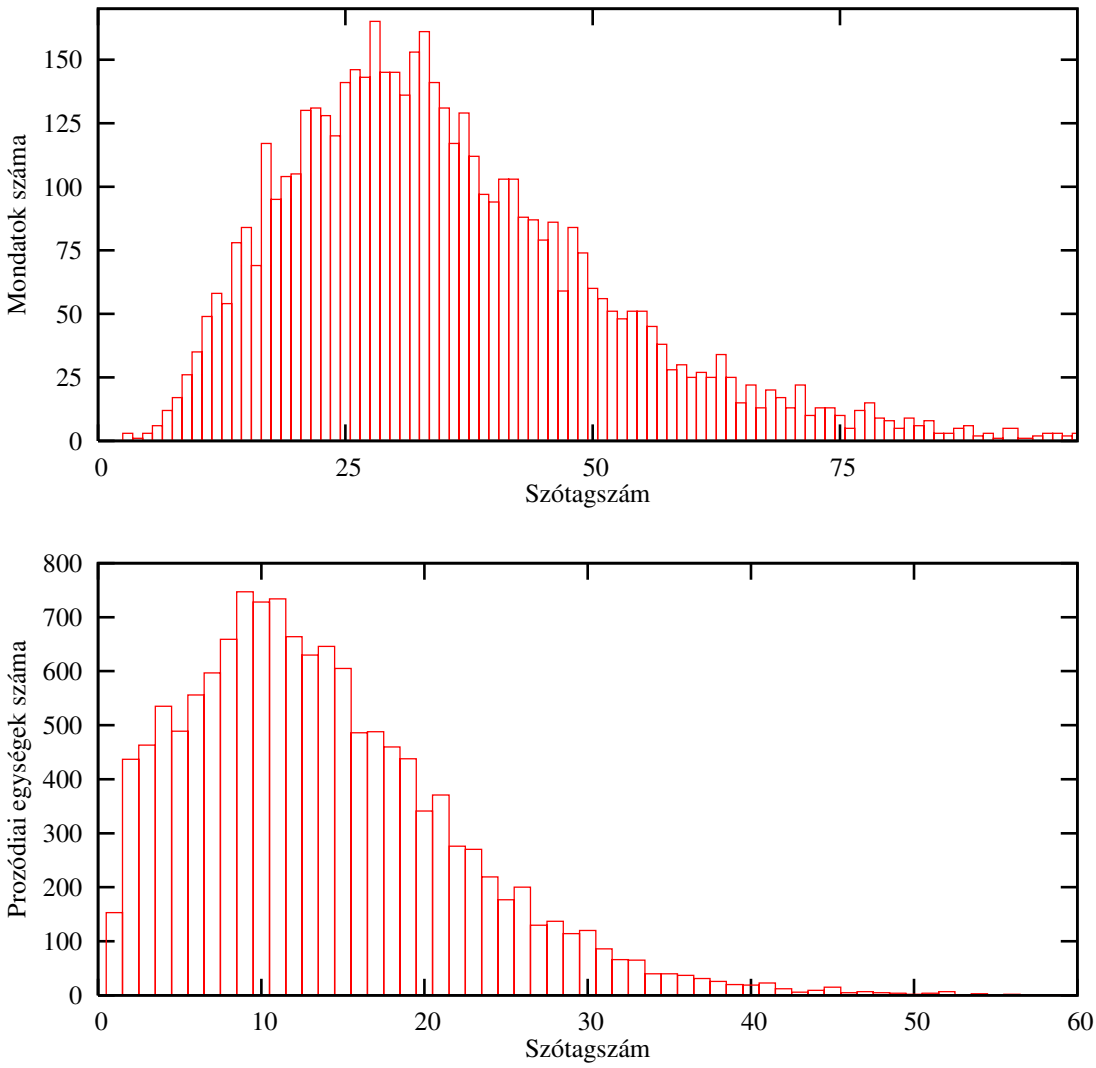
A dolgozatban használt beszéddallam-adatbázist a BME-TMIT bocsátotta rendelkezésünkre, mivel ez egy korábbi kutatáshoz készült el. A jelenlegi munkánkhoz szükséges volt ennek kiegészítése.

3.1.1. Az eredeti adatbázis

A beszéddallam-adatbázis a következőkben Fék és társai [3] munkája alapján kerül bemutatásra. A korpusz egy professzionális bemondótól származó, időjárás-előrejelzés témájú, magyar nyelvű kijelentő mondatokból áll. Egy-egy mondathoz a következő részek tartoznak:

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

A korpusz létrehozásakor tehát a kiindulási egységek a mondatokhoz tartozó, természetes beszédet tartalmazó hullámforma fájlok voltak, az adatbázis elkészítésekor ezek címkézése történt meg automatikusan. Minden mondathoz tartozik egy szöveges átírás, ami a felolvasott mondatot tartalmazza. Ebből lehetett létrehozni a magyar nyelv megfelelő hasonulási szabályainak segítségével a fonetikus átírást, amiben a kiejtett fonémák jelennek meg. Az adatbázisbeli mondatok hang-, szó- és szünethatárainak jelölése automatikusan történt meg egy beszédfelismerő segítségével. A zöngeperiódus-határok, vagyis a mondat dallammenetének meghatározása a Praat fonetikai beszéd-analizátor programban implementált alapfrekvencia-detektálás alapján történt [18].



8. ábra. Mondatok és frázisok szótagszámának gyakorisága az adatbázisunkban.

3.1.2. Az adatbázis kiegészítése

A beszédkorpuszbeli mondatokhoz tartozó hangsúlycímkék nem álltak rendelkezésre. Ezeknek meghatározása a Profivox beszédszintetizátor segítségével történt [19] a szöveges átírás alapján, szintén automatikusan. A rendszer öt féle hangsúlytípust különböztet meg, ezek:

- fókusz, mondatközpont
- nyomaték, értelmi hangsúly
- normál szóhangsúly
- semleges szó
- negatív hangsúly

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

A beszéddallam-adatbázis 5200 kijelentő mondatot tartalmaz. A mondatok szótagszámának hisztogramja a 8. ábra felső részén látható. Észrevehetjük, hogy nagyon sok hosszú mondat van, ami több mint 25 szótagból áll. Azért van ez így, mert a mondatok speciális témájúak, időjárás-előrejelzésekből lettek származtatva. Ez nem tipikus a magyar nyelvre, ezért a mondatokat prozódiai egységekre bontottuk, és azok alapján is vizsgáltuk őket, mely a 8. ábra alsó felén található. Mivel a prozódiai egységek hosszú mondatok (akár 75 szótag), esetén is viszonylag rövidek (tipikusan 2-20 szótag), jobban reprezentálják az általános témájú beszédet.

