Szintetizált beszéd természetesebbé tétele

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

2006. október

Tartalomjegyzék

1. Bevezetés

A mai életben egyre fontosabb szerepet játszik az ember-gép kapcsolat, hiszen a világ az információs társadalom létrejötte felé halad. Egyre több mindent várunk el a számítógépektől, hogy egyszerűbbé tegyék mindennapjainkat. Ebben a folyamatban a beszédtechnológia, ezen belül a beszédszintézis is fontos szerepet játszik.

Az elmúlt években fokozatosan előtérbe került a jó minőségű gépi beszédkeltés egyik kulcskérdése, a megfelelő, de ugyanakkor változatos prozódia megvalósítása. A prozódiai változatosság alatt érthetjük az intonáció, a hangsúlyozás, vagy a ritmus variálását. Ma már nem elég, ha adott szöveghez mindig azonos prozódiát rendelünk valamilyen szabályrendszer segítségével. A cél az, hogy a gépi megoldás jobban modellezze az emberi beszédet olyan szempontból, hogy ugyanazt a mondatot mindig máshogy mondja, hiszen a valóságban sincs két egyformán kiejtett mondat, mert a prozódia egy adott személy beszédében is folyamatosan változik.

A dolgozat első felében áttekintjük a témához tartozó szakirodalmat. A prozódia legfontosabb összetevőinek részletes vizsgálata után a beszédszintetizátorok fejlődését mutatjuk be a kezdetektől napjainkig. Ismertetjük az irodalomban megtalálható prozódiai modelleket több szempont szerint csoportosítva. Megvizsgáljuk a magyar nyelv kijelentő mondatainak változatosságát egy konkrét példa során. Bemutatjuk azt is, hogyan végezhető el a természetes beszéd dallammenetének módosítása.

Saját kísérletek keretében első lépésben azonos bemondó hasonló jellegű bemondásait elemezzük, ezekből próbálva a mondatok dallamszerkezetére jellemző információt kinyerni. Ezen információ alapján kiválasztunk néhány mondatot, amiken különféle prozódiai módosításokat végzünk.

Koncepciónk ellenőrzésére egy egyszerűsített kísérletsorozatot tervezünk meg, majd meg is valósítjuk ezt. Levonjuk a következtetéseket, megvizsgáljuk az idő közben felmerülő problémákat.

Befejezésül kitérünk munkánk értékelésére és a további kutatási lehetőségekre és feladatokra.

2. Elméleti háttér

Ebben a fejezetben áttekintjük a dolgozat megértéséhez szükséges alapfogalmakat.

A 2.1. alfejezetben a prozódia három fő összetevőjének a leírása található.

A 2.2. alfejezetben a beszédszintetizátorok fejlődését és jelenlegi helyzetét mutatjuk be röviden.

A 2.3. alfejezetben az alapvető prozódiai modellek ismertetése történik meg, néhány konkrét típusról rövid áttekintés olvasható.

A 2.4. alfejezet egy konkrét magyar beszédkorpuszon végzett vizsgálatok egy részének eredményét mutatja be.

A 2.5. alfejezet a dallammenet módosításának elméletét, és a megvalósítás egy lehetőségét tartalmazza.

2.1. A prozódia legfontosabb összetevői

A prozódia (a dallam, a ritmus, a tempó, a hangsúlyozás, a hangerő és a hangszínezet változtatása) adja a beszéd kifejező erejét. Érzékeltethetjük vele a mondat fajtáját, a lelki állapotunkat, azt is, hogy mit tartunk fontosnak a mondanivalóból vagy mit nem, valamint, hogy milyen szituációt akarunk hangban a hallgató elé tárni. Ha szöveget olvasunk fel, a szöveg tartalma is befolyásolja a prozódiai megformálást. Ezek közül a három legfontosabb összetevő szubjektív szempontok szerint a dallam, a hangsúly, és a ritmus, amiket Olaszy és társai műve alapján ismertetek [1].

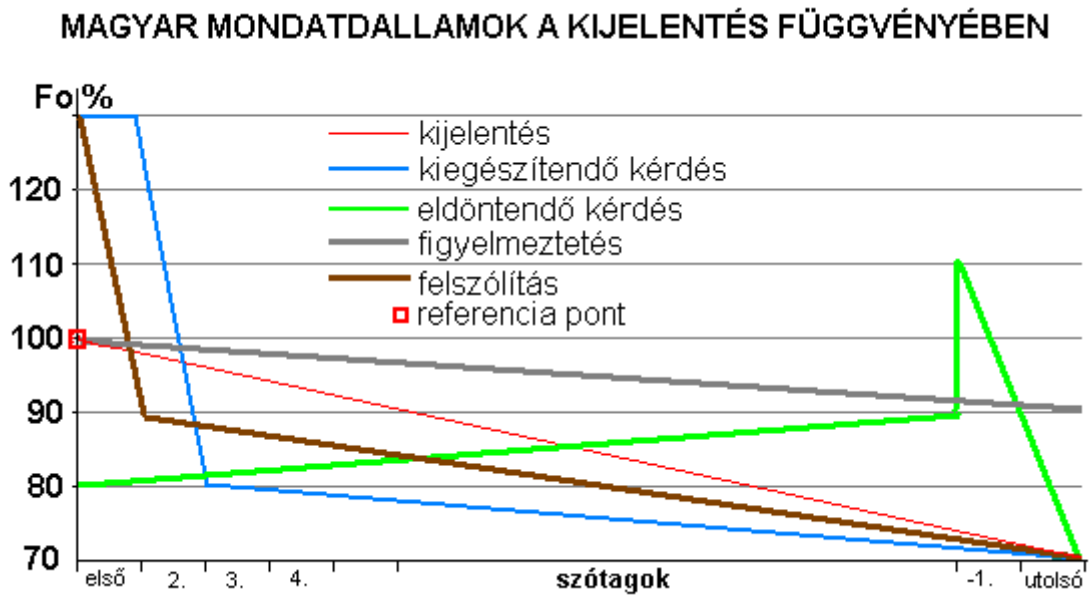
Ez a három összetevő objektív paraméterekkel is jellemezhető. A dallam a beszéd alapfrekvenciájának[1] változásával azonosítható. A hangsúly az intenzitás- és az alapfrekvencia-menet csúcsainak helyén érzékelhető. A ritmus a beszédrészletek időtartamainak változását jelenti. A dolgozatban ezen összetevők szubjektív és objektív elnevezését is fogjuk használni, mivel ezek megfeleltethetőek egymásnak.

2.1.1. Dallam

Az emberi beszédkeltés egyik legfontosabb eszköze a hanglejtés [1, 242. oldal], amit az alapfrekvencia változtatásával hoz létre a beszélő. Ez a változás adja többek között a beszéd dallamát is. A beszéddallamot alapvetően háromféle formával lehet jellemezni: emelkedő, ereszkedő és lebegő. A hanglejtés vizsgálatának, ismeretének, modellezésének fontos szerepe van a beszédszintézisben. A magyarban a hanglejtésnek mondat szinten van jelentésmegkülönböztető szerepe. A kijelentő mondat általános dallamformája az ereszkedő, a kérdő mondatoknál a végleges dallamforma a dallamépítő elemek legkülönfélébb kombinációjából áll elő, a felszólító és óhajtó mondatok is sajátos dallamformával rendelkeznek. A mondatszintű dallamok kezdő- és végpontjai szoros összefüggésben vannak egymással, és ezt a nyelv határozza meg.

Az 1. ábrán a leggyakoribb mondatfajták alapvető dallamformáinak egymáshoz való viszonya látható, relatív alapfrekvencia-értékekkel ábrázolva. A dallamformák erősen stili-

[1] A beszéd zöngés részei kvázi-periodikusak. Az alapfekvencia a periódusidő a reciproka.



1. ábra. A leggyakoribb mondatfajták dallamformái. Forrás: Olaszy és társai ([1, 242. oldal]).

záltak, csak a mondatok sémáját mutatják. A függőleges tengely az alapfrekvencia százalékos értékét mutatja, a vízszintes tengely a mondat szótagjaira utal. Az egyes mondatok dallammenetének relatív értékét a kijelentő mondat indulási frekvenciájához viszonyítják, ez a referencia pont. A magyar mondatdallamok általános tendenciájának jellemző összetevője az ábra alapján az eső dallamforma. Egyedüli kivétel az eldöntendő kérdés.

2.1.2. Hangsúly

A hangsúlyozásnál [1, 246. oldal] a beszélő nyomatékot tesz a hangsúlyozott szótagra. Fizikai paraméterekkel ez a nyomatékképzés három elemmel jellemezhető (F_0 emelés, intenzitás emelés, hangidőtartam nyújtás). A magyar hangsúlyozási szabály azt mondja ki, hogy a szavak első szótagja a hangsúlyos. A hangsúlyozást befolyásolja a beszélő érzelmi állapota, valamint a szöveg tartalma, annak értelmezése is. Ebből kifolyólag meghatározott esetekben hangsúlyozhatunk más szótagokat is, sőt többet is egy-egy szón belül. A szövegben a szótag lehet hangsúlyos, és lehet kifejezetten hangsúlytalan is. E két szint között több fokozat helyezkedhet el, ami a nyelvtől is és a beszélőtől is függ.

A folyamatos beszédben nem egymástól elválasztott szavakat mondunk, hanem a szavakat egymásba fűzzük és úgynevezett szólamokban, prozódiai egységekben beszélünk. A mondat egy vagy több szólamból áll. A szólam olyan szófüzér, amely logikailag egy egységet alkot. Ezt folyamatos egységként mondjuk ki, szünet nincs benne. A szólam egésze egy hangsúlyegységet képvisel és csak az első szaván van normál hangsúly. A legerősebb szólamhangsúlyt mondathangsúlynak nevezzük.

2.1.3. Tempó, ritmus

A beszéd ritmusa [1, 249. oldal] a beszédtempó változását, hullámzását és a szüneteket foglalja magába. A beszédtempó és a ritmus számos tényezőtől függ: a nyelvtől, a beszélő egyéniségétől, a pillanatnyi érzelmi állapottól, a témától, a beszédhelyzettől stb. Ezért beszédtechnológiai szempontból is fontos a beszéd ritmusszerkezetének részletes vizsgálata és esetleges modellezése. A folyamatos beszédben a hangsor különböző részein változtatjuk az artikulációs sebességet (a hangokat hol gyorsabban, hol lassabban ejtjük). A hangidőtartamok változása ilyenkor 10-20 %-on belül mozog.

2.2. Beszédszintetizátorok generációi

A beszédszintézis célja a bemeneti információ beszéddé alakítása. Ez a bemenet legtöbbször írott szöveg, ekkor szövegfelolvasó rendszerről beszélünk. Egy ilyen rendszer két alapvető részből áll: először a bemeneti szöveget alakítja szimbolikus információvá, majd ez alapján hangfájllá. A közbenső szimbolikus információ általában a szöveg tartalmát megadó fonémasorozatból és a szöveg prozódiai jellemzőit leíró információkból (hanglejtés, hangsúlyok, ritmika) áll. A beszédszintetizátoroknak három fő generációját különböztethetjük meg, amit Fék és társai munkája alapján foglalunk össze [2].

A szintézis során a zöngés és zöngétlen hangokat eltérő módon kell kezelni. A zöngés gerjesztés alapfrekvenciájának a vezérlését az úgynevezett prozódiai modul végzi, aminek bemenete a szövegből előállított fonémasorozat és a szimbolikus prozódia. Ez utóbbi a mondatok modalitását (kijelentő/kérdő) és a hangsúlyok helyét és típusát adja meg. A prozódiai modul előírja az előállítandó hullámforma alapfrekvencia- és intenzitásmenetét, illetve az egyes hangok időtartamait.

2.2.1. Formánsszintézis

Az első olyan technológia, amellyel egy szöveget automatikusan jól érthető beszéddé lehet alakítani, a formánsszintézis volt. Ez egy gerjesztett szűrőrendszer kimeneteként állítja elő a beszédjelet, ami az emberi beszédkeltést modellezi. Mivel a formánsszintézishez szükséges jó minőségű paraméterek automatikus előállítását nem tudják megvalósítani, az ilyen rendszerek hangzása érthető ugyan, de többnyire nagyon „robotos”. Ezért a formánsszintézis háttérbe szorult.

