Steigerung der Natürlichkeit synthetisierter Sprache

TDK Arbeit

Anfertigt von:

Csapó Tamás Gábor

csapszi@sch.bme.hu

Konsulent:

Clemens Prinz

clemens.prinz@gmx.at

Oktober 2007

Inhaltsverzeichnis

1. Auszug

Heutzutage spielt die Mensch-Maschine-Beziehung, dabei auch die Sprachsynthese eine immer wichtigere Rolle. Sprachsynthese nennt man die Erzeugung von gesprochener Sprache durch einen Computer. Eine der Schlüsselfragen der Sprachtechnologie ist die Realisierung von entsprechender, gleichzeitig abwechslungsreicher Prosodie (Satzmelodie, Betonung, Sprechrhythmus). Heute ist es nicht mehr genug, mit einem Regelsystem zu einem festgelegten Input-Text immer die gleiche Prosodie zu fügen. Das Ziel ist, der menschlichen Stimme nahe zu kommen. In dieser Arbeit wird erstens eine Übersicht über die zum Thema gehörende Fachliteratur gegeben, danach wird eine Methode dargelegt, wie man abwechslungsreiche Prosodie erzeugen kann. Ein solches Text-To-Speech-System kann bei zahlreichen praktischen Anwendungen verwendet werden, wie z.B. SMS-, E-Mail-, Buchvorlesautomat. Abwechslungsreiche Satzmelodie ist hauptsächlich bei der Vorlesung von längeren Texten vorteilhaft, weil dabei die Monotonie des TTS-Systems störend wäre.

2. Theoretischer Hintergrund

In diesem Abschnitt wird ein Überblick über die Grundbegriffe, die zum Verständnis der Arbeit notwendig sind, gegeben. In Unterabschnitt 2.1. wird auf die Prosodie näher eingegangen. Der Unterabschnitt 2.2. stellt den Aufbau der Vorlesautomaten dar. Im letzten Unterabschnitt wird die Generierung von Prosodie durch die Maschine genauer behandelt.

2.1. Komponenten der Prosodie

Prosodie (Satzmelodie, Betonung, Sprechrhythmus [1]) ist eine ganz wichtige Eigenschaft der Sprache. Damit werden verschiedene Gefühle ausgedrückt, die nicht konkret im geschriebenen Text enthalten sind. Natürlich beeinflusst der Text auch den Ton: wir betonen einen Aussagesatz anders als einen Fragesatz.

Die Komponenten der Prosodie können durch subjektive und objektive Parameter charakterisiert werden. Subjektive Eigenschaften sind jene, die jeder hört (Satzmelodie, Betonung und Sprechrhythmus), objektive sind die mit der Maschine messbaren Merkmale. Zur Satzmelodie gehört also die Veränderung der Grundfrequenz (F_0) der Sprache. F_0 bedeutet die Schwingungszahl der Stimmbänder. Die Betonung ist mit drei physischen Parametern

beschreibbar: Erhöhung von F_0, Intensität und Dehnung der Lautlänge. Sprechrhythmus bedeutet die Variation der Zeitdauer der Sprechpartien.

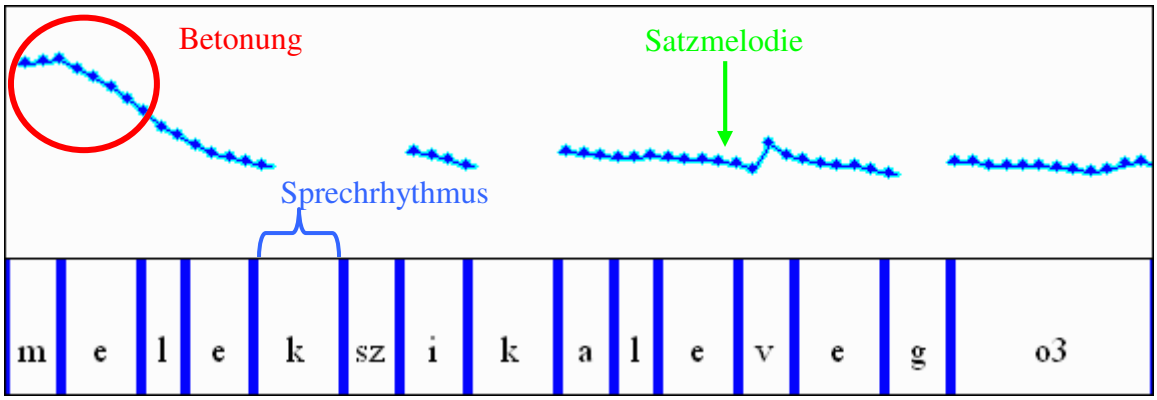

Abbildung 1.: Die drei Komponenten der Prosodie (Satz: „Melegszik a levegő“)

2.1.1. Satzmelodie

Eines der wichtigsten Mittel der menschlichen Sprache ist die Intonation [2], die der Sprecher mit der Variation der Grundfrequenz erzeugt. Grundsätzlich gibt es drei Formen von Sprachmelodie: aufsteigend, fallend oder schwebend. Bei der Sprachsynthese ist es wichtig, diese Intonation gut zu modellieren. Ein Beispiel für eine fallende Satzmelodie kann man auf Abbildung 1. sehen, wo ein Aussagesatz abgebildet ist.