3.2. Alapfrekvencia beállítása a Profivoxban

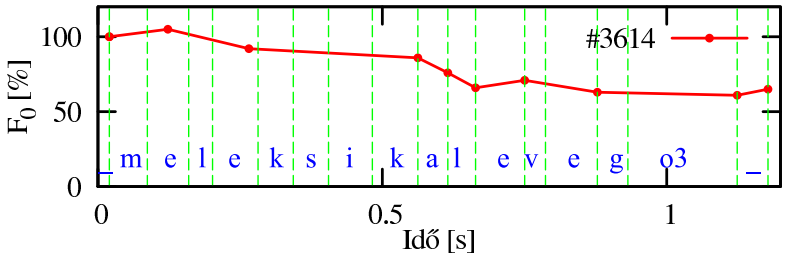
A Profivox beszédszintetizátor a bemeneti szövegből először egy köztes fájlt, úgynevezett intonációs mátrixot hoz létre, amelyből a kimeneti beszéd szintetizálható. A prozódia módosítására ezen köztes fájl segítségével van lehetőségünk. A 2. táblázat és a 9. ábra egy intonációs mátrixra, és az abban definiált dallammenetre mutat példát. A mátrix minden sorában egy-egy hang paraméterei találhatóak: többek között a hanghoz tartozó fonéma, az alapfrekvencia magassága egy alapértékhez képest százalékban kifejezve (ez az alapérték csak a beszéd szintézisekor kerül beállításra), illetve az, hogy az adott hangban hol kell elérni az előbbi magasságot (a hang hosszához képest, százelékban kifejezve). Az intonációs mátrixban megadható a hang tervezett időtartama és intenzitása is.

A dallamra tehát egy olyan töröttvonalat tudunk definiálni, 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.

2. táblázat. Profivox intonációs mátrix. Minden sorban egy-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>

...



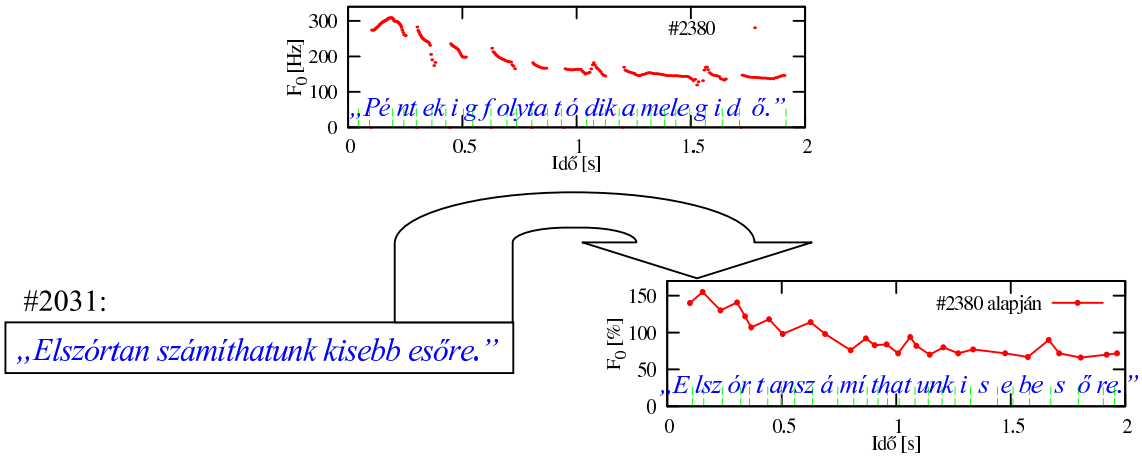
9. ábra. Profivox intonációs mátrix által definiált dallammenet: hangonként egy töréspont adható meg, amikkel az F_0-görbét tudjuk meghatározni.

3.3. Dallammenet létrehozása minták alapján

A beszédszintézis során szöveg alapján hozunk létre beszédet különböző lépésekben. Ezek egyike a prozódia tervezése, amire sokféle módszer ismert a szakirodalomban (2.3. alfejezet). A cél tehát az, hogy a bemeneti szöveghez minél természetesebb dallammenetet tudjunk rendelni, amit jelen munkánkban korpusz alapon, azaz a természetes mondatokat utánozva próbáltunk megvalósítani.

3.3.1. Dallammásolás ötlete

Ahhoz, hogy a beszédkorpuszunk alapján definiálni tudjuk egy-egy szintetizálandó mondat dallamát, először annak eldöntésére volt szükség, hogyan lehet a természetes mondatok adatbázisából olyan részeket kiválasztani, amik hozzárendelhetőek a bemeneti szöveghez. Korábbi munkánk során a bemeneti szöveg és az adatbázisbeli minták összerendelését a szótagszámaik alapján oldottuk meg [1]. Azért választottuk a szótagszámot, mert a mondatbeli hangsúlyok általában a szótagokhoz kötődnek, és azt feltételeztük, hogy hasonló szótagszerkezet választásával várhatóan jó helyre kerülnek a hangsúlyok a szintetizálandó mondatban is. Ehhez fel kellett mérni az összes adatbázisbeli mondat szótagszerkezetét. Szótagszerkezeten azt értjük, hogy a mondatbeli szavak hány szótagból állnak. Például az *„Elszórtan számíthatunk kisebb esőre.”* mondat első szava három szótagból, a második négyből, a harmadik kettőből, a negyedik ismét három szótagból áll, így szótagszerkezete: 3+4+2+3. Ezekben a kísérletekben csak egy prozódiai egységből álló mondatokat vizsgáltunk, így nem volt szükség részletesebb



10. ábra. Dallammásolás teljes mondat alapján: a #2031-es mondat szövegéhez szótagonként hozzárendeltük a #2380-as mondat dallamát, egy intonációs mátrixot létrehozva.

felbontásra.

Először kiválasztottunk két szótagszerkezet alapján hasonló mondatot az adatbázisból. Ekkor még csak a jelenlegi adatbázis 200 mondatos részhalmazával dolgoztunk, és ebben nem találtunk teljesen egyező szótagszerkezetű mondatokat, ezért olyanokat választottunk, amik közel álltak egymáshoz (#2031: *„Elszórtan számíthatunk kisebb esőre.”*, 3+4+2+3, #2380: *„Péntekig folytatódik a meleg idő.”*, 3+4+1+2+2). Az egyik mondat szövegéhez hozzárendeltük a másik mondat dallammenetét, ahogy a 10. ábrán is látható. A #2380-as mondat dallammenetét szótagonként hozzáillesztettük a #2031-es mondat szótagjaihoz (ezt azért tehettük meg, mert a két mondat szótagszáma egyező). A szótagonkénti F_0-hozzárendelés a Profivox intonációs mátrix segítségével történt meg. Az időzítés- és intenzitásértékeket a Profivox rendszer hozta létre szabály alapon, ezeket nem módosítottuk.

A tavalyi munka tehát csak az első lépés volt a prozódiai változatosság megvalósításának irányába, főleg arra szolgált, hogy ellenőrizzük, koncepciónk megvalósítható-e. Az alapfrekvencia-menetek másolása ekkor még kézi és félautomatikus módszerekkel történt. Mivel az elvégzett kísérletek eredményesek voltak, folytattuk munkánkat a dallammásolás részleteinek pontosabb kidolgozásával.