2.2.2. Elemösszefűzés

Az elemösszefűzésen alapuló beszédszintézisben a természetes beszédből kivágott hullámforma elemeket fűzünk össze. Az egyik alapvető kérdés, hogy melyek legyenek a gépi beszédet előállító elemek. Több szempontot is figyelembe kell venni: egyrészt teljes fedés kell, azaz az adott nyelv tetszőleges hangsorozatát elő kell tudni állítani, másrészt az előállított beszédnek minél természetesebben kell szólnia. Ha a fonémáknak megfelelő hangokat használjuk elemekként, az teljes fedést biztosít, de az összefűzés során előállított jel nem hangzik folytonosnak, mert a fonémák egymásra is hatással vannak, amit így nem tudunk biztosítani. Az előálló

hang minősége tehát gyenge. A megoldás környezetfüggő hangok használata is lehet, amikor minden egyes hangot minden lehetséges hangkörnyezetének megfelelő változatban tárolunk. Az ilyen elemeket triádoknak nevezzük. A gyakorlatban bevált megoldás a két egymás utáni félhang együtteseként előálló diád alkalmazása, amiket további triád-elemekkel lehet bővíteni. Az elemek összefűzése után gondoskodni kell arról, hogy az előálló beszédjel kövesse a prozódiai modulban előírt jellemzőket: az alapfrekvencia- és intenzitásmenetet, hangidőtartamokat. 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 előállított beszéd érthető, de még nem teljesen természetes hangzású.

A dolgozat során a BME-TMIT[2]-en Olaszy és társai által kifejlesztett Profivox [3] beszédszintetizátort fogjuk használni különböző tesztek 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 és egy női változatot fogunk alkalmazni.

2.2.3. Korpusz alapú, elemkiválasztásos

Az elemösszefűzéses technológia továbbfejlesztése a korpusz alapú, elemkiválasztásos beszédszintézis. A korábbiakhoz képest itt új elvekkel találkozhatunk. Az egyik az, hogy a korpusz alapú szintézis elve szerint a beszédszintetizátor hangadatbázisa teljes mondatokat tartalmaz, nem pedig kivágott diád- illetve triádelemeket. A hagyományos elemösszefűzéssel ellentétben tehát a korpuszos adatbázisban egy adott hangsorhoz tartozó beszédelem általában többféle formában is előfordulhat, amiknek eltérő lehet a dallammenete, intenzitása, vagy a hangszíne. Így egy adott szintetizált beszédszakasz több lehetséges módon előállítható, amik közül a legtermészetesebbet kell kiválasztani.

A korpusz alapú elemkiválasztásos beszédszintézis két fő ok miatt eredményez javulást a hagyományos elemösszefűzéshez képest: egyrészt kevesebb összefűzési pontot tartalmaz, valamint a megfelelő prozódia kiválasztásához kevesebb jelfeldolgozási művelet szükséges. Ha mégis szükség van a prozódia módosítására, viszonylag egyszerűen eltolással biztosítható az egybefűzött darabok illeszkedése. A korpusz alapú szintézis ugyanakkor nagyobb tárigényt és számítási kapacitást igényel, mint elődei. A rendszer minőségét az is nagyban befolyásolja, hogy a szintetizálandó szöveg mennyire van közel a szintetizátor beszédkorpuszához, emiatt célszerű a korpuszt az adott alkalmazáshoz igazítani.

2.3. Prozódiai modellek csoportosítása

A 2.2. alfejezetben említettük a szövegfelolvasók két alapvető modulját. Az első létrehozza a bemeneti szövegből a szimbolikus leírást, míg a második a szimbolikus leírás alapján hozza létre a kimeneten a hullámformát, vagyis megadja a fizikai paramétereket. Ez a két lépés ugyanakkor egyben is történhetne, hiszen a köztes szimbolikus információra nem mindig van szükségünk.

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

A prozódiai modelleket tehát osztályozhatjuk aszerint, hogy milyen bemenet alapján milyen kimenetet képesek előállítani. Eszerint lehetnek:

1. fizikai paraméterek → szimbolikus leírás
2. szöveg → szimbolikus leírás
3. szimbolikus leírás → fizikai paraméterek
4. szöveg → fizikai paraméterek

Az első típusú modellel az a probléma, hogy a megfelelő paraméterek származtatása nehéz művelet, pedig a feladat a beszélő prozódiai szándékának leírása lenne. A második célja, hogy a természetes beszéd F_0 menetét és más jellemzőit modellezze minél pontosabb eszközökkel. A harmadik típus ad lehetőséget a prozódiai paraméterek (pl. az F_0 görbe) generálására megfelelő bemenet alapján. Hagyományosan a fizikai prozódia paramétereket létrehozó szövegfelolvasó komponenst úgy tervezik meg, hogy csak a szimbolikus információt használja bemenetként, ezért csak olyan dolgok kifejezésére képes, ami a szimbolikus reprezentációban definiált.

A különböző intonációs (F_0 változás leíró) módszereket osztályozhatjuk Kochanski és társai kategorizálása alapján [4] aszerint, hogy mennyire részletesek. Az alulspecifikált rendszerek minél kevesebb címkével próbálják leírni az intonációt. A részletesebb rendszerek több mintát engedélyeznek például a hangsúlyok közelében, hiszen ott fontosabb az alapfrekvencia pontos ismerete, a többi helyen pedig aszimptotikus közelítéssel, vagy lineáris approximációval hozzák létre a dallammenetet. A másik oldal, a teljesen specifikált rendszerek lényege, hogy gépi tanulási technikákat alkalmaznak, ami sűrű mintájú F_0 értékeket eredményez, így a nagyon komplex helyeket is részletesen le lehet írni.

Az alulspecifikált hangsúlymenet használatának előnye, hogy szabadabban kezelhetjük a specifikált részek közötti egységeket, mégpedig tipikusan interpolációs közelítéssel. A hátránya viszont, hogy nem foglalkozik a specifikált helyek közötti egységek változásaival. Másrészről a teljesen specifikált rendszerekben nincs elég mozgástér a konfliktusok megfelelő kezelésére. Ha teljesen specifikált hangsúlyokat egyszerűen egymás után helyezünk, az olyan beszédhez vezet, amiben természetellenes ugrások vannak az összefűzési pontoknál. Sok rendszerben szűrőket alkalmaznak, hogy kikerüljék az ilyen hirtelen ugrásokat.

Más csoportosítás szerint a prozódiai modellek lehetnek leíró jellegűek, szabály alapúak, adatvezéreltek, vagy szuperpozíciósak, ami ez előbbiek kombinációját jelenti.

2.3.1. Leíró jellegű modellek

ToBi A ToBi[3] rendszer [5] alulspecifikáltnak számít, mert ennek a leíró jellegű modellnek az a célja, hogy az intonációt egy minimális címkehalmazzal írja le. A ToBi leírás [6] minimálisan a felvett beszédből, a hozzárendelt alapfrekvencia-menetből, valamint a szimbolikus információ négy kategóriájából áll. Ezekből az egyik a hanglejtés, ami a jellegzetes alapfrekvencia változásokat jelenti, amiket alapvetően magas (H, high) és alacsony (L, low) szimbólumokkal jelölnek. Egyéb kiegészítő jelöléseket alkalmaznak a szóhangsúlyra (L*), frázis hangsúlyra (H-), illetve a határhangokra (H%).

[3] Tones and Break indices

A ToBi a bemenetből szimbolikus információt hoz létre. Azonban ez a lépés korántsem triviális, mert eddig nem ismert olyan megoldás, ami automatikusan létre tudná hozni a címkéket. A ToBi tehát kézi címkézéssel működik, ami időigényes és drága is.

GToBi A GToBi[4] [7] a német intonáció fonológiai strukturájának címkézését összefoglaló konvenciók halmaza. A célja az, hogy könnyen tanulható, megbízható legyen, és lehessen használni különböző címkézési metódusokkal. Közel áll az angol ToBi rendszerhez. Spontán és felolvasott beszédre is alkalmazható, illetve nemrég módosították, hogy a rendszer fonetikailag transzparensebb legyen, és az intonációs fonológia legújabb eredményeit is magába foglalja.

A GToBi minimálisan három címke osztályból áll: hangok, szünetek és szavak. A hangok kategóriájában az alapfrekvencia görbét módosító jelekkel lehet változtatni, úgymint emelés ('!') és csökkentés ('^'), amit közvetlenül a vonatkozó hang elé kell helyezni. A címkék teljes leltára tartalmaz két egy egytagú (H*, L*) és négy kéttagú hangsúlyt (L+H*, L*+H, H+L*, H+!H*), valamint olyan hangokat, amik a prozódiai egységek határán vannak (L- vagy H-), illetve nagyobb egységek határán (L% vagy H%). Így tehát az eredeti ToBi szimbólumokhoz újakat vettek hozzá a német követelményeknek megfelelően. A szimbolikus reprezentáció így pontosabb eredményt ad, de még mindig nem megoldott a címkék automatikus vagy legalább félautomatikus módszerekkel történő származtatása.

IViE Az IViE[5] [8] egy gép által olvasható prozódiai címkéző rendszer. Az angol prozódia korábban ismertetett de facto szabványán, a ToBi-n alapul. Elsősorban két dologban különbözik a ToBi az IViE-től. Először is, az eredeti ToBi rendszert az angol nyelv hivatalos változatára fejlesztették ki. Az IViE egy rendszerben kezeli a hivatalos és az egyéb angol nyelvi változatokat. Másodsorban a ToBi-ban az intonációt csak egy kategória írja le. Az IViE-ben az intonációs szerkezetet három külön osztályra szedték szét. Az első osztály címkéi a ritmusról és a hangképzésről hordoznak információt. A másodikban az alapfrekvencia változásai vannak leírva, míg a harmadik a kiejtések intonációjáról tartalmaz adatokat. Az angol nyelv változatai különbözhetnek a ritmusban, bizonyos hangsúlyok kiejtésbeli megvalósításában, illetve a hangok tárában. Éppen ezért az IViE-ben mindhármat be lehet állítani.

2.3.2. Szabály alapú modellek

A prozódia modelljét szabályok segítségével is meg lehet adni. A szöveget a korábbiakhoz hasonlóan címkézni kell, például a mondatokat típusuk szerint, a szavakat, szótagokat a rajtuk eső hangsúly szempontjából kell kategorizálni. A hangokat pedig szóban elfoglalt helyük szerint, hangrendjük és környezetüktől függően osztályozhatjuk.

Ezen modellek nagy előnye a kiszámíthatóság, vagyis mindig hasonló minőségű prozódiát hoznak létre. Azért fontos ez, mert az ember többnyire rosszul tűri a változást, ha már megszokott egy adott minőséget.

[4] German Tones and Break indices

[5] Intonational Variation in English

A szabály alapú modellek hátrányai közé tartozik, hogy a szabályok megalkotása nehéz, több tudományterület ismerete szükséges jó minőségű rendszer készítéséhez. Ugyanakkor ha hibát fedezünk fel a módszerben, akkor a szabályok javításával megoldható a probléma. A természetes nyelvek nem reguláris szerkezetűek, vagyis mindig vannak kivételek, amiket külön kezelni kell. Az is megemlítendő, hogy az emberek által beszélt nyelvek folyamatosan változnak, amihez a szabályoknak alkalmazkodniuk kell.

A gyakorlatban általában valamilyen másik módszerrel ötvözve alkalmazzák a szabály alapú megközelítést.

2.3.3. Adatvezérelt modellek

A korábban, 2.2 alfejezetben látott kétlépcsős megközelítés mögött az az ésszerű magyarázat, hogy a nyelvi jellemzők jobban korrelálnak a szimbolikus prozódiával mint az akusztikus megvalósítás. Így nem csak egyszerűbb az ember számára olyan szabályokat írni, amik becslik a prozódiát, hanem gép számára is könnyebb megtanulni a szabályokat az adatbázisból.

Sajnos a ToBi címkézés illetve változatainak kézi és félautomatikus megvalósítása lassú és drága, ezért nagyon hiányzik egy teljesen automatikus eljárás. Gépi tanulás segítségével lehetségessé válhat az automatikus prozódiai címkéző létrehozása.

ToBi nélkül, CART-okkal Volker Strom prozódiai modellje [9] négy CART-ból[6] áll. A CART módszer lényege, hogy minden lépésben kettéosztjuk az adathalmazt úgy, hogy a keresett változó értéke minél megjósolhatóbb legyen. A négyből kettő CART bináris döntéseket hoz a hangsúlyok helyéről és határaikról. A másik kettő meghatározza a szótagonkénti F_0 értéket. A két CART pár tehát a szimbolikus és az akusztikus prozódia becslőt jelképezi.

Az ily módon működő prozódiai modellben a hangok egy részére sikerült kimutatni, hogy jelentős javulás állt elő a prozódiai előrejelzés hagyományos változataihoz képest.

Neurális háló Jianhua Tao és társai neurális háló alapú prozódiai modelljével [10] kellő méretű tanító adatbázis esetén jobb minőségű alapfrekvencia menet generálása lehetséges, mint a hagyományos modellekkel, mivel ez a struktúra a tesztek szerint pontosabban jellemzi a prozódiát. Az eredmény természetessége sokat fejlődött, és a rendszer megpróbál rugalmas lenni a gyakorlatban. Egy fuzzy csoportosító algoritmus segítségével állítják be a szótagok F_0 értékeit. Az algoritmus bizonyíthatóan optimalizálja a neurális hálót és alkalmas a Mandarin nyelvű beszéd alapfrekvencia menetének előállítására.

2.3.4. Szuperpozíciós modellek

A szuperpozíciós modellek jellemzője, hogy a prozódia összetevőinek különböző szintű (szegmentális, szupraszegmentális, esetleg a kettő közötti) megvalósítását adja össze, azaz szuperponálja egymásra.

[6] Classification And Regression Tree

Automatikus prozódia generálás, modell a magyar nyelvre Olaszy és társai modelljében [11] egy bonyolult szabály-halmaz írja le a felolvasandó szöveg három prozódiai komponensét (időszerkezet, alapfrekvencia-menet és intenzitás). Mindhárom komponens külön-külön egy három lépéses eljárás során áll elő.

A prozódiai modellek nyelvfüggőek bizonyos értelemben. Például egy nyelv időbeli szerkezetéből csak az általános jellemzők hasonlíthatóak össze más nyelvekkel. Az időbeli struktúra részletes vizsgálatához nyelvspecifikus szabályokra van szükség. Az itt bemutatott időszerkezeti modell más nyelvekben is használható, de a részletes szabályok csak a magyar nyelvre igazak. Ugyanez vonatkozik az alapfrekvenciára és az intenzitásra is.

Az idő-struktúra komponens lényege, hogy a beszédet szét tudjuk választani szegmentális és szupraszegmentális szintre. A szegmentális szint a beszélő szándékától független jellemzőket jelenti. A szupraszegmentális szinten a szószerkezet és a mondat szintű hatások befolyásolják a hangok időtartamait.

Az F_0 modellezése a legmagasabb (szupraszegmentális) szinten kezdődik, először a mondat szintű dallam létrehozása történik meg. A mondat dallamának három típusú összetevője lehet: emelkedő, egyenletes és eső. A középső szinten a szó és szótag szintű alapfrekvencia változásokat képezik egymásra. Ezek a változások dinamikusan a mondat szintű egységek alapján történnek. A szavak szintjén két típusú struktúra van: semleges szó, és negatív hangsúllyal rendelkező szó. A szótagok F_0-menete emelkedő, eső, vagy emelkedő-eső lehet. Ezen szabályok segítségével létre lehet hozni a hangsúlyokat, és más szótag szintű változásokat. A szótag szintű változtatások egy vagy két szótagra vonatkoznak. Végül a legalacsonyabb (szegmentális) szint következik, aminek során a mikrointonációs változások is rákerülnek az alapfrekvencia görbére.

Az intenzitás modellezése a szegmentális szinten kezdődik és szupraszegmentális szinten végződik. A magyarban a szavak általános intenzitásszerkezete eső jellegű, ezt kell létrehozni a szintézis során is. A negatív hangsúlyú szavak, és a mondatok utolsó szavai általában kisebb intenzitásúak. Szótag szinten egy hangsúly kifejezése emelkedő intenzitást eredményez. A végső intenzitásszerkezetet a teljes mondat szintjén határozzuk meg.

A bemutatott modell mondat és szöveg szinten leírja a prozódia három alapvető elemét. A módszer előnyei: személyfüggetlen, statisztikai mérések segítségével tovább finomítható, valamint szabályok definiálása is könnyen megy. A modellt sikeresen tesztelték magyar nyelvű szintetizált beszéd létrehozása során, és alkalmazzák is a Profivox beszédszintetizátorban.

2.4. Prozódiai változatosság elemzésének első lépései

Ebben az alfejezetben egy a magyar nyelvre elvégzett kutatás bizonyos részleteit mutatjuk be, amely bár korpusz alapú beszédszintézis (2.2.3. alfejezet) témájú, de számunkra is vannak benne fontos megállapítások.

A korpusz alapú beszédszintézis nyelvi, fonetikai kérdéseinek elemzése során Olaszy [13] vizsgálatokat végzett beszédadatbázisokon a bennük szereplő mondatok alapfrekvencia- és intenzitásmenetének elemzésével.

2.4.1. Kijelentő mondatok vizsgálata

A vizsgált beszédadatbázisokban a bemeneti hullámforma fájlok mellett a fonemikus átírás is szerepel, illetve az elhangzott beszédjel hang- és szószintű címkéi, valamint az ezekhez tartozó időtartam, alapfrekvencia és intenzitás adatok. A célkitűzés megvalósítására egyszerű kijelentő mondatok alapfrekvencia- és intenzitásszerkezetét elemezték. A mintamondatok mindegyikén jellemezték az alapfekvencia változást annak töréspontjaival, valamint az intenzitások alakulását.

A vizsgált kijelentő mondatokra egyenként elvégezték az F_0 menet elemzését és a jellemző pontok hozzárendelését a szöveghez. 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 mondat belsején nem lehet jellemző F_0 karakterisztikát találni. Ez 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.

Az intenzitáskép egységesebb, mint az alapfrekvencia. A mondat kezdetén kialakuló intenzitás jellemző a mondat nagy részére, a befejező szakaszban az intenzitás csökken.

Látható tehát, hogy a kijelentő mondatokban a prozódia és a szöveg kapcsolatának kijelölése bonyolult szövegelemzést is igényelhet, hogy a megfelelő prozódiai részeket függetlenítsük a szöveg tartalmától.

2.5. A dallammenet módosítása

2.5.1. Dallammenet beállítása a Profivox-ban

A Profivox magyar nyelvű beszédszintetizátorban [14] a prozódia mindhárom fő paramétere (intenzitás, időtartam és alapfrekvencia) külön állítható.

Az amplitúdó értékek minden hangra %-ban adhatóak meg. Ennek az alapja a szegmentális szintű intenzitás, amit előzetesen definiáltak a hangadatbázis elemeire. A lineáris ereszkedő jelleg beállítása a szavakban és a mondat egészén is megtörténik.

Az alapfrekvencia generálás két alapfrekvencia beállító eljárás kombinációjával valósul meg. Az első az alapvető dallamot reprezentálja, ami a teljes mondatra vagy mondatrészekre vonatkozik. A második pedig szavak szintjét fejezi ki. Ezek a minták a hangsúlyok megvalósításából, és egyéb dallammenetbeli csúcsok (például kérdésekben) létrehozásából állnak. A szó szintű alapfrekvenciaváltozásokat rávetítik az első struktúrára a következőképpen: például ha a mondat egészének dallama eső jellegű 100 Hz-től 70 Hz-ig, és definiálunk egy 10 %-os

pozitív szó szintű alapfrekvencia emelést, akkor ez 10 Hz-es emelkedést jelent a mondat elején és 7 Hz-eset a mondat végén. A rendszer 30 különböző mondat-szintű dallamot tud kezelni.

2.5.2. A TD-PSOLA eljárás

A 2.5.1. szakaszban említettük, hogy a Profivox rendszerben szükség van az alapfrekvencia módosítására, ami az itt leírtakhoz hasonlóan történik meg. A beszéd alapfrekvenciájának módosítását úgy kell megoldani, hogy az eredeti hangszín ne változzon. Ennek egyik megvalósítási módja a TD-PSOLA[7] eljárás, ami zöngeszinkron átlapolásos összeadást jelent.

Lényege, hogy rövid (célszerűen egy-három periódusnyi) hullámforma-szegmenseket elemzünk úgy, hogy a beszédjel zöngés szakaszában minden hangperiódusra átlapolt ablakfüggvényt fektetünk, majd ezt a hangszakaszt spektrálisan jellemezzük. Ezután ezeket a kiablakolt jelrészleteket egymásra csúsztatjuk és összeadjuk. Így, a zöngés hangszakaszok periódusainak az időtartamát, tehát az alapfrekvenciát változtathatjuk meg, vagyis a beszédjel dallamát valósíthatjuk meg jelfeldolgozás segítségével.

A PSOLA eljárás alkalmazásának két fontos kritériuma van. Az első, hogy a beszédjelet el kell látni zöngeszinkron jelekkel, vagyis minden hangperiódusban ki kell jelölni a periodicitást mutató jelzőt. Ezt célszerû a periódus legnagyobb energiájú pontjára tenni. A második, az ablak alakjának és hosszának optimális meghatározása. Az időfüggvényen végzett PSOLA-transzformáció esetén két periódusnyi hosszú (Hamming, illetve Hanning) ablakot célszerű alkalmazni, ami egy haranggörbéhez hasonlítható.

Ez a kijelölési forma azt biztosítja, hogy az egyes ablakolt jelrészek között mindig lesz átlapolás. Ha ezt az ablakolást minden hangperiódusra elvégezzük, akkor olyan adatsorozatot kapunk, amelyben minden ablakra elvégzett analízis egy-egy hangperiódust képvisel. Ha ezeket összeadjuk (az idő tengelyen változatlan periodicitással), akkor visszakapjuk az eredeti időfüggvényt.

Ha az ablakokat időben például közelebb csúsztatjuk egymáshoz az összeadás előtt, akkor a végeredményként kapott időfüggvényben a hangszínezet nem változik (a formánsstruktúra ugyanaz marad, mint az eredeti jelben), a hangmagasság azonban magasabb lesz, mint a kiinduló jelben. Ugyanígy, ha távolítjuk egymástól az ablakokat és úgy adjuk össze őket, a beszédjel alapfrekvenciája mélyebb lesz. Az eljárásból adódik a korlát is: az alapfrekvenciát elvileg maximum kétszeresére, illetve a felére lehet megváltoztatni, azonban a gyakorlatban 30 %-nál nagyobb módosítás már erős torzulásokat eredményez.

[7] Time Domain Pitch Synchronous Overlap Add

3. Szintetizált beszéd dallammenetének változatosabbá tétele

A 3.1. alfejezetben megfogalmazzuk azokat a követelményeket, célokat, amiket kísérleti munkánk során megvalósítani próbáltunk.

A 3.2. alfejezet a kutatásunkban felhasznált beszéddallam-adatbázisokat mutatja be.

A 3.3. alfejezetben részletezzük, hogy milyen szempontok szerint kerestünk mondatokat a beszédkorpuszokból, és melyeket választottuk ki további feldolgozásra.

A 3.4. alfejezetben az olvasható, hogyan lehet megvalósítani a 3.1. alfejezetben megjelölt célunkat, az előző alfejezetben kiválasztott mondatok dallamának módosítását.

3.1. Alapvető cél

Napjainkban a jó minőségű gépi beszédkeltés egyik kulcskérdése a megfelelő, de egyben változatos prozódia (intonáció, hangsúlyozás, ritmus) megvalósítása. Ma már nem elég az, ha egy szabályrendszer segítségével a nyelvi elvárásoknak megfelelő, adott szöveghez mindig azonos prozódiát rendelünk. A cél az, hogy a gépi megoldás abból a szempontból is hasonlítson a természeteshez, hogy ugyanazt a mondatot ne mindig ugyanúgy mondja, hanem legyen benne némi (sokszor az adott személyre jellemző) változatosság.

A 2. fejezetben, a szakirodalom áttekintése során nem találkoztunk olyan megvalósítással, ami ezt a problémát orvosolni tudná, ezért kezdtünk kísérletezni a prozódia változatosabbá tételével.

A változatosságot nem könnyű elérni, hiszen a beszédszintézis során a bemeneti információ az írott szöveg, amiben nem jelennek meg a prozódiai összetevők. Így a prozódiát a szintetizátornak kell létrehoznia a bemenet alapján úgy, hogy legyen lehetőség többféle variáns közül választani.

Sokféleképpen lehet értelmezni a változatosságot: pl. dallam, hangsúlyok, ritmus, vagy hangerősség átalakításával. A jelenlegi dolgozatban kizárólag a dallam, vagyis a mondatok alapfrekvencia-menetének módosításával foglalkoztunk, a többi paramétert változatlanul hagyva.

3.2. Beszéddallam-adatbázis kiválasztása

A dolgozatban kétféle beszéddallam-adatbázist vizsgáltunk. Fék és társai [2, 24–25. oldal] munkája alapján röviden bemutatjuk ezek szerkezetét.