3.3.2. Dallammásolás nagyobb adatbázisban

Mivel a 200 mondatos részadatbázisban nem találtunk teljesen egyező szótagszerkezetű mondatokat, az egész, 5200 mondatból álló beszédkorpuszt kezdtük vizsgálni [20]. Olyan mondatcsoportokokat kerestünk benne, melyeknek szótagszerkezete teljes mértékben megegyezik. Ilyen teljesen egyező szerkezetű halmazok csak az adatbázis rövidebb mondatai között szerepeltek. Kiválasztottuk a 3. táblázatban látható csoportokat, és ezen mondatokat többféle változatban szintetizáltuk. Kidolgozásra került egy módszer az alapfrekvencia-menet automatikus másolására, aminek segítségével lehetővé vált az egy prozódiai egységből álló mondatok dallamának másolása természetes mondat alapján.

Ezen munkánk során 42 szintetizált mondatot hoztunk létre, melyek többek között Profivox

3. táblázat. Teljesen egyező szótagszerkezetű mondatok az adatbázisunkban.

mondat azonosító	szótagszám	szószám	szavak szótagszáma
#3053	7	3	3+1+3
#3614	7	3	3+1+3
#3855	7	3	3+1+3
#3056	10	3	4+3+3
#3373	10	3	4+3+3
#1773	11	4	3+2+2+4
#2565	11	4	3+2+2+4
#3517	17	6	3+4+1+4+2
#3953	17	6	3+4+1+4+2
#2551	17	6	1+5+3+5+2+1
#3966	17	6	1+5+3+5+2+1

szabály alapú dallammal, illetve különböző dallammásolási módszerrel készültek. Az újabb meghallgatásos teszt eredményéből az derült ki, hogy a alapfrekvencia-másolás javította a szintetizált mondatok minőségét.

3.4. Dallammásolási lehetőségek a változatosabb prozódia érdekében

A 2.4. alfejezetben bemutattunk néhány lehetőséget a dallam létrehozására beszédminták alapján. Miután kidolgoztunk egy saját módszert a természetes mondatok dallamának másolására (3.3. alfejezet), jelenlegi munkánkban az került fókuszba, hogyan lehetne ennek segítségével a szintetizált beszéd prozódiai változatossgát növelni. Az lenne a cél, hogy a TTS egy-egy bemeneti mondatához ne mindig ugyanolyan prozódiájú mondatot szintetizáljunk. Úgy valósíthatjuk ezt meg, ha a bemeneti szöveghez többféle dallammenetet tudunk generálni, és ezek közül a rendszer szintéziskor egyet kiválaszt. Ekkor ugyanis csökken a monotonitás, hiszen nem determinisztikusan ugyanaz a dallammenet rendelődik a mondatokhoz.

A prozódiai változatosság eléréséhez az szükséges, hogy egy-egy mondathoz legalább 3-4 lehetséges dallammenetet tudjunk definiálni, a természetes mondataink adatbázisának segítségével. Az, hogy egy mondathoz hány teljes dallammintát tudunk előállítani, függ attól, hogy mekkora F_0 egységekkel dolgozunk, és mekkora a beszéddallam-adatbázis mérete. A beszédkorpuszunk 5200 mondattal már elég nagynak számít, így az F_0 egységek méretén változtattunk. A 3.3. alfejezetben bemutatottak szerint korábbi munkánkban teljesen egyező szótagszerkezetet vizsgáltunk, ami csak rövid mondatcsoportok megtalálására volt jó (hosszabb mondatokban ugyanis a szótagszerkezet nagyon változatos lehet, így kicsi az esélye, hogy találunk két egyformát).

Ahhoz, hogy a hosszabb, több frázisból álló mondatokhoz is találhassunk prozódia-mintát, a mondatok felbontására volt szükség. Mivel a prozódiai egységek rövidebbek, egy-egy ilyenhez nagyobb valószínűséggel lehet találni egyező szótagszerkezetűt. Ha például egy szintetizálandó mondat három frázisból áll (*„Hétfőn a csípős, fagyos reggelt követően több-*

órás napsütésre számíthatunk, amelyet csak északkeleten zavar átmeneti felhősödés."), egyben kezelve egyáltalán nem biztos, hogy találunk hozzá szerkezetileg hasonlót. A 8. ábrán látható hisztogram azt mutatja, hogy prozódiai egységből sok van. Tegyük fel, hogy az első (*„Hétfőn a csípős,"*) és második szintetizálandó frázishoz (*„fagyos reggelt követően többórás napsütésre számíthatunk,"*) két-két egyező szerkezetűt találunk, a harmadikhoz (*„amelyet csak északkeleten zavar átmeneti felhősödés."*) pedig csak egyet. Ekkor a mondatot a frázisok dallamainak kombinációjából $2 \cdot 2 \cdot 1 = 4$-féle F_0-menettel tudjuk szintetizálni.

3.4.1. Prozódiai egységek vizsgálata

Az adatbázisbeli mondatokat tehát felbontottuk prozódiai egységekre a szöveges áítás alapján automatikus módszerrel, ahogy ezt már a 3.1.2. részben is bemutattuk. Az egyes prozódiai egységekhez a következő információkat tároltuk el (zárójelben egy-egy példával):

- mondat sorszáma, amiből származik (#5241)
- pozíció a mondatban (első)
- összes szótagszám (10)
- szótagszerkezet (3+1+2+4)
- szótagonként egy-egy alapfrekvencia-érték (239 Hz, 218 Hz, 194 Hz . . .)
- átlagos alapfrekvencia (183 Hz)
- hangsúlyszerkezet (N E N N)

Az adatbázisbeli mondatokat sorszámukkal jellemeztük. A frázis pozíciója a mondatban lehet első, középső vagy utolsó. A szótagszerkezet a korábbiakban definiáltak alapján került meghatározásra. Szótagonként egy-egy alapfrekvencia-értéket tároltunk el, ami a dallam másolásához volt szükséges. A hangsúlyszerkezetet a Profivox rendszer leírása szerint jelöltük (F = fókusz, E = nyomaték, W = normál, N = semleges, - = negatív).

Összesen 13 415 prozódiai egységre bontottuk így az adatbázist. Átlagosan egy mondat 2,57 frázisból, egy frázis pedig 13,78 szótagból áll az egész korpuszt figyelembe véve.

A hangsúlyszerkezetet, a pozíciót és az átlagos F_0 értéket azért tároltuk el, hogy a prozódia létrehozásakor a frázisok kiválasztásában ezeket is figyelembe lehessen venni. A prozódiaminták kiválasztásakor és egymás után fűzésekor tehát különböző kényszerek segítségével biztosíthatjuk a természeteshez hasonló dallammenetet:

- hasonló F_0 érték a frázisok határán
- hasonló átlagos F_0 az egymás után következő frázisokban (10-20%-os különbség lehet)
- hangsúlyok figyelembe vétele a szótagszerkezet mellett
- frázisok sorrendje

Az első kényszer azért lehet fontos, mert a természetes beszédben sincsenek hirtelen nagy F_0-ugrások, így a szintetizálás során sem célszerű ilyet létrehozni. Az emberi beszéd kijelentő mondatai ereszkedő dallamúak, így például az átlagos F_0 alapján csökkenő prozódiai egységeket kiválasztva várhatóan természetesebb lesz a dallammenet. A hangsúlyok figyelembevételével a dallammásolás hatékonysága és természetessége tovább növelhető. Az utolsó kényszer lényege, hogy például az első szintetizálandó frázishoz olyan prozódia-mintát keressünk, ami az adatbázisban is mondat elején állt, a középső szintetizálandókhoz középső adatbázisbelit, az utolsó szintetizálandó frázishoz pedig utolsó korpuszbeli minta legyen.