Az adatbázisok elsődlegesen hullámforma fájlokból állnak, mindegyik fájl egy-egy mondatot tartalmaz. Minden egyes hullámforma fájlhoz tartozik szöveges átírás, ami a mondat felolvasásához használt szövegből lett származtatva. A mondatokban az elemhatárok, illetve zöngeperiódus-határok jelölése automatikusan történt, amik az adatbázisban külön fájlokként jelennek meg.

Először a szöveg fonetikus változatára volt szükség. A fonetikus átírás automatikusan készült el, a magyar nyelv hasonulási szabályai szerint. A szavakon átívelő hangegybeolvadások miatt előfordul, hogy egy hang egyszerre két szóhoz is tartozik. A szavak közötti átfedés speciális jelölést igényelt (például: <számíthatun<k>isebb> , ahol a "<" egy szó kezdetét, a ">" egy szó végét jelöli).

A hang-, szó- és szünethatárok jelölése egy, a BME TMIT-en Mihajlik és társai által kifejlesztett beszédfelismerő segítségével lett megoldva [15]. A határok jelölése a beszédjel szintjén történt, azaz a hanghatárokat leíró fájlba az került, hogy hányadik mintán kezdődnek a hangok.

A zöngeperiódusok a beszéd zöngés részének egy-egy periódusát jelentik. A zöngétlen részeken a beszéd nem periodikus, ezért ott 5 ms-onként vannak jelölések. A zöngeperiódus-határok jelölése egyrészt az alapfrekvencia pillanatnyi értékének meghatározásához kell, másrészt az alapfrekvencia-menet módosításához szükséges. A jel alapfrekvencia-menete kiszámítható a zönge periódus időtartamok reciprokaként. A zöngeperiódus-határok bejelölése a Praat[8] fonetikai és beszéd-analizátor szoftverben implementált alapfrekvencia-detektálás alapján történt [16].

3.2.1. Időjárás előrejelzés hangkorpusz

Ez egy időjárás-jelentés témájú nagy korpusz [2, 24–25. oldal] 200 mondatos részhalmaza. A részadatbázist a BME-TMIT bocsátotta rendelkezésünkre, mivel ott egy korábbi kutatás során már megtörtént a hullámforma fájlok és a szöveges átírás alapján az elemhatárok és zöngeperiódus-határok megjelölése. Az adatbázis egy prozódiai egységgel rendelkező kijelentő mondatokból áll, ami azért fontos, mert az ilyen mondatoknak lényegesen egyszerűbb a dallammenetét vizsgálni. A 2.4.1 részben leírtak alapján tudjuk ugyanis, hogy a kijelentő mondat F_0 menetében változást okoznak a prozódiai egységek határai is.

Az adatbázis szerkezetére jellemző, hogy minden mondat (és a címkézése) külön fájlban van eltárolva. Például a 2031-es számú mondathoz tartozó fájlok adatai:

2031.wav - a 2031-es számú mondatot tartalmazó hullámforma fájl

2031.txt - a 2031-es számú mondat szöveges átírását tartalmazó szövegfájl

2031.ssw - a 2031-es számú mondat elemeinek határai, szünetek, hangok és szavak kezdete mintaszámban megadva

2031.TextGrid - a 2031-es számú mondat elemeinek határai, a Praat által értelmezhető formában, szünetek, hangok és szavak kezdete és vége másodpercben megadva

2031.pit - a 2031-es számú mondat zöngeperiódus-határai, vagyis a mondat alapfrekvencia-menete

3.2.2. Harangok története hangkorpusz

Szintén a BME-TMIT bocsátotta a kutatás rendelkezésére ezt az adatbázist, ami 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.

Ez a beszéd-adatbázis még nem volt előkészítve, felcímkézve, mint az időjárás előrejelzéses hangkorpusz. Négy harangismertetés alapján hoztuk létre a végső adatbázist.

[8] A Praat egy beszédanalizáló és módosító program, mely a nemzetközi gyakorlatban széleskörűen elterjedt.

Első feladatként az összefoglalókat mondatokra kellett bontani, majd a beszédet kellett visszaalakítani szöveges formába, vagyis leírni a hallott szöveget. Így a fonetikus átírás már elvégezhető volt, ami alapján a BME-TMIT-en kifejlesztett beszédfelismerőt [15] kényszerített módban alkalmazva létrehoztuk a mondatok elemhatárait leíró címke fájlokat, amik az időjárás előrejelzés hangkorpuszhoz hasonlóak. A zöngeperiódus-határok jelölése is a korábbiakhoz hasonlóan automatikus módszerekkel történt.

Az adatbázis szerkezete egyezik a 3.2.1 részben leírtakkal.

3.3. Mondatok vizsgálata

A beszéd-adatbázisok mondatait összehasonlítottuk abból a célból, hogy a későbbiekben felhasználható jellemzőket tudjunk kinyerni a mondatokból. Először hasonló hosszúságú mondatokat kerestünk, majd ezek között hasonló dallamszerkezetűeket. Az volt a feltevésünk, hogy ha találunk hasonló alapfrekvencia-menetű mondatokat, akkor ezek dallamcseréjével meg lehet valósítani a prozódiai változatosságot.

3.3.1. Hasonló mondatok keresése

A hasonló mondatok keresése a következő módon ment végbe: első célunk az volt, hogy olyan mondatokat találjunk, amiknek időtartama közel áll egymáshoz. Ezt viszonylag könnyű elérni, hiszen a beszédhang-adatbázisokban lévő hanghatárokat leíró fájlokból meg lehet tudni a megfelelő információt. Azonban a két mondat időtartamának egyformasága nem elég a két mondat hasonlóságához, mivel a beszédtempó[9] különbségei miatt egészen különböző mondatok is lehetnek hasonló időtartamúak.

Ennél jobb megoldáshoz vezet, ha azt a feltételt adjuk meg, hogy a keresett mondatok szótagszáma egyezzen. A legtökéletesebb az lenne, ha lennének olyan mondatok, amiknek megegyezik a szótagszáma, szavaik száma, és külön-külön a szavak szótagjainak száma is, azonban az adatbázisaink mérete ezt nem tette lehetővé, mert a mondatok meglehetősen eltérőek voltak.

Az 1. táblázat az időjárás előrejelzés hangkorpuszra vonatkozik. Látható, hogy az adatbázisban vannak egyforma szótagszámú mondatok, de olyan mondatpár nincs, amire teljesülne, hogy az egyes szavak szótagszáma is egyezzen. A 2031-2380 mondatpárt választottuk ki a további kísérletekhez, mert ezeknek az első két szava ugyanannyi szótagú, így feltételezhető, hogy a szóhangsúlyok is hasonló helyen vannak a mondatokban.

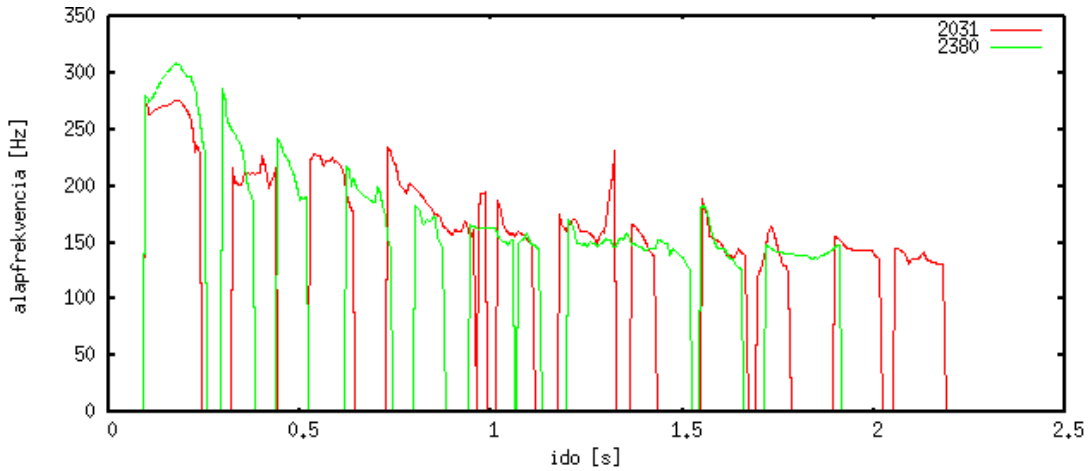
A célunk az volt, hogy össze tudjuk hasonlítani a két mondat dallammenetét. Ennek a megoldására egy egyszerű C programot készítettünk, ami az adatbázisban lévő pit (alapfrekvencia-értékek) és ssw fájlok (elemhatárok) alapján létrehoz egy Matlab/Octave szkriptet. A szkriptet Octave-ban lefuttatva kirajzolódik a két mondat alapfrekvencia-menete.

A 2. ábrán a két mondat egymásra rajzolt alapfekvenciamenete látható. A mondatok nincsenek szinkronban, és bár szótagszámuk megegyezik, hosszúságuk nem teljesen egyforma.

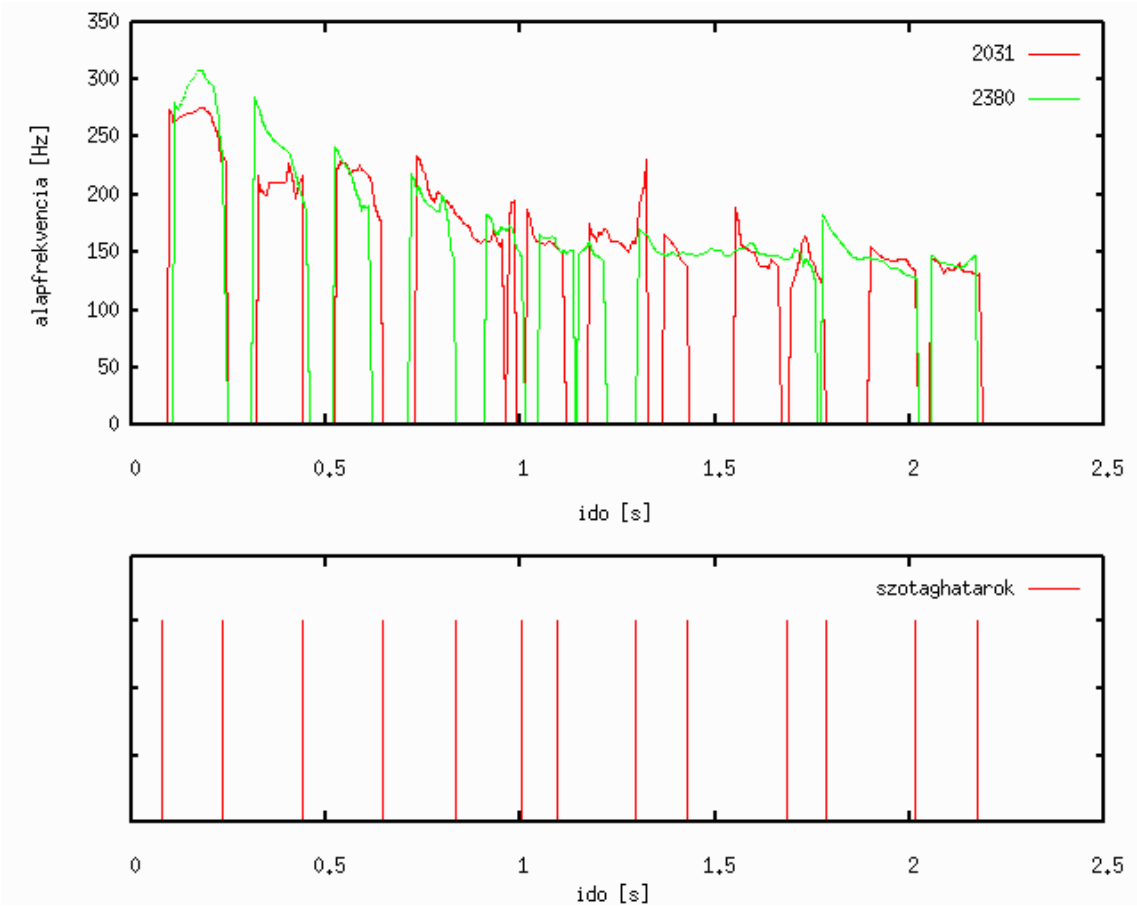
[9] A beszédtempó az időegységre jutó beszédesemények mennyiségét fejezi ki. Beszédesemény a hang kiejtése és a szünet tartása. ([1, 249. oldal] alapján)

mondat	szótagszám	szószám	szavak szótagszáma
2203	10	4	6 1 1 2
2563	10	4	3 4 2 1
3056	10	3	4 3 3
2019	11	5	1 1 3 3 3
2150	11	4	4 1 5 1
2565	11	4	3 2 2 4
3024	11	4	4 1 3 3
2031	12	4	3 4 2 3
2380	12	5	3 4 1 2 2
3031	12	7	3 1 1 1 3 1 2

1. táblázat. Mondatok szótagszám szerinti csoportosítása



2. ábra. A 2031-es és 2380-as mondat alapfrekvencia-menetének összehasonlítása időbeli vetemítés nélkül



3. ábra. Felső rész: a 2031-es és 2380-as mondat alapfrekvencia-menetének összehasonlítása időbeli vetemítéssel, alsó rész: a mondatok szótaghatárai

Fontos volt az egymásra rajzolt alapfrekvencia-menetek időbeli vetemítése, hogy a két mondat azonos részei egy helyen legyenek az ábrán. Ezt úgy tudtuk elérni, hogy az első mondatot referenciaként használva az ábrázolás során a második mondat hanghatárait megváltoztattuk úgy, hogy az azonos szótagok azonos helyen kezdődjenek és végződjenek, tehát az összeillesztést szótag szinten végeztük el. A második mondathoz tartozó alapfrekvencia-értékeket eközben nem változtattuk, így azok eltolódtak úgy, hogy az azonos szótagokhoz tartozó alapfrekvencia-menet nagyjából egy helyen legyen.

A 3. ábrán a két mondat egymásra rajzolt alapfrekvenciamenete látható időbeli vetemítéssel, az első mondatot használva referenciaként. Mivel a két mondat szótagjait egymásra toltuk, az alapfrekvencia-menetek is jobban illeszkednek egymásra. Az ábra alsó részén a referenciaként használt első mondat szótaghatárai jelennek meg, ami a vetemítés miatt megegyezik a második mondat szótaghatáraival.

A 2. és a 3. ábrán a 2031-es mondatban 1,4 mp-nél látható alapfrekvencia-csúcsot valószínűleg címkézési hiba okozta, mivel a mondatok meghallgatása során nem tapasztaltunk ilyen dallambeli ugrást. A szótagok szerinti vetemítés sikeresnek bizonyult, az ábrákon látható, hogy az azonos szótagokban a hasonló tartalom nagyjából azonos helyre esik minden esetben.

Miután megtaláltuk ezt a mondatpárt, további hasonló mondatokat kerestünk, hogy részle-

mondat	szótagszám	szószám	szavak szótagszáma
01_deszk_003	31	9	1 5 1 8 3 3 4 1 5
03_szurdokpuspoki_001	31	15	2 1 2 2 2 1 5 1 2 2 3 1 1 2 4
03_szurdokpuspoki_004	31	15	1 2 3 2 2 2 1 2 2 1 1 3 1 3 5

2. táblázat. Harangos beszédhang-adatbázis mondatainak szótagszám szerinti csoportosításának részlete

tesebb vizsgálatokat tudjunk végezni.

3.3.2. Hasonló mondatok dallammenetének összehasonlítása

A 3.3.1 részben leírtak szerint sikerült tehát ábrázolni egy mondatpár alapfrekvenciamenetét időbeli vetemítéssel úgy, hogy az azonos szótagokhoz tartozó értékek illeszkedjenek. Ezek után olyan további mondatokat kerestünk, amik jobban hasonlítanak egymásra a dallammenet szempontjából, vagyis különböző ábrákat vizsgálva próbáltuk megtalálni a kísérletekhez megfelelő mondathalmazokat.

A harangok története beszédhang-adatbázisban is megvizsgáltuk a mondatokat szótagszámuk szerint. A 2. táblázatban láthatóak az általunk kiválasztott és a további kísérletek során felhasznált 31 szótagból álló mondatok.

A két adatbázisban tehát több a céljainknak megfelelő mondathalmazt is találtunk, ezek:

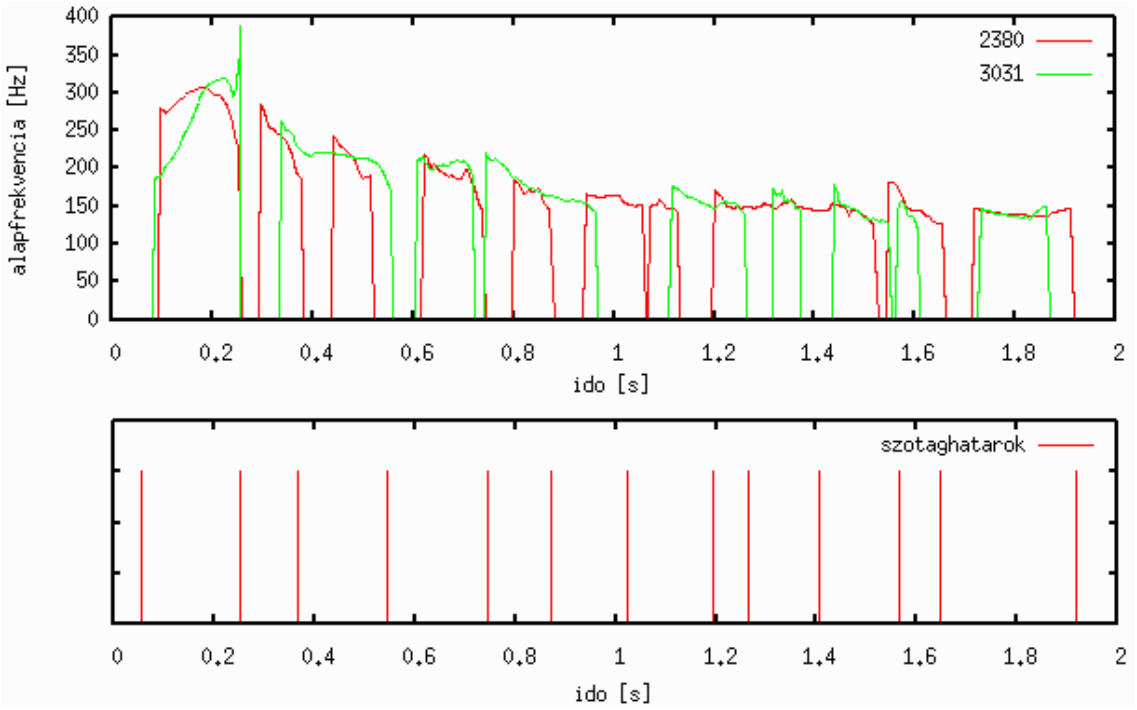
2031 - 2380 - 3031 : A mondatokat az időjárás előrejelzés hangkorpuszból válogattuk. A 2031 - 2380 párt már az első kísérlethez is kiválasztottuk. A 3031-es számú mondat szótagszáma szintén 12. A 4. ábrán a 3031-es mondat a 2380-assal összevetve látható. A két mondat dallammenete szinte teljesen megegyezik, csak az első szótagban van jelentősebb eltérés. A mondatok tartalma:

- 2031: *Elszórtan számíthatunk kisebb esőre.*
- 2380: *Péntekig folytatódik a meleg idő.*
- 3031: *Légköri front nem lesz fölöttünk kedd estig.*

2019 - 2150 : A mondatokat az időjárás előrejelzés hangkorpuszból válogattuk. A 2019 - 2150-es mondatok mindegyike 11 szótagból áll, így alkalmasak az összehasonlításra. Az 5. ábrán tekinthető meg az egymásra rajzolt dallammenetük, ami az első másodpercig jelentősen különböző. Mégis felhasználtuk a mondatpárt a kísérletekben, mert a mondatok dallamának többi része hasonló jellegű. A mondatok tartalma:

- 2019: *Ma nem várható közvetlen fronthatás.*
- 2150: *Megélénkül az északnyugati szél.*

01_deszk_003 - 03_szurdokpuspoki_001 - 03_szurdokpuspoki_004 : A mondatokat a harangok története hangkorpuszból válogattuk. Ezek kb. hat másodperc hosszú mondatok, amik 31 szótagból állnak. Mindhárom mondat alapfrekvenciája 60 és 120 Hz



4. ábra. Felső rész: a 2380-as és 3031-es mondat alapfrekvencia-menetének összehasonlítása időbeli vetemítéssel, alsó rész: a mondatok szótaghatárai

között ingadozik. Mivel kijelentő mondatokról van szó, mindegyiknek eső jellegű a dallammenete, és hasonlóságok fedezhetőek fel köztük. A mondatok tartalma:

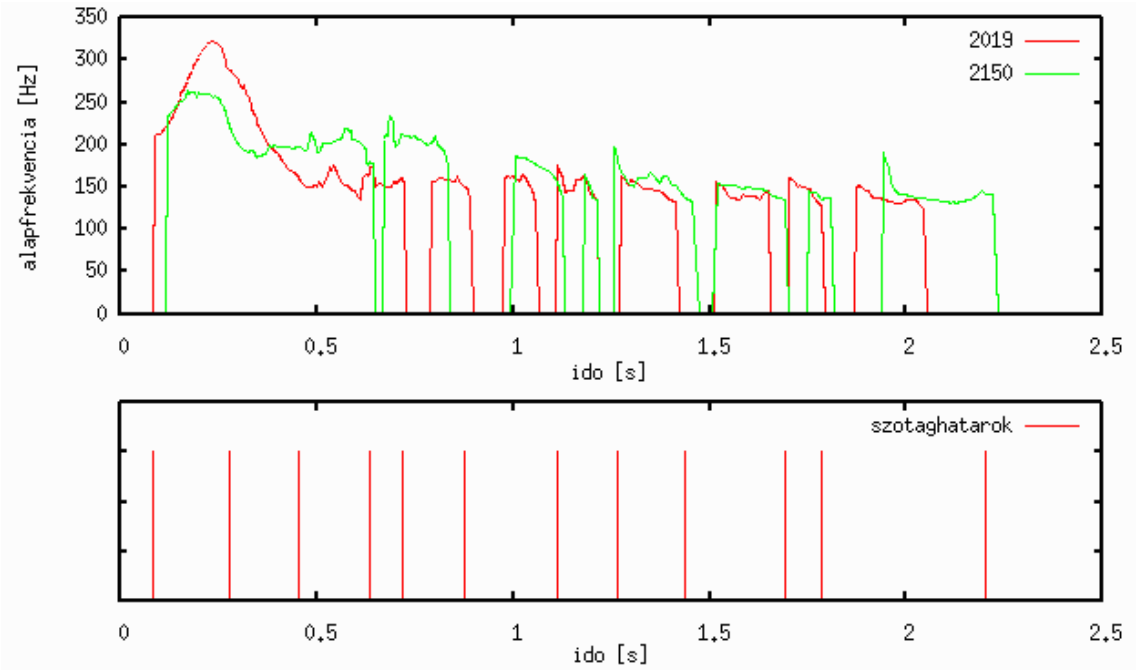
- 01_deszk_003: *A településről az ezernégyszázkilencvenes évektől találunk adatokat az oklevelekben.*
- 03_szurdokpuspoki_001: *Ezen a héten minden délben a szurdokpüspöki Szent Kereszt templom harangja szól a Kossuth rádióban.*
- 03_szurdokpuspoki_004: *A falu nevének első tagja utal a keskeny völgyre, míg a második az egykori tulajdonosra.*

3.4. Kiválasztott mondatok dallamcseréje

Az eddig ismert beszédszintetizátor rendszerekben egy-egy mondathoz mindig ugyanaz a prozódia tartozik. A rövidesen ismertetett kísérletekben megpróbáljuk megvalósítani a prozódiai változatosságot (konkrétan a dallam változatosságát).

Az az alapfeltételezésünk, hogy ha két hasonló mondat alapfrekvencia-menetét részben vagy teljesen lecseréljük a másikéra, akkor még mindig természetesnek hangzó beszédet kapunk. A kérdés az, hogy észlelhető-e egyáltalán ez a változás, és ha igen, akkor nem rontjuk-e el vele a mondat eredeti hangzását.

A kísérleteket természetes és szintetizált beszéden is elvégeztük.