3.4.2. Dallammásolás prozódiai egységek alapján

A beszédkorpusz prozódiai egységekre bontása után előttünk állt a lehetőség, hogy a dallamok másolását kiterjesszük hosszabb mondatokra is. Módszerünk működése a 11. ábrán látható egy példán keresztül.

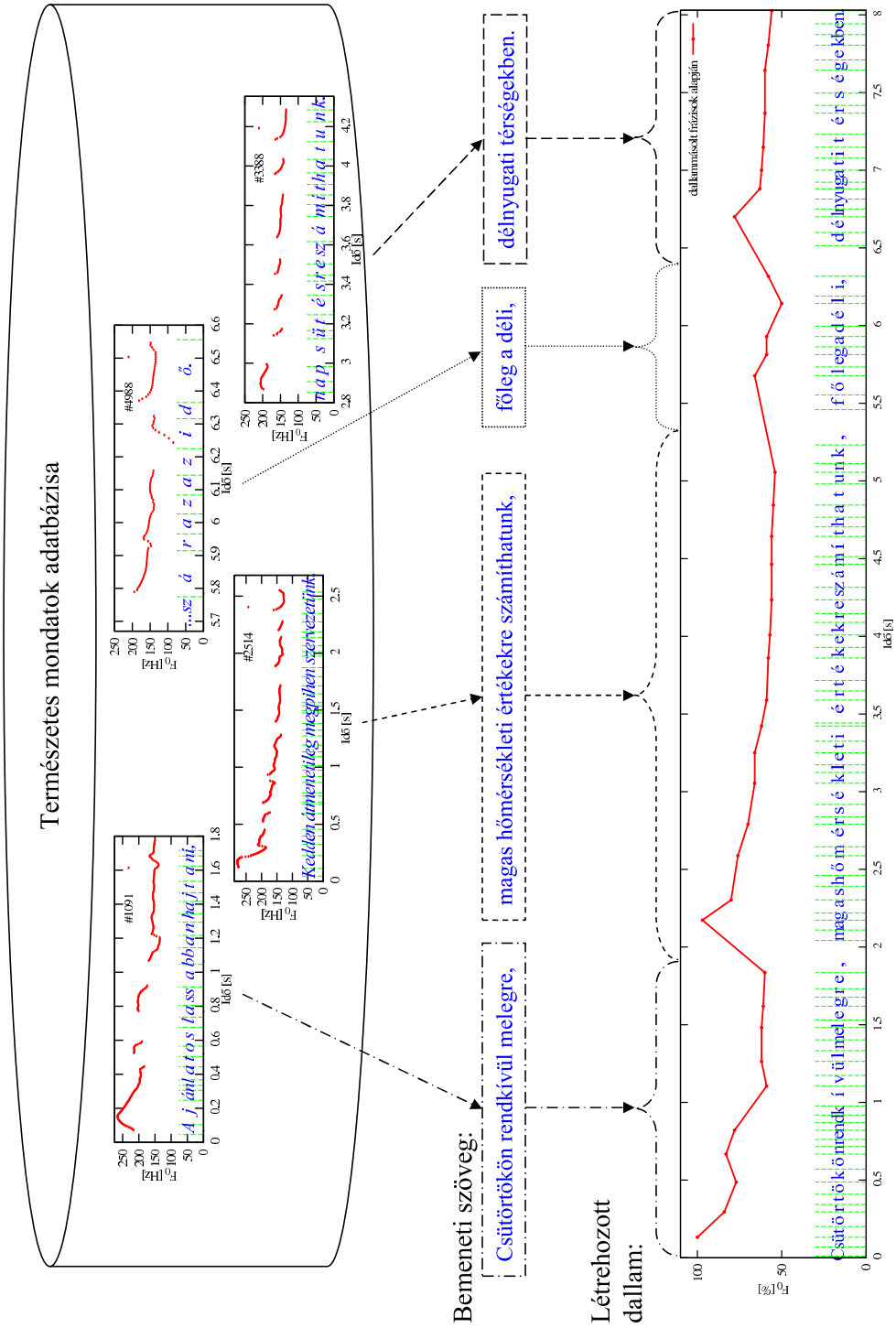
A bemeneti szöveget (*„Csütörtökön rendkívül melegre, magas hőmérsékleti értékekre számíthattunk, főleg a déli, délnyugati térségekben.”*) először prozódiai egységekre bontjuk a vesszők mentén, ebből négy darab van a mondatban. Minden egyes frázisnak meghatározzuk a szótagszerkezetét (4+3+3, 2+5+4+4, 2+1+2, 4+4), és ez alapján keresünk illeszkedő egységeket a természetes mondatok adatbázisából. Egy-egy részhez általában több illeszkedőt is lehet találni az adatbázisban, így el kell dönteni, hogy melyik frázisokat használjuk az adatbázisból. Fontos szempont lehet, hogy az eredeti dallam mondata ereszkedő jellegű legyen, illetve hogy a prozódiai minták összeillesztésénél ne legyen nagy eltérés, mert az természetellenes dallamugrást okozna a szintetizált mondatban. Az egyes kiválasztott prozódiai egységek (1. frázishoz a #1091-es, 2.-hoz a #2514-es, 3.-hoz a #4988-as, 4.-hez a #3388-as természetes mondatok részei) szótagonkénti alapfrekvenciáját felhasználva a bemeneti szöveghez rendeljük azokat, így létrehozva egy intonációs mátrixot, mely alapján a Profivox létre tudja hozni a beszédhangot.

A bemeneti szöveghez a módszer segítségével teljesen automatikusan történik meg a teljes mondatra vonatkozó dallammenet meghatározása. Az előbbiekben említett kényszerek közül először csak azt használtuk, hogy az egymás után következő frázisok átlagos F_0-ja ne legyen nagyon eltérő (azaz 10-20%-on belül), mert enélkül meglehetősen nagy ugrások lettek volna a szintetizálandó mondat dallammenetében.

3.5. Prozódiai változatosság megvalósítása

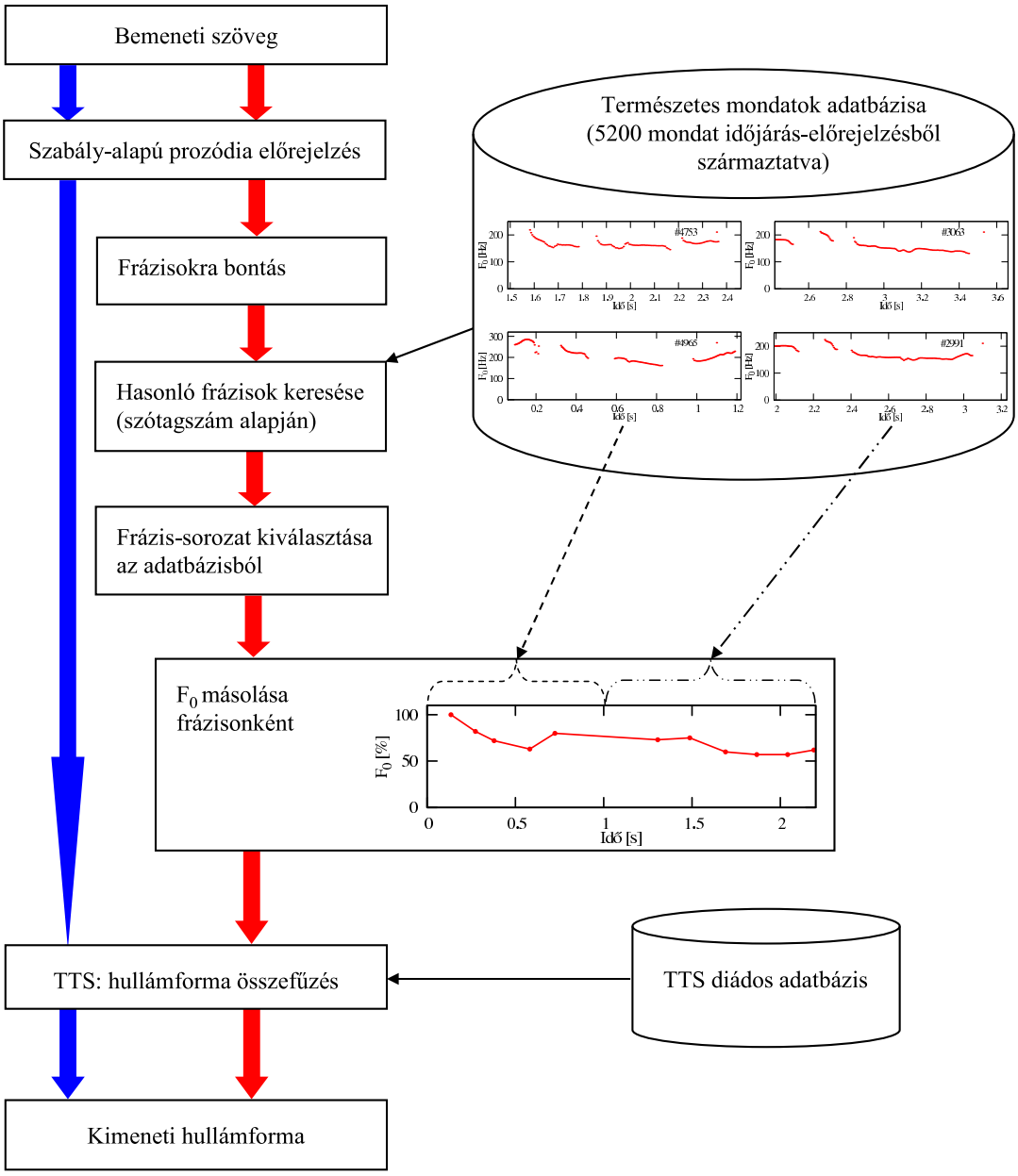
A korábbi, tisztán szabály alapú beszédszintetizátorok hátránya a determinisztikusság, azaz egy konkrét bemeneti szöveghez mindig azonos beszédet hoznak létre, ami így hosszabb szöveg szintetizálása esetén monotonitást eredményez. Jelen dolgozatunkban kidolgozásra került egy olyan módszer, mellyel ez a hátrány kiküszöbölhető. Az új módszerben egy konkrét bemenethez szintézis során több lehetőségből nemdeterminisztikus módon kerül meghatározásra a dallammenet.

A két módszer összehasonlítása a 12. ábrán látható. A régi, szabály alapú megvalósítást a kék nyilak, míg az új, dallammásoláson alapuló, változatosságot megvalósító módszert a



11. ábra. Dallammásolás frázisok alapján.

piros nyilak mutatják. A bemeneti szöveg alapján a hangidőtartamok és az intenzitás meghatározása mindkét esetben szabály alapon történik. Innentől külön válik a két módszer: a kék nyilak a szabály alapú dallam létrehozását jelentik, míg a piros nyilak a dallammásolás lépésein haladnak. Az új módszerben az előző alfejezetben bemutatott módon történik a dallam meghatározása: először frázisokra bontjuk a bemenetet, majd mindegyikhez keresünk prozódiamintát az adatbázisból. A lehetséges minta-sorozatok közül egyet véletlenszerűen kiválaszt a rendszer (bizonyos korlátok figyelembe vételével), és megtörténik az F_0-szakaszok másolása frázisonként. Végül mindkét módszer esetében egy diádos adatbázis segítségével történik meg a hullámforma összefűzés, vagyis a szintetizált beszéd létrehozása.



12. ábra. Szabály alapú (kék nyilak) és változatosságot megvalósító módszer (piros nyilak) működésének összehasonlítása.

4. Vizsgálatok, teszt és eredmények

Ebben a fejezetben bemutatjuk, hogyan minősíttettük a módszerünkkel létrehozott szintetizált mondatok minőségét tesztelők segítségével. Először a kísérletben meghallgatott mondatok tulajdonságairól írunk, majd a teszt körülményei után az eredmények kiértékelése következik.

4.1. Vizsgált mondatok

Kiválasztottunk a természetes mondatok adatbázisából 10 mondatot, és ezeket szöveges átírásuk alapján újraszintetizáltuk a korábban bemutatott módszer (3.4.2. rész) segítségével, különféle dallammenetekkel. A változatok között szerepelt mondatonként 1-1 szabály alapú F_0-görbével rendelkező változat, illetve 2-3 olyan variáns is, amelynek dallama prozódiai egység alapján történő másolással lett létrehozva.

A tesztben vizsgált mondatok szövege:

#0205 *„Hétfőn a csípős, fagyos reggelt követően többórás napsütésre számíthatunk, amelyet csak északkeleten zavar átmeneti felhősödés."*

#0211 *„Bár szeles, hideg időnk lesz, rétegesen, jó melegen öltözve sétáljunk nagyokat a szabad levegőn."*

#0413 *„Csütörtökön rendkívül melegre, magas hőmérsékleti értékekre számíthattunk, főleg a déli, délnyugati térségekben."*

#0515 *„Alig-alig mozdul környékünkről a ciklon, azaz a kettős (egyidejű hideg és meleg) front hatásával még mindig számolnunk kell."*

#1438 *„Az égei- tenger térségében, leszálló légmozgások érvényesülnek, kevés a felhő, csapadékról nem érkezik jelentés. "*