5. ábra. Felső rész: a 2019-es és 2150-es mondat alapfrekvencia-menetének összehasonlítása időbeli vetemítéssel, alsó rész: a mondatok szótaghatárai

3.4.1. Természetes bemondás változtatása

Először természetes bemondásokon akartuk tesztelni, hogy a dallammenetbeli változások mennyire érezhetőek meghallgatás során. Ehhez a beszédhang-adatbázisainkban lévő hullámforma fájlokból indultunk ki, hiszen ezek természetes beszéd digitalizált változatai. A korábbiakban kiválasztott hasonló mondatok dallamcseréje a következőkben leírtak szerint történt. A mondatpár egyik tagja a referenciamondat, a másikon végeztük a változtatásokat. Először átlagoltuk a mondatok minden szótagjának alapfrekvenciáját. Ezt a szótagon belüli zöngés részeken értelmezett alapfrekvencia-értékek átlagaként számoltuk ki. Végül a két mondat szótagjainak átlagos alapfrekvenciájának különbségeivel eltoltuk a változtatandó mondat megfelelő szótagjának alapfrekvencia-értékeit.

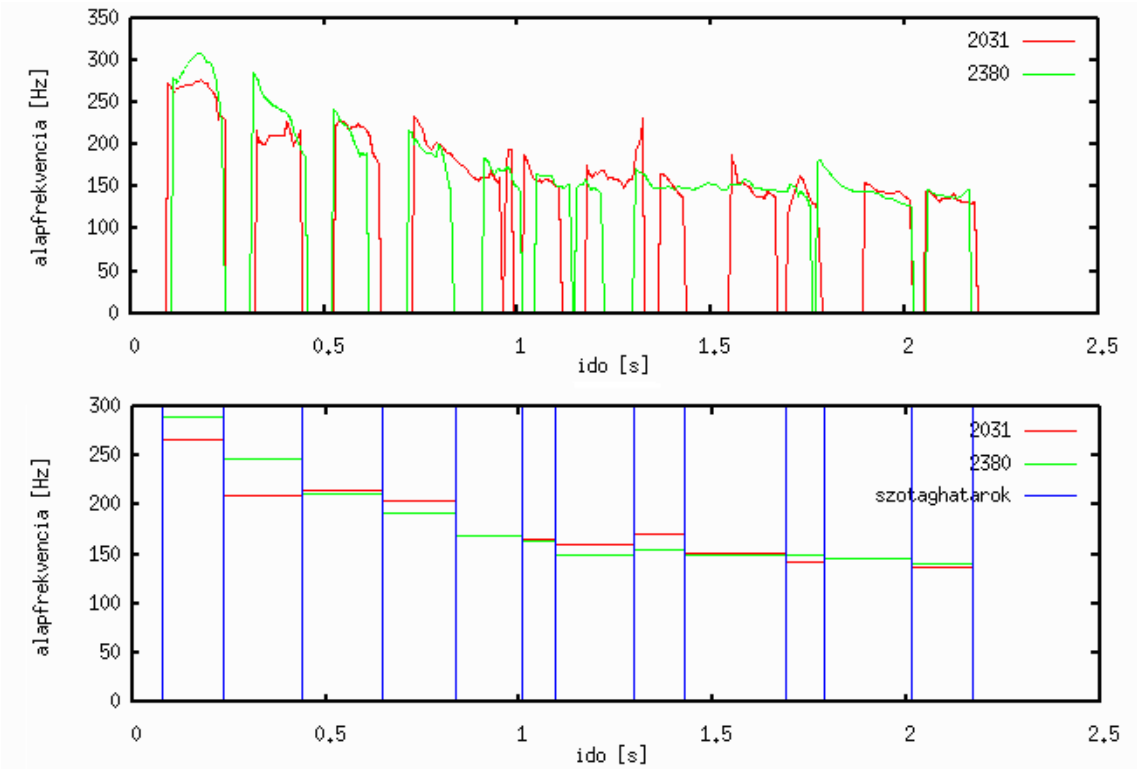
Első példa erre a 2031-2380-as mondatpár. A referenciamondat a 2031-es, a változtatandó a 2380-as. A mondatok szótagonkénti átlagos alapfrekvenciái, illetve ezek különbségei a 3. táblázatban láthatóak. A táblázatban észrevehetjük, hogy a különbségek minden esetben 40 Hz-nél kisebbek. Ez azért fontos, mert az alapfrekvencia módosítására alkalmazott TD-PSOLA eljárás (amit a 2.5.2 részben ismertettünk) akkor működik elfogadható minőségben, ha a megváltoztatott érték 30 %-os küszöbön belül van. Ez jelen esetben teljesül (hiszen a legnagyobb változtatás kb. 15%-os), a dallammenet módosítása tehát elvileg nem okoz rossz minőséget.

A 6. ábra a 3. ábrához hasonló, az eltérés csak annyi köztük, hogy a 6.-on már szerepelnek a szótagok átlagos alapfrekvenciái is.

Miután megvan, hogy mennyivel kell változtatni szótagonként az alapfekvenciát, a Praat program segítségével elvégezhető a dallammenetek eltolása. Eleinte kézzel végeztük ezeket

mondat	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
2031	266	210	215	205	169	164	160	171	151	141	145	137
2380	289	247	211	191	169	162	150	155	150	149	146	140
különbség	23	37	-4	-14	0	-2	-10	-16	-1	8	1	3

3. táblázat. A 2031 - 2380-as mondatok szótagjainak átlagos alapfrekvenciája és ezek különbsége Hz-ben megadva



6. ábra. Felső rész: a 2031-es és 2380-es mondat alapfrekvencia-menetének összehasonlítása időbeli vetemítéssel, alsó rész: a mondatok szótaghatárai és a szótagonkénti átlagos alapfrekvenciák

	2031	2380	3031
2031	x	2031_(2380_alapjan)	2031_(3031_alapjan)
2380	2380_(2031_alapjan)	x	2380_(3031_alapjan)
3031	3031_(2031_alapjan)	3031_(2380_alapjan)	x

4. táblázat. A 2031 - 2380 - 3031 mondathalmaz változatai. Az első sorban a referenciamondatok szerepelnek, az első oszlopban a változtatandó mondatok. A táblázat celláiban a referenciamondatok alapján megváltoztatott mondatok találhatóak.

	2019	2150
2019	x	2019_(2150_alapjan)
2150	2150_(2019_alapjan)	x

5. táblázat. A 2019 - 2150 mondathalmaz változatai. Az első sorban a referenciamondatok szerepelnek, az első oszlopban a változtatandó mondatok. A táblázat celláiban a referenciamondatok alapján megváltoztatott mondatok találhatóak.

az eltolásokat a program grafikus felületének segítségével, azonban ez rövidesen kényelmetlenné vált. Ennek a problémának a megoldására készítettünk egy Praat szkriptet[10], amivel már csoportosan, több hullámforma fájlon is meg lehet valósítani a dallammenet módosítását.

Ezeket a kísérleteket elvégeztük a többi hasonló mondathalmazban is. Így tehát keletkeztek olyan hullámforma fájljaink, amikben a beszéd dallammenete változott meg egy referenciamondat alapján. A létrehozott fájlokat a 4., 5. és 6. táblázatban foglaljuk össze. A táblázatokban az "x"-ek azt jelentik, hogy minden mondathoz csak másikat használunk referenciaként, önmagát nem. A 2031-es, 2380-as és 3031-es mondatoknak így két-két változata, a 2019-es és 2150-es mondatoknak egy-egy változata született az eredetin kívül. A 01_deszk_003, 03_szurdokpuspoki_001, 03_szurdokpuspoki_004 mondatok mindegyikéből szintén két új változat készült.

Az így létrejött mondatok előnye, hogy a természetességüket megőrzik, hiszen a kiindulási alap felvett beszéd volt. A dallammenetbeli módosításokat könnyű volt elvégezni. Az eredményt meghallgatva eldönthető, hogy sikeres volt-e a kísérlet, egy ilyen teszt bemutatására a 4. fejezetben kerül sor. Ez a kísérletsorozat csak a természetes beszéd módosíthatóságát vizsgálta, a szintetizált beszéd prozódiájának módosítása a következő részben olvasható.

3.4.2. Szintetizált mondatok változtatása

A természetes mondatok prozódiájának módosítása után azért tértünk át a szintetizált mondatok vizsgálatára, mert a megfelelő szegmentális szerkezet így jobban biztosítható. A szegmentális szerkezet a beszélő akaratától független paramétereket jelenti (pl. specifikus időtartamok és intenzitások).

Az itt leírt kísérletek során a 2.5.1. részben ismertetett diádos és triádos elemekkel rendel-

[10] A Praat programhoz készíthető kötegelt adatfeldolgozást segítő szkript. Ennek lényege, hogy ugyanazokat az utasításokat végzi el, amit a felhasználó végezne a grafikus felületen, de a bemenet egy szkriptfájl, és nem egérkattintások.

	01_deszk_003
01_deszk_003	x
03_szurdokpuspoki_001	03_szurdokpuspoki_001_(01_deszk_003_alapjan)
03_szurdokpuspoki_004	03_szurdokpuspoki_004_(01_deszk_003_alapjan)

	03_szurdokpuspoki_001
01_deszk_003	01_deszk_003_(03_szurdokpuspoki_001_alapjan)
03_szurdokpuspoki_001	x
03_szurdokpuspoki_004	03_szurdokpuspoki_004_(03_szurdokpuspoki_001_alapjan)

	03_szurdokpuspoki_004
01_deszk_003	01_deszk_003_(03_szurdokpuspoki_004_alapjan)
03_szurdokpuspoki_001	03_szurdokpuspoki_001_(03_szurdokpuspoki_004_alapjan)
03_szurdokpuspoki_004	x

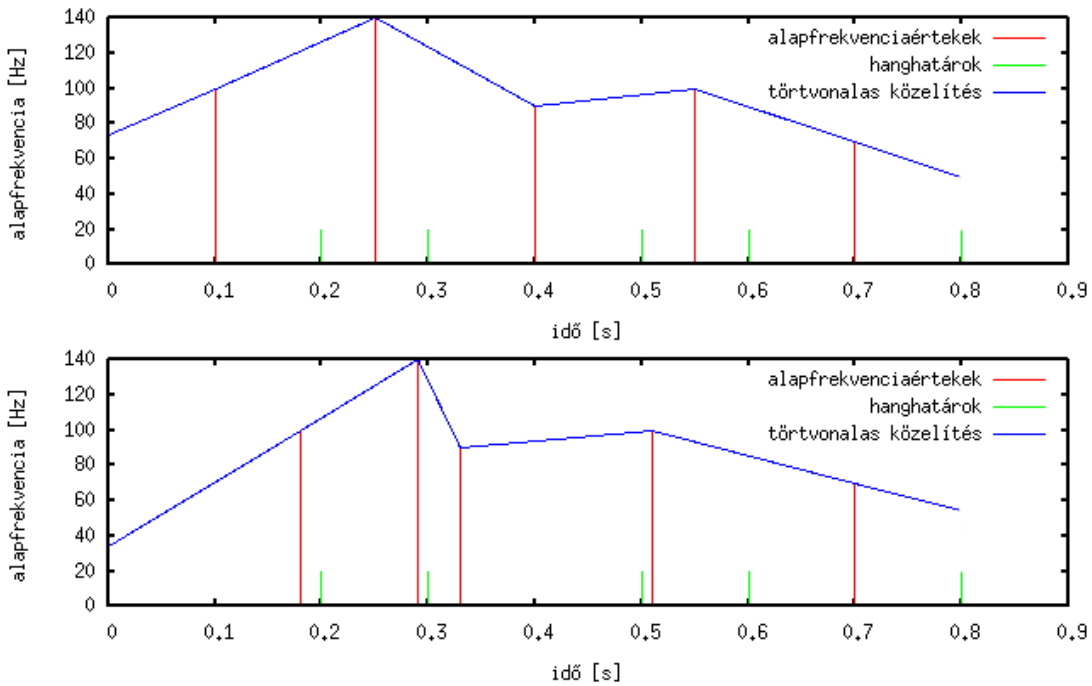
6. táblázat. A 01_deszk_003 - 03_szurdokpuspoki_001 - 03_szurdokpuspoki_004 mondathalmaz változatai három kisebb táblázatban. Az első sorokban a referenciamondatok szerepelnek, az első oszlopokban a változtatandó mondatok. A táblázat celláiban a referenciamondatok alapján megváltoztatott mondatok találhatóak.

kező Profivox beszédszintetizátort használtuk, amit Olaszy és társai fejlesztettek ki [14]. A Profivox szintetizátor többféle bemenettel használható: ha csak szöveges bemenetet adunk meg, maga a szintetizátor határozza meg az alapfrekvencia értékeket, a hangok időtartamát és az intenzitást. Egy köztes fájl segítségével lehetőség van azonban arra, hogy az ebben meghatározott értékekkel készüljön el a szintetizált mondat. Ez a köztes fájl egy intonációs mátrixot tartalmaz, amiben a hangok kódjai, valamint ezeknek alapfrekvencia, intenzitás és időtartam értékei definiáltak.

Legelőször az adatbázisban lévő szöveges fájlok alapján készítettünk szintetizált mondatokat, hogy későbbi tesztjeink során ezeket össze tudjuk hasonlítani a változatos dallamú mondatokkal.

A különböző dallammenetű szintetizált mondatok létrehozásához először a szöveges fájlokat át kellett alakítani a beszédszintetizátor köztes formátumába. Mivel célunk az volt, hogy csak az alapfekvencia menetét változtassuk, ezért egy a Szabó munkájában [17] szereplő Praat szkript (*Sound2Imf.praat*) segítségével a szöveges fájlokból, és a hanghatárokat tartalmazó TextGrid fájlokból létrehoztuk a szintetizátor számára értelmezhető köztes fájlokat. Az így létrejött fájlokban tehát az egyes hangok, időtartamaik és intenzitásaik már be voltak állítva, amiket a későbbiekben sem módosítottunk.

A Profivox szintetizátorban az alapfekvencia úgy jön létre, hogy minden hanghoz megadhatunk egy százalékos értéket, hogy az alaphanghoz (pl. 110 Hz) képest mennyire legyen magas vagy mély az adott hang. Ez az alaphang a szintetizált beszéd létrehozása előtt kerül beállításra. Megadhatjuk azt is, hogy az adott hangban hol kell elérni az előírt frekvencia értéket (szintén százalékos formában, a hang hosszához képest). Ezután a beszédszintetizátor a mondat alapfrekvencia-menetét az általunk kijelölt pontok (időpont és alapfrekvenciaértékek) lineáris interpolációjával hozza létre. Az alapfrekvencia csak zöngés hangokra értelmezett, de zöngétlen hangokhoz is írhatunk értékeket, így az interpolációt pontosabbá lehet tenni,



7. ábra. A Profivox intonációs mátrixában egy hang alapfrekvenciájának beállítása során meg kell adnunk azt is, hogy az adott hangban hol kell elérni ezt az értéket. Az ábra felső részén ez minden hangra 50 %, míg az alsó részen különbözőek a százalékos értékek.

hiszen több pont van megadva. A dallammenet egyenes szakaszokból áll elő, amiben a töréspontok a beállított értékek. A 7. ábrán az látható, hogyan lehet azonos hangidőtartamok és alapfrekvencia-értékek mellett különböző dallammenetet létrehozni az adott alapfrekvenciák hangon belüli helyének eltolásával.

Első kísérletként azt próbáltuk elérni a 2031-es számú mondattal, hogy a természetes bemondást utánozva hozzunk létre szintetizált mondatot. A *Sound2Imf.praat* szkript segítségével elkészítettük a hangokat, intenzitásokat és időtartamokat tartalmazó intonációs mátrixot. Ezután megvizsgáltuk a természetes bemondás dallammenetét, és kézzel átírtuk az intonációs mátrixba az egyes hangokhoz tartozó relatív alapfrekvencia értékeket és ezek helyét. Az adott alapfrekvencia értékek hangon belüli helyét úgy állítottuk be, hogy minél pontosabban közelítse az eredeti dallammenetet. Azonban az alapfrekvenciák helyének beállítása egy-egy hangon belül meglehetősen bonyolult volt, és nehéz lett volna automatikus módszerekkel létrehozni. A továbbiakban ezért nem kézzel állítottuk be az alapfrekvencia paramétereket, hanem készítettünk egy C programot a feladat elvégzésére. A program az beszédhang-adatbázisban lévő zöngeperiódushatárokat tartalmazó fájlból indul ki, ebbben megkeresi minden hang átlagos alapfrekvenciáját, és beírja azt az intonációs mátrixba. Az intonációs mátrixban az alapfrekvenciák helyét egységesen 50 %-ra állítottuk, ahogy ez a 7. ábra felső részén is látható. Úgy tapasztaltuk, hogy a kézi és az automatikus átírás között nem hallható lényeges különbség, ezért a továbbiakban csak az utóbbit alkalmaztuk.

Miután sikerült megoldani, hogy a mondatokat a saját dallammenetükkel szintetizáljuk, már az is megoldhatóvá vált, hogy egy másik mondat alapján állítsuk be az alapfrekvencia-menetet.

A 3.3. alfejezetben megtalált mondathalmazok mindegyik mondatát létrehoztuk a halmazban lévő többi mondat dallammenetével. Így hasonló fájlokhoz jutottunk, mint amit a 4., 5. és 6. táblázatban bemutattuk, azzal a különbséggel, hogy ezek a hullámforma fájlok szintetizált beszédet tartalmaznak.

Sikerült tehát előállítani egy-egy mondat többfajta szintetizált változatát. Például a 2031-es mondat variánsai:

2031_profivox : a 2031-es mondat szöveges változatából a Profivox alapbeállításaival készült hullámforma

2031_kezzel : a 2031-es mondat *Sound2Imf.praat*-tal elkészített intonációs mátrixában az alapfrekvencia-menet kézzel történő beállítása után Profivox-szal szintetizált hullámforma, amelynek dallammenete az eredeti 2031-es mondaté

2031_geppel : a 2031-es mondat *Sound2Imf.praat*-tal elkészített intonációs mátrixában az alapfrekvencia-menet automatikus módszerrel történő beállítása után Profivox-szal szintetizált hullámforma, amelynek dallammenete az eredeti 2031-es mondaté

2031_(2380_alapjan) : a 2031-es mondat *Sound2Imf.praat*-tal elkészített intonációs mátrixában az alapfrekvencia-menet automatikus módszerrel történő beállítása után Profivox-szal szintetizált hullámforma, amelynek dallammenete a 2380-as mondaté

2031_(3031_alapjan) : a 2031-es mondat *Sound2Imf.praat*-tal elkészített intonációs mátrixában az alapfrekvencia-menet automatikus módszerrel történő beállítása után Profivox-szal szintetizált hullámforma, amelynek dallammenete a 3031-es modaté

A többi mondathalmazban a kézzel történő beállítással szintetizált hullámforma már nem szerepel, de így is elő tudtuk állítani minden mondatnak legalább három változatát: Profivox alapbeállításokkal, a mondat saját dallammenete alapján, illetve egy másik mondat alapfrekvencia-menete alapján.

4. Teszt-eredmények

A fejezet az elvégzett tesztek tervét, megvalósítását, eredményeit és értékelését tartalmazza.

A 3.4.1 és 3.4.2 részekben sikerült előállítanunk minden vizsgált mondatnak több változatát. Ezekből 28 mondatpárt összeválogatva létrehoztunk egy tesztet. A tesztben azonos mondatok különböző módon előállított változatait kell összehasonlítani egy-egy mondatpárban. A két mondat meghallgatása után el kell dönteni, hogy az 1. vagy a 2. jobb minőségű, vagy esetleg egyformának tűnnek a mondatok. A teszt célja az volt, hogy megtudjuk, mennyire voltak sikeresek a kutatás során a beszédhang-adatbázisok mondatain végzett prozódiai módosítások.

A teszteket a BME-TMIT-en kifejlesztett webes tesztelő rendszerben végeztük: `http://ss20.tmit.bme.hu/dallamteszt`. A honlapra ellátogatva a tesztelőnek el kell olvasnia egy rövid ismertetőt a teszt jellegéről. Eközben a teszthang meghallgatásával meggyőződhet róla, hogy minden megfelelő-e a teszt tényleges elvégzéséhez.

Először néhány adatot kérünk be a tesztelőtől (becenév, életkor és nem), majd ezután indul az éles teszt a 28 mondatpár meghallgatásával, ami kb. 10 percig tart. Végül a személyes megjegyzést is lehet leírni a teszttel kapcsolatban.

A tesztet 2006. október 18. és 2006. október 25. között 23-an végezték el, ami az eredmények kiértékeléséhez nagyjából elegendő. A tesztelők a 20-30 éves korosztály tagjai, mindannyian ép hallású, magyar anyanyelvű emberek voltak. Egy részük a BME-TMIT-en beszédkutatással fogalkozó munkatárs, akik a témához értvén valószínűleg más szempontok szerint osztályozták a mondatpárokat, mint a tesztelők másik része, akik kevésbé jártasak ebben a témakörben.

A teszt elvégzése után írt megjegyzésekből kiderült, hogy egyesek szerint a kérdést pontosabban kellett volna feltenni. Problémát jelentett ugyanis számukra, hogy nem tudták eldönteni, milyen szempontok szerint kellett volna a minőséget figyelniük (hangsúlyozás, érthetőség, hangok elcsúszása).

Az alábbiakban összefoglaljuk az elvégzett tesztek eredményeit. Összesen hét különböző mondat szerepelt a tesztben, mindegyik több dallampárban is előfordult. Ezek a mondatok:

Az időjárás előrejelzés hangkorpuszból:

- 2019, 5 mondatpár
- 2031, 5 mondatpár
- 2380, 4 mondatpár
- 2565, 4 mondatpár

A harangok története hangkorpuszból:

- 01_deszk_003, 3 mondatpár
- 03_szurdokpuspoki_001, 4 mondatpár
- 03_szurdokpuspoki_004, 4 mondatpár

1. mondat 2. mondat	szint., 2380 term., eredeti	term., 2380 term., 3031	szint., 2380 szint., kézzel	szint., géppel szint., profivox	term., 3031 term., 3031 /2
1. jobb	0	4	2	11	14
2. jobb	22	11	11	6	4
egyforma	1	8	10	6	5

7. táblázat. A 2031-es mondatváltozatok összehasonlításának eredménye.

A mondatpárokban találhatóak természetes bemondások, természetes bemondások egy hasonló mondat alapján módosítva, szintetizált mondatok a Profivox alapbeállításaival, szintetizált mondatok saját dallammenetükkel (kézzel másolva), szintetizált mondatok saját dallammenetükkel (automatikusan másolva) és szintetizált mondatok egy hasonló mondat dallammenetével.

A 7. táblázatban példaként a 2031-es mondat különböző változataival elért teszteredmények láthatóak. A felső részben található, hogy a 2031-es mondatnak melyik két változata van a mondatpárban. A természetes mondatot *term.*-mel jelöltük, *szint.*-tel a szintetizáltat. Az ez után következő szó a dallam módosítására utal: *2380, 3031* azt jelenti, hogy a dallammenet ezen mondat alapján lett származtatva, a *kézzel* készült változatot a mondat eredetijéből hoztuk létre kézi átírással, a *géppel* készült változatot a mondat eredetijéből hoztuk létre automatikus átírással, a *profivox* pedig arra utal, hogy a Profivox alapbeállításaival szintetizáltuk a mondatot. A táblázat celláiban az látható, hogy a tesztelők hogyan értékelték a meghallgatott mondatokat.

Az első esetben az eredeti, természetes bemondást hasonlítottuk össze egy módosított dallamú szintetizált változattal. Látszik, hogy a természetes bemondás egyértelműen jobb, bár ez várható is volt, hogy ennél jobbat nem tudunk létrehozni. A második eset két dallamában módosított természetes bemondást hasonlít össze. A tesztelők itt már nem tudtak egyértelműen dönteni. Ebből az következik, hogy a dallambeli módosítás érzékelhető, de ugyanakkor egyik változat sem sokkal jobb a másiknál. A harmadik mondatpár két szintetizált mondatból áll. A kézzel létrehozott dallammenet kicsit természetesebbnak bizonyult (ez várható is volt, hiszen ez a lehető legjobb általunk Profivox-szal megvalósított minőség), ugyanakkor itt is látszik, hogy a módosított alapfrekvencia-menettel rendelkező mondat sem sokkal rosszabb dallamú. A negyedik teszteset szintetizált mondatokban hasonlítja össze az automatikus módszerrel másolt eredeti dallammenetű változatot a Profivox alapbeállításaival létrehozottal. Az alapbeállításokkal készült mondat itt alulmaradt, mivel a másolt dallammenetű mondat jobban közelíti a természetes beszédet. Végül a 2031-es mondat utolsó tesztjében olyan mondatpárt kellett összehasonlítani, amelyek mindegyike a természetes bemondást manipulálva jött létre. A két mondatban ugyanazt, a 3031-es mondatot használtuk a dallam referenciájaként, de a másodikban kicsit módosítottuk a dallam másolásának paramétereit, ami nem bizonyult jónak a tesztelők véleménye alapján.

A többi mondatváltozat tesztjének eredményét már nem elemezzük ilyen részletesen, csak az eddig leírtaktól különböző észrevételeket soroljuk fel.