#2411 *„Hideg idő lesz, kedd reggel északon."*

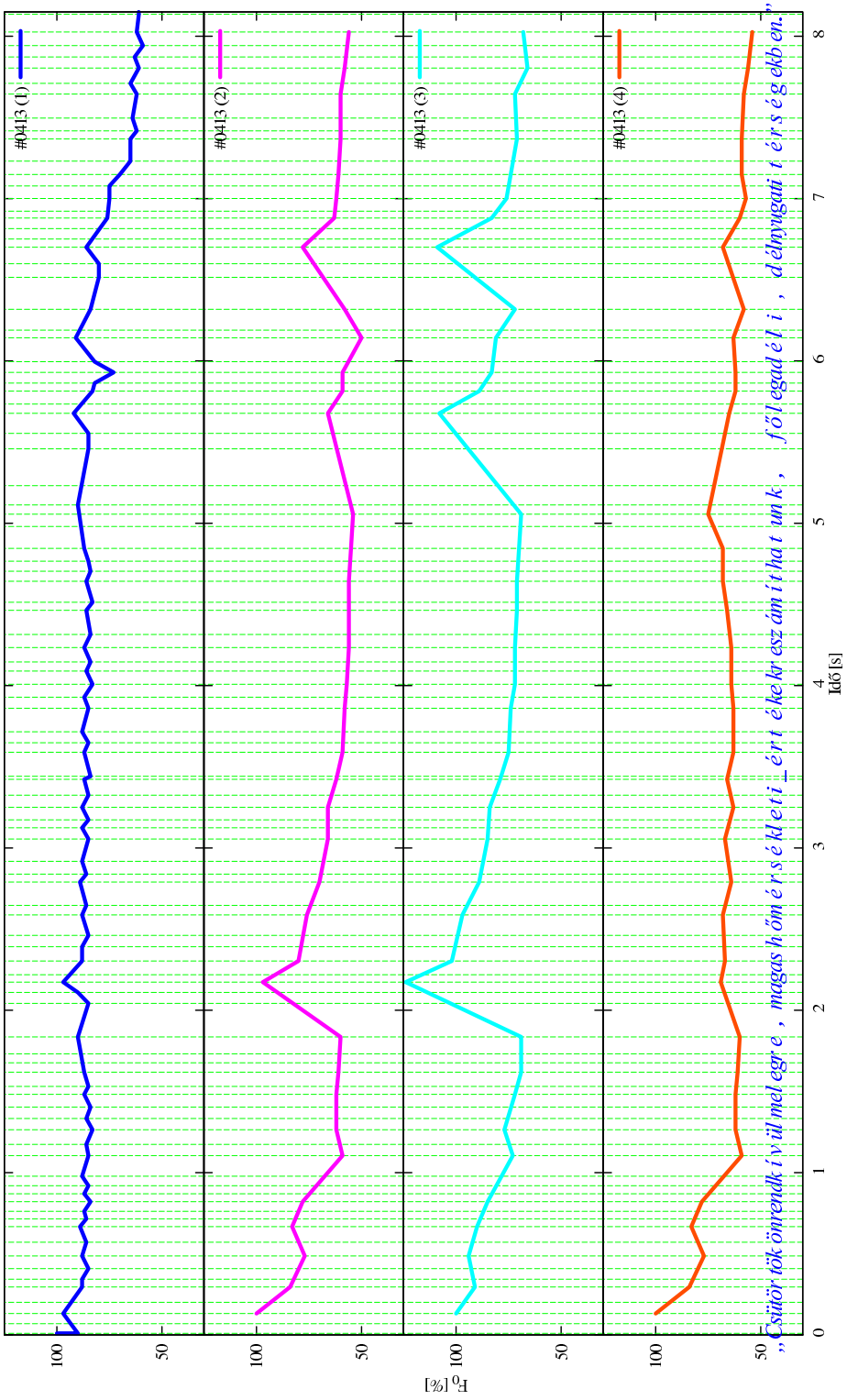
#2546 *„Erős az ibolyántúli sugárzás, ezért a déli, közvetlen napsütéstől óvnunk kell magunkat."*

#3733 *„Szombaton országszerte várható havazás, délen havas eső, eső is lehet, majd ezt követően szerdáig többnyire napos, száraz, de hideg, téli idő várható."*

#4120 *„Az északi szelet sokfelé élénk, időnként erős, zivatar környeztében átmenetileg viharos lökések kísérhetik."*

#4965 *„Útját erős szél, és sokfelé jelentős mennyiségű csapadék kíséri."*

Egy példaként, a #0413-as mondat 4 változatát a 13. ábrán láthatjuk. A legfelső, 1-es számú a szabály alapú dallammenettel létrehozott, míg a többi, 2, 3, 4-es számú különböző dallamminta sorozatokról másolt alapfrekvencia-menettel rendelkezik. A hangok időtartamai mind a négy változatban egyformák, mivel az szabály alapon készült az összes esetben.



13. ábra. A #0413-as mondat négy különböző dallammenettel szintetizált változata.

4.2. Tesztkörnyezet

A létrehozott mondatok tesztelését a BME-TMIT-en kifejlesztett webes tesztelő rendszerben végeztük. A létrehozott mondatokból mondatpárokat hoztunk létre, melyek egy-egy mondat két változatát tartalmazták. Összesen 37 ilyen mondatpár készült el. A tesztet elvégzők feladata az volt, hogy eldöntsék, a mondatpár első vagy második tagját tartják természetesebbnek, vagy nem tudnak különbséget tenni a két változat között, egyforma minőségűnek hallják azokat. Egy-egy mondatot többször is meghallgathattak, hogy döntésüket könnyebben meg tudják hozni. A mondatok lejátszása véletlen sorrendben történt. A tesztelőknek a `http://speechlab.tmit.bme.hu/csapo/` oldalt meglátogatva egy rövid ismertetőt kellett elolvasniuk a teszt menetéről, majd néhány információt kértünk be róluk (becenév, életkor, nem). Ezután megkezdődhetett a mondatpárok meghallgatása. A szintetizált hangok meghallgatása után a tesztelők megjegyzést is írhattak észrevételeikről.

4.3. Tesztelők

A mondatpárok meghallgatását 2007. október 29. és 2007. november 4. között 13 tesztelő végezte el. A tesztelők mindannyian ép hallású, magyar anyanyelvű emberek voltak, 20-64 év közötti korral. Egy részük tanszéki munkatárs, a témához értő volt, míg a többiek egyetemi hallgatók köréből kerültek ki. A rendszer rögzítette a teszt elkezdésének és befejezésének időpontját, így azt a tesztelőt kizártuk az eredmények kiértékeléséből, aki 10 percnél rövidebb idő alatt végezte el a tesztet (hiszen ennyi idő minimálisan szükséges lett volna az összes mondat meghallgatásához). A teszt átlagos meghallgatási ideje 19 perc volt.

4.4. Eredmények

A teszt kiértékeléséből az derült ki, hogy a tesztelők az esetek többségében a adatbázisbeli frázisok másolásával létrehozott dallamot preferálták a szabály alapú változathoz képest. A 4. táblázatben az összes mondatpár értékelése látható. A táblázat bal oldalán az egyes mondatpárokról szerepel információ (melyik milyen dallammenettel készült), a jobb oldalon pedig az, hogy a tesztelők melyiket tartották jobbnak.