A 2019-es mondat variánsai közül kitűnik, hogy az a szintetizált mondat, ami a 2150-es dallammenete alapján készült, kimondottan rossz tulajdonságokkal rendelkezik, a tesztelők egyértelműen gyenge minőségűnek ítélték. Ebből látszik, hogy egyáltalán nem mindegy, hogy

a dallammenetek átültetése során mennyire közel álló mondatokat vizsgálunk. A 18. oldalon az 5. ábra elemzésekor meg is említettük, hogy a 2019-es és 2150-es mondat eleje nagy mértékben különbözik. Most látható, hogy ez a különbség nem teszi lehetővé, hogy a dallamcsere során jó minőségű mondatok keletkezzenek.

A 2380-as mondatban a 2031-eshez hasonlóan érvényesült az, hogy a két különböző mondatról (2031-es és 3031-es) másolt dallammenetű változatok hasonlóak minőségi szempontokból. Ez a természetes bemondás dallamcseréjére és a szintetizált változatokra is igaz.

A 2565-ös mondat szótagszám szempontjából a 2019-esre hasonlított, az alapfrekvencia-menetét is arról másoltuk. Így a 2019-es mondat változatainál ismertetett problémák itt is előfordulnak.

A harangos beszédkorpuszban kissé más eredmények születtek, mivel ezen adatbázis mondatainak jellege is más volt. A három kiválasztott mondat 31 szótagú, ami sokkal hosszabb, mint az időjárás előrejelzés hangkorpuszból kiválasztott mondatok 11 illetve 12 szótagja.

A 01_deszk_004 mondat dallammásolással létrehozott variánsait a tesztelők hasonlónak ítélték, de ez valószínűleg azért van így, mert az eredeti, természetes bemondáshoz képest sem érezhető jelentősebb különbség. A jelenség abból fakad, hogy a beszédhang-adatbázis mondatait felolvasó férfihang meglehetősen monoton, és a különböző bemondások között nem sok dallambeli különbség mutatható ki. Az sem vezet tehát kielégítő eredményre a kutatásunk szempontjából, ha olyan mondatok alapfrekvencia-menetét próbáljuk megcserélni, amik eredetileg is nagyon közel álltak egymáshoz.

A 01_szurdokpuspoki_001 és 01_szurdokpuspoki_004 mondatokban a tesztelők a Profivox alapbeállításai szerint létrejött változásokat jobbnak ítélték, mint azokat a variánsokat, amiket másik mondat alapfrekvencia-menete alapján szintetizáltunk. Ez azért fordulhatott elő, mert a hosszú mondatokban az egyes szavak meglehetősen eltérő hosszúságúak, így a szóhangsúlyok is más-más helyen vannak, amik az alapfrekvencia-menet másolásakor rossz helyre kerülnek. A természetes változatok módosításánál ez nem volt jelentős különbség, de a szintetizált variánsokban jobban hallható.

Az eredmények kiértékelése alapján tehát felállíthatunk egy minőségi sorrendet a mondatok különböző variációi között, a természetes és a szintetizált variánsokat külön tárgyalva.

Az eredeti bemondás a legjobb minőségű, ehhez képest gyengébb a természetes, másik mondat dallama alapján módosított változat. Különböző mondatok dallammenetét másolva a minőség hasonló lesz. A szintetizált mondatok közül a saját dallammenete alapján létrehozott a legjobb, a másik mondatok dallammenete alapján készült változat kevésbé kellemes hangzású, a Profivox alapbeállításokkal szintetizált variáns pedig a leggyengébb minőségű.

Összességében elmondhatjuk, hogy az esetek egy részében sikerült megvalósítanunk a prozódiai változatosságot. Egy-egy mondatot többféle prozódiával elő tudtunk állítani úgy, hogy a minőségük nem különbözött jelentősen. Ugyanakkor az is szembetűnő, hogy ha nem megfelelően választjuk meg a hasonló dallammenetűnek ítélt mondatok halmazait, akkor rossz eredményekre jutunk. Nem jó, ha a kiválasztott halmaz mondatainak dallama túlságosan eltér egymástól, de az sem vezet kielégítő eredményre, ha monoton mondatokról van szó, amiknek dallama túl közel van egymáshoz. Az eddigiek során nem találtunk olyan algoritmust, amivel egyértelműen meghatározható lenne, hogy melyik mondatok felelnek meg a dallamcserére.

5. Felhasználási, továbbfejlesztési lehetőségek

A beszéd kutatása nagyon sok tudományágat foglal magába, ezen belül a beszédszintézis rendszerek tanulmányozása is rendkívül kiterjedt témájú. A prozódiai modellek tanulmányozása során világossá vált számunkra, hogy milyen sokféle módszerrel próbálják meg minél jobb minőségű beszédszintetizátor létrehozását. A jobb minőséget dolgozatunkban a prozódia változtatásával próbáltuk elérni.

A tesztek eredményének kiértékelése során kiderült, hogy a dolgozatban bemutatott kutatást sokféle irányban lehet folytatni. Először is szükség lenne olyan algoritmusok konstruálására, amikkel egy adott beszédkorpuszban optimálisan megtalálhatjuk az egymáshoz hasonlító mondatokat. A „hasonlóság" fogalmat megfelelően definiálni kellene. Optimalizálásra azért van szükség, mert a 4. fejezetben leírtak szerint olyan mondatokat keresünk, amiknek dallammenete elég közel áll egymáshoz abból a szempontból, hogy a mondatok közötti alapfrekvencia-menet csere ne okozzon hirtelen ugrásokat a dallamban. Ugyanakkor az sem jó, ha a mondatok túl egyformák, mert akkor dallamjuk cseréjével nem hozható létre érzékelhető változás. Az ideális megoldáshoz úgy lehetne közelíteni, ha megfelelően nagy beszédadatbázis állna rendelkezésre. Ekkor ugyanis nagyobb az esélye annak, hogy találunk olyan mondatokat, amik megfelelnek a hasonlósági követelményeinknek. A 3.3.1 szakaszban megállapítottuk, hogy a kísérleteinkben felhasznált mondatok között nem találtunk olyanokat, amik az egyforma szótagszám mellett az egyes szavaik szótagjainak számában is megegyeztek volna. Tegyük fel ugyanis, hogy a beszédkorpuszunkban van két mondat, aminek ezek a tulajdonságai megegyeznek:

1. *Ma reggel esni fog az eső.*

2. *Ma este esni fog az eső.*

Ha például egy hasonló szerkezetű 3. mondatot (*Ott délben jönni fog a felhő.*) szeretnénk szintetizálni, akkor az 1. és a 2. mondatok bármelyikét használhatnánk prozódiai mintaként.

A további kutatások során ezt a feltételezést bizonyítani, ellenőrizni kellene és minél szélesebb körben (lehetőleg az egész magyar nyelvre) kidolgozni.

Ha a jelenlegi dolgozatban körvonalazott kérdések pontosabb kidolgozása megtörténik, lehetőség van arra, hogy a Profivox magyar nyelvű beszédszintetizátorban alkalmazásra kerüljön egy, a prozódia változatosságáért felelős modul. Ez a beszédszintézis során véletlenszerűen választana több lehetséges prozódia közül, így megvalósítva a természetesebbé tett beszédet.

6. Összefoglalás

A fejezetben összefoglaljuk az eddig elvégzett munkánkat a szakirodalom feldolgozásától kezdve a kísérletekig és tesztekig.

Munkánk célja az volt, hogy a beszédszintézis során létrejövő beszédet természetesebbé tegyük. A természetesség a dolgozat folyamán a prozódia változatosabbá tételét jelentette.

A munkát a szakirodalom áttekintésével kezdtük. Részleteztük a prozódia legfontosabb összetevőit, majd a beszédszintetizátorok különböző változatait ismertettük. Az irodalomban megtalálható prozódiai modelleket többféle szempontból csoportosítottuk. Egy konkrét példa során megvizsgáltuk a magyar nyelv kijelentő mondatainak prozódiai változatosságát. Végül a dallammenet módosításának elvét mutattuk be.

A kísérleti munka az alapvető cél definiálásával kezdődött. Miután ez megtörtént, megfelelő beszéddallam-adatbázisokat kerestünk, illetve hoztunk létre. Ezután elemeztük az adatbázisok mondatait, és különböző szempontok szerint hasonlóságokat kerestünk bennük. Kiválasztottunk néhány mondathalmazt, amelyek hasonló jellegű bemondásokat tartalmaztak. Megvalósítottuk a kiválasztott mondatok prozódiájának változatosabbá tételét dallammenetük módosításával.

A munka lezárásaként egy tesztet hoztunk létre a módosított mondatok halmazából, amelyet értékeltünk is, miután elegendő tesztelő elvégezte azt. A tesztek kiértékelése során megvizsgáltuk a felmerült problémákat.

Végül a továbbfejlesztési lehetőségeket is áttekintettük.

7. Köszönetnyilvánítás

Ezúton mondok köszönetet Dr. Németh Géza és Dr. Fék Márk konzulenseimnek a munkám során nyújtott segítségükért, észrevételeikért és tanácsaikért, Zainkó Csabának a Harangok története beszédhang-adatbázis rendelkezésemre bocsátásáért, és Bartalis István Mátyásnak a webes tesztelő rendszer beállításáért. A munka a BME-TMIT Beszédtechnológiai Laboratóriumában készült.

8. Irodalomjegyzék

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

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

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

[4] Greg Kochanski and Chilin Shih, *Prosody Modeling with Soft Templates*, Bell Laboratories, Lucent Technologies Technical Report, 2001, pp. 3–5., http://prosodies.org/papers/SpeechComm1_2001.pdf

[5] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J., „TOBI: A standard for labelling English prosody", in *ICSLP'92. Proceedings of the Second International Conference on Spoken Language Processing.* Banff, October 1992. pp. 867–870., http://www.ling.ohio-state.edu/~tobi/ame_tobi/ /Silverman_etal1992.pdf

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

[7] Stefan Baumann, Martine Grice and Ralf Benzmüller, „GToBi - a phonological system for the transcription of German intonation", in *Prosody 2000: Speech Recognition and Synthesis*, Krakow, Poland. Adam Mickiewicz University, 2000., pp. 21–28.

[8] Maria E. Graba, Brechtje Post and William F. Nolan, „Modelling intonational variation in English: The IViE system" in *Prosody 2000: Speech Recognition and Synthesis*, Krakow, Poland. Adam Mickiewicz University, 2000., pp. 51–57.

[9] Volker Strom, „From text to prosody without ToBi", in *In Proceedings [ICSLP-2002] 7th International Conference on Spoken Language Processing (ICSLP2002 - INTERSPEECH 2002)*, International Speech Communication Association, Denver, 2002 http://www.cstr.ed.ac.uk/downloads/publications/2002/ /paper.icslp02.pdf

[10] Jianhua Tao, Lianhong Cai, Herbert Tropf, „An optimized neural network based prosody model of chinese speech synthesis system", *The 17th IEEE Region 10 International Conference on Computers, Communications, Control and Power Engineering*, Beijing, 2002, http://hcsi.cs.tsinghua.edu.cn/Paper/Paper02/200214.pdf

[11] Olaszy, G., Németh, G., Olaszi, P., „Automatic Prosody Generation - a Model for Hungarian”, *Proc. of the Eurospeech 2001*, Vol. 1., pp. 525–528.

[12] Németh Géza, Olaszy Gábor, *Beszédinformációs rendszerek tantárgy előadás anyaga*, 2. fejezet, 2005., pp. 17–18. `http://speechlab.ttt.bme.hu/postnuke/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=92`

[13] Olaszy Gábor, „A korpusz alapú beszédszintézis nyelvi, fonetikai kérdései”, *Híradástechnika*, Vol. LXI., no. 3, pp. 43–50, 2006.

[14] Gábor Olaszy, Géza Kiss and Géza Németh, „Hungarian audiovisual prosody composer and TTS development environment”, in *Prosody 2000: Speech Recognition and Synthesis*, Krakow, Poland. Adam Mickiewicz University, 2000., pp. 167–177.

[15] Mihajlik P., Révész T., Tatai P., „Phonetic Transcription in Automatic Speech Recognition”, *Acta Linguistica Hungarica*, Vol. 49. (3-4), 2002., pp. 407–425.

[16] Boersma, Paul & Weenink, David, *Praat: doing phonetics by computer*, (Version 4.4.34) [Computer program], 2006, `http://www.praat.org/`

[17] Szabó János, *Érzelmi beszédadatbázis feldolgozása és összehasonlító elemzése*, Budapesti Műszaki és Gazdaságtudományi Egyetem, Budapest, pp. 16–24., 56–63., 2006.