Látható, hogy például a #0413-as mondat három különböző F_0-másolt változatát a tesztelők jobbra értékelték, mint a szabály alapú párjaikat. Az 1. és 3. F_0-másolt változat láthatóan jól teljesített, a másodiknál már nem ennyire egyértelmű a helyzet. Azt is észrevehetjük, hogy a két dallammásolt változat összehasonlítása során a tesztelők nem tudtak különbséget tenni köztük annak ellenére, hogy a 13. ábrán határozott eltérést fedezhetünk fel a két mondat dallammenetében. Ez arra enged következtetni, hogy a két változat eltérő ugyan, de mindkettő elképzelhető a valós beszédben is, például más helyzetben. Tehát sikerült olyan dallamú mondatokat létrehoznunk, amelyek az emberi prozódiai változatossághoz közelebb állnak, mint a szabály alapú dallammenet.

Olyan mondatpárok is voltak, ahol a tesztelők a szabály alapú változatot jobbnak minősítették (pl. #4120-as mondat). Ezeket megvizsgálva kiderült, hogy egyes esetekben az okozta a dallammásolás gyengébb minőségét, hogy a mondat végén túl magasak lettek az F_0-értékek. Ez akkor fordulhatott elő, amikor nem vettük figyelembe a frázisok adatbázisbeli pozíciójára

4. táblázat. A mondatok szintetizált változatainak értékelése.

Mondatpár				Tesztelők választása		
Mondat	1. változat	2. változat		1. jobb	egyforma	2. jobb
#0205	szabály alapú	F_0-másolt/1		5	4	3
#0211	szabály alapú	F_0-másolt/1		7	3	2
#0211	szabály alapú	F_0-másolt/2		3	2	7
#0211	szabály alapú	F_0-másolt/3		3	4	5
#0211	F_0-másolt/1	F_0-másolt/2		0	0	12
#0413	szabály alapú	F_0-másolt/1		1	0	11
#0413	szabály alapú	F_0-másolt/2		2	4	6
#0413	szabály alapú	F_0-másolt/3		1	0	11
#0413	F_0-másolt/1	F_0-másolt/3		2	10	0
#0515	szabály alapú	F_0-másolt/1		2	2	8
#0515	szabály alapú	F_0-másolt/2		2	1	9
#0515	szabály alapú	F_0-másolt/3		1	0	11
#0515	F_0-másolt/1	F_0-másolt/2		2	5	5
#1438	szabály alapú	F_0-másolt/1		3	4	5
#1438	szabály alapú	F_0-másolt/2		6	3	3
#1438	szabály alapú	F_0-másolt/3		2	9	1
#1438	F_0-másolt/2	F_0-másolt/3		1	6	5
#2411	szabály alapú	F_0-másolt/1		5	2	5
#2411	szabály alapú	F_0-másolt/2		11	1	0
#2411	szabály alapú	F_0-másolt/3		5	2	5
#2411	F_0-másolt/1	F_0-másolt/2		11	0	1
#2546	szabály alapú	F_0-másolt/1		4	1	7
#2546	szabály alapú	F_0-másolt/2		3	0	9
#2546	szabály alapú	F_0-másolt/3		0	2	10
#2546	F_0-másolt/1	F_0-másolt/2		3	5	4
#3733	szabály alapú	F_0-másolt/1		0	1	11
#3733	szabály alapú	F_0-másolt/2		1	1	10
#3733	szabály alapú	F_0-másolt/3		1	0	11
#3733	F_0-másolt/2	F_0-másolt/3		5	2	5
#4120	szabály alapú	F_0-másolt/1		9	2	1
#4120	szabály alapú	F_0-másolt/2		7	2	3
#4120	F_0-másolt/1	F_0-másolt/2		3	2	7
#4965	szabály alapú	F_0-másolt/1		0	0	12
#4965	szabály alapú	F_0-másolt/2		4	6	2
#4965	szabály alapú	F_0-másolt/3		3	4	5
#4965	F_0-másolt/1	F_0-másolt/2		3	4	5
#4965	F_0-másolt/2	F_0-másolt/3		3	7	2

vonatkozó kényszereket, így megtörténhetett, hogy a szintetizált mondat utolsó frázisa egy adatbázisbeli első frázisról kapta a dallamot, ami így túl magas lett. Az a következtetés vonható le ebből, hogy a dallammásolás során mindenképpen figyelembe kell venni a frázisok pozícióját.

Olyan eset is előfordult, amikor a módszer a dallamot túl magasra próbálta állítani, és ez már természetellenes torzulásokat okozott a szintetizált beszédben. További munkánk során tehát jobban kell figyelnünk arra, hogy az F_0 beállítása egy bizonyos határ fölé (illetve alá, hiszen a nagyon alacsony F_0 érték is természetellenes) ne történhessen meg.

Összességében elmondhatjuk, hogy a 10 mondatból 5 esetben egyértelműen az új, frázisok alapján működő F_0-másolási módszer volt jobb (#0413, #0515, #2546, #3733, #4965), 3 esetben nem lehetett dönteni a tesztelők véleménye alapján (#0205, #0211 és #1438), és 2 mondat esetében a szabály alapú megoldás minőségét értékelték jobbnak (#2411 és #4120).

A tesztelők megjegyzései közül fontos kiemelni, hogy egyesek nagyon zavarónak tartották a mondat végi dallam emelést, mert ott mindenképpen a legmélyebb hangot várja a hallgató. Mások szerint a mondatok meglehetősen hosszúak voltak, így nagyon kellett koncentrálni, hogy el lehessen dönteni közülük, melyik a természetesebb. A későbbiekben tahát figyelnünk kell arra, hogy összehasonlítási kísérleteinkben rövidebb mondatokat vizsgáljunk.

5. Felhasználási, továbbfejlesztési lehetőségek

Az általunk kidolgozott módszer segítségével természetesebbé tehető a szöfegfelolvasók által létrehozott prozódia. Ez az előny számos gyakorlati alkalmazásban használható, mint például SMS-, email-, könyv-felolvasó, vagy telefonos tudakozó. A változatosabb prozódia főleg hosszú szövegek felolvasása esetén előnyös, hiszen ekkor zavaró lenne a beszédszintetizátor monotonitása. A fő cél tehát az, hogy a módszert a Profivox beszédszintetizátorba beépítbe szélesebb körben használni lehessen azt.

Érdekes lenne megvizsgálni, hogy más beszédadatbázissal milyen eredményeket tudunk elérni. Olyan korpuszt lenne célszerű választani, amiben rövidebb mondatok vannak, amik jobban közelítik az általános beszéd mondathosszát. Ehhez azonban egy új adatbázis felvétele és címkézése szükséges, ami hosszú és fáradságos munka.

A módszert más nyelvekben is lehetne alkalmazni. A finn és lengyel nyelv a magyarhoz hasonló fix hangsúlyt használ, ami alkalmassá teszi módszerünk használatára. A magyar hangsúlyozási szabály szerint mindig a szó első szótagján van a nyomaték. Más, változó hangsúlyt használó nyelvekre (pl. angol, német) a dallammásolás megvalósítása bonyolultabb, de szintén lehetséges.

Azt az irányt is érdemes megvizsgálni, mi lenne ha a prozódia többi komponensét (elsősorban az időtartamokat) is korpusz alapján hoznánk létre. Ennek megvalósításához a már elkészített rendszerek működésének érdemes utánanézni a szakirodalomban.

Fontos megjegyeznünk újra, hogy a tesztelők rossz minőségűnek ítélték azokat a mondatokat, amiknek a végén szokatlanul magas volt a dallammenet. A továbbiakban tehát mindenképpen figyelembe kell vennünk ezt is.

Módszerünk pontosabb értékeléséhez több tesztelővel elvégzett kísérletekre van szükség.

6. Összefoglalás, eredmények összegzése

Munkánk célja az volt, hogy növeljük a szövegfelolvasók által létrehozott mondatok prozódiájának természetességét.

Jelen dolgozatban először áttekintettük a beszédszintetizátorok működésének megértéséhez szükséges szakirodalmat. Ezután olyan módszereket kerestünk az irodalomban, melyek korpusz alapon generálják a szintetizált beszéd prozódiáját. Összehasonlítottuk ezeket, megvizsgálva a módszerekben felhasznált prozódia minták méretét. Bemutatásra került egy prozódiai változatossággal foglalkozó cikk is.

A szakirodalomban ismert módszerek tárgyalása után áttekintettük korábbi munkánkat és ennek eredményét. Ismertettük a kidolgozott módszer továbbfejlesztési irányait. Egyrészt automatikussá tettük a prozódia másolását, másrészt nagyobb méretű beszédkorpuszban vizsgáltuk a korábbi módszer eredményességét.

Újabb kutatásainkhoz a beszéddallam-adatbázis módosítására volt szükség: a hangsúlyok címkézését kellett végrehajtnunk, valamint frázisokra bontottuk a korábban teljes mondatokból álló adatbázist. A dallammásolási módszert átalakítottuk úgy, hogy ezen prozódiai egységek alapján történjen a másolás. Vizsgáltuk azt is, hogy a prozódiai egységek dallamainak egymás után illesztésével mikor lehet a legtermészetesebb mondatdallamot létrehozni. Ezzel megnyílt

a lehetőség arra, hogy egy-egy szintetizálandó mondat alapfrekvencia-menetét többféle módon megtervezzük. Így a rendszer a szintézis során dönthet arról, hogy melyik F_0-görbét használja, amivel csökkenthető a hosszabb szintetizált beszéd monotonitása.

A módszerünkkel létrehozott mondatok minőségét egy webes tesztben ellenőriztük. Mondatpáronként kellett a tesztelőknek értékelniük a különböző dallamváltozatú mondatokat. Az eredmények kiértékeléséből kiderült, hogy a dallammásolással létrehozott szintetizált mondatok az esetek többségében jobbak a szabály alapú változatoknál.

7. Köszönetnyilvánítás

Ezúton mondok köszönetet konzulenseimnek, Dr. Németh Gézának és Dr. Fék Márknak a munkám során nyújtott segítségükért, hasznos tanácsaikért és észrevételeikért. Köszönettel tartozom továbbá Bartalis István Mátyásnak a webes tesztelő rendszer beállításáért, valamint a tesztet elvégzőknek a mondatpárok meghallgatásáért és értékeléséért, illetve a jövőben hasznosítható megjegyzéseikért. A munka a BME-TMIT Beszédtechnológiai Laboratóriumában készült.

8. Irodalomjegyzék

[1] Csapó Tamás Gábor, *Szintetizált beszéd természetesebbé tétele*, TDK dolgozat, BME-VIK, *Információs rendszerek* szekció, Budapest, 2006.

[2] Olaszy Gábor, Kovács Magdolna, Nikléczy Péter, Gósy Mária, *Magyar nyelvi beszéd-technológiai alapismeretek. (600 oldal CD-ROM-on).* Szerk.: Olaszy Gábor Nikol Kiadó, Budapest, 2002., `http://alpha.tmit.bme.hu/pub/beszinf/start.html`.

[3] Fék Márk, Pesti Péter, Németh Géza, Zainkó Csaba, „Generációváltás a beszédszintézisben", in *Híradástechnika*, Vol. LXI., no. 3., pp. 21–30., 2006.

[4] „Sprachsynthese", Bausteinauswahl, `http://www.ias.et.tu-dresden.de/sprache`.

[5] Olaszy Gábor, Németh Géza, Olaszi Péter, Kiss Géza, Gordos Géza, „PROFIVOX – A Hungarian Professional TTS System for Telecommunications Applications", *International Journal of Speech Technology*, Vol. 3, Numbers 3/4, December 2000, pp. 201–216.

[6] Mary E. Beckman and Gayle Ayers Elam, *Guidelines for ToBi Labelling*, The Ohio State University Research Foundation, pp. 8–12., 1993. `http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf`.

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

[8] Olaszy, G., Németh, G., Olaszi, P., „Automatic Prosody Generation – a Model for Hungarian", *Proc. Eurospeech 2001*, Vol. 1., pp. 525–528., 2001.

[9] Dong, M., Lua, K. T. „An Example-based Approach for Prosody Generation in Chinese Speech Synthesis", in *Proc. ISCSLP 2000*, Beijing, pp. 303–307., 2000.

[10] Meron, Joram, „Prosodic unit selection using an imitation speech database", in *Proc. SSW4-2001*, paper 113., 2001.

[11] Huang X., Acero A., Hon H., Ju Y., Liu J., Meridith S., Plumpe M., „Recent Improvements on Microsoft's Trainable Text-to-Speech System – Whistler", in *Proc. ICASSP97*, pp. 959–962, 1997.

[12] Raux, A., Black, A. „A Unit Selection Approach to F0 Modeling and its Application to Emphasis", in *Proc. ASRU 2003*, pp. 700–705., 2003.

[13] *The Festival Speech Synthesis System*, `http://www.cstr.ed.ac.uk/projects/festival/`.

[14] Van Santen, J., Kain, A., Klabbers, E., and Mishra, T. „Synthesis of Prosody using Multi-level Unit Sequences”, in *Speech Communication*, Volume 46, Issues 3-4, pp. 365–375., 2005.

[15] Takashi Saito, „Generating F0 Contours by Statistical Manipulation of Natural F0 Shapes”, in *IEICE Trans. Inf. & Syst.*, Vol. E89-D, No. 3., pp. 1100–1106., 2006.

[16] Min Chu, Yong Zhao, Eric Chang, „Modeling stylized invariance and local variability of prosody in text-to-speech synhtesis”, in *Speech Communication*, Vol. 48., pp. 716–726., 2006.

[17] Dacheng Lin, Yong Zhao, Frank K. Soong, Min Chu, Jieyu Zhao, „Iterative Unit Selection with Unnatural Prosody Detection”, in *Proc. Interspeech 2007*, pp. 2909–2912., 2007.

[18] Boersma, Paul, Weenink, David, *Praat: doing phonetics by computer*, (Version 4.6.34) [Computer program], 2006., `http://www.praat.org/`.

[19] Anne Tamm, Kálmán Abari, Gábor Olaszy, „Accent Assignment Algorithm in Hungarian, Based on Syntactic Analysis”, in *Proc. Interspeech 2007*, pp. 466–469., 2007.

[20] Géza Németh, Márk Fék, Tamás Gábor Csapó, „Increasing Prosodic Variability of Text-To-Speech Synthesizers”, in *Proc. Interspeech 2007*, pp. 474–477., 2007.