2.1.2. Betonung

Durch die Betonung [2] kann der Sprecher auf eine Silbe, ein Wort oder einen Satzteil einen besonderen Akzent legen. Die Akzentuierung wird beeinflusst vom Gefühlszustand des Sprechers und vom Inhalt des Textes. Das bedeutet, dass in bestimmten Fällen man auch andere Silben akzentuieren kann. Ein Wort kann betont oder auch neutral sein. Die Abbildung 1. zeigt einen Akzent am Beginn eines Satzes, der durch die Erhöhung der Grundfrequenz realisiert wurde.

2.1.3. Sprechrhythmus

Der Sprechrhythmus [2] meint die Veränderung, Schwankung des Sprechtempos und die Pausen. Sprechtempo und Rhythmus hängen von zahlreichen Faktoren ab: Sprache,

Persönlichkeit des Sprechers, Gefühlszustand in dem Moment, Thema usw. Bei flüssiger Rede verändert sich die Artikulationsgeschwindigkeit bei verschiedenen Teilen der Lautfolge (einige Laute werden schneller, andere langsamer ausgesprochen). Die Variation der Lautdauer kann zwischen 10-20% betragen. Auf der Abbildung 1. sieht man die Lautdauer des abgebildeten Satzes.

2.2. Generationen von Vorlesautomaten

Sprachsynthese nennt man die Erzeugung von gesprochener Sprache durch einen Computer. Das TTS[1] System (Vorlesautomat) bekommt einen Text, der in verschiedenen Schritten verändert wird: Erstens produziert ein solches System symbolische Informationen auf Grund der zum Text gehörenden Phonemkette. Diese Information besteht aus prosodischer Information (z.B. mögliche Akzente), wodurch ein TTS den Ausgangston erzeugen kann.

Es existieren verschiedene Methoden zur Sprachsynthese. Bei fast jeder Methode wird eine Datenbank benutzt, deren Elemente zu der gewünschten Äußerung verknüpft werden. Man kann drei Generationen von Vorlesautomaten unterscheiden, die ich hier mit Hilfe der Arbeit von Fék et al. [3] und Wikipedia [4] vorstellen möchte.

2.2.1. Formantsynthese

Die erste Technologie, die Texte automatisch zu gut verständlicher Sprache formen konnte, war die Formantsynthese. Formant ist ein Bereich im Spektrum der Sprache, der mehr Energie aufweist, als seine benachbarten Frequenzbereiche. Abbildung 2. veranschaulicht ein Beispiel für menschliche Formantfrequenzen. Die Formantsynthese versucht sich der menschlichen Stimme mit akustischen Mitteln zu nähern. Eine Stimmquelle sorgt für ein Grundsignal, welches durch einen variablen Filter verändert wird. Die von solchen Systemen gebildeten Laute sind verständlich, klingen aber meistens robotisch, deshalb verlor diese Art der Sprachsynthese rasch an Bedeutung.

[1] Text-To-Speech

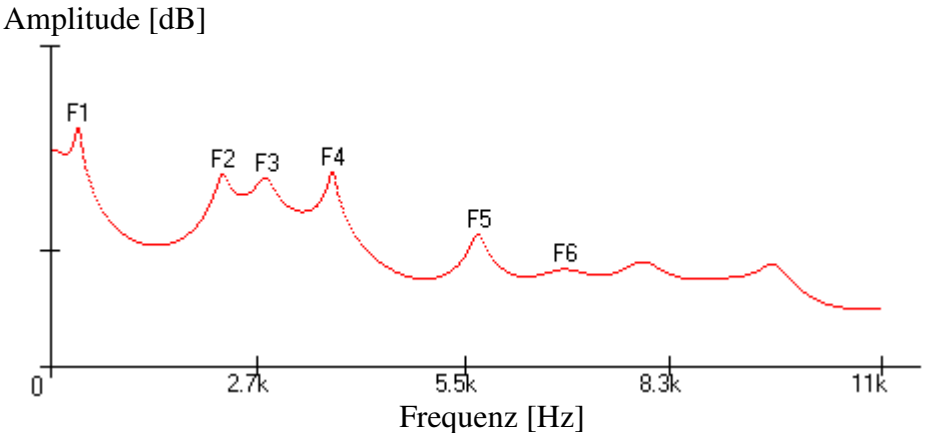

Abbildung 2.: Formantfrequenzen der menschlichen Sprache. Quelle: [5].

2.2.2. Diphonesynthese

Experimente Anfang des 20. Jahrhunderts haben gezeigt, dass die korrekte Wiedergabe der Lautübergänge wesentlich für die Verständlichkeit synthetisierter Sprache ist. Deshalb stellt man heute Datenbanken zusammen, in denen die Lautübergänge aller Lautpaare (z.B. a-b, a-c) gespeichert werden. In der Diphonesynthese werden diese, aus natürlicher Sprache ausgeschnittenen Wellenformen verbunden. Auf der Abbildung 3. wird der Prozess der Diphoneverkettung veranschaulicht.

Es ist auch möglich, andere Teile der Sprache zu nutzen: bei Triphonsynthese werden die Laute in allen möglichen Varianten (die von dem Kontext abhängen, z.B. b-a-b, b-a-c) gespeichert. Die so erzeugte Sprache ist verständlich, hat aber immer noch keine natürliche Lautung.

In dieser Arbeit wurde zu verschiedenen Experimenten ein Vorlesautomat verwendet, der von Olaszy et al. [6] entwickelt wurde. Profivox ist ein ungarisch sprechendes System, das 1444 Diphone und 6000 Triphone Elemente enthält. Der Vorlesautomat hat mehrere Vorlesetöne, wovon ich eine Mann-Variante verwendet habe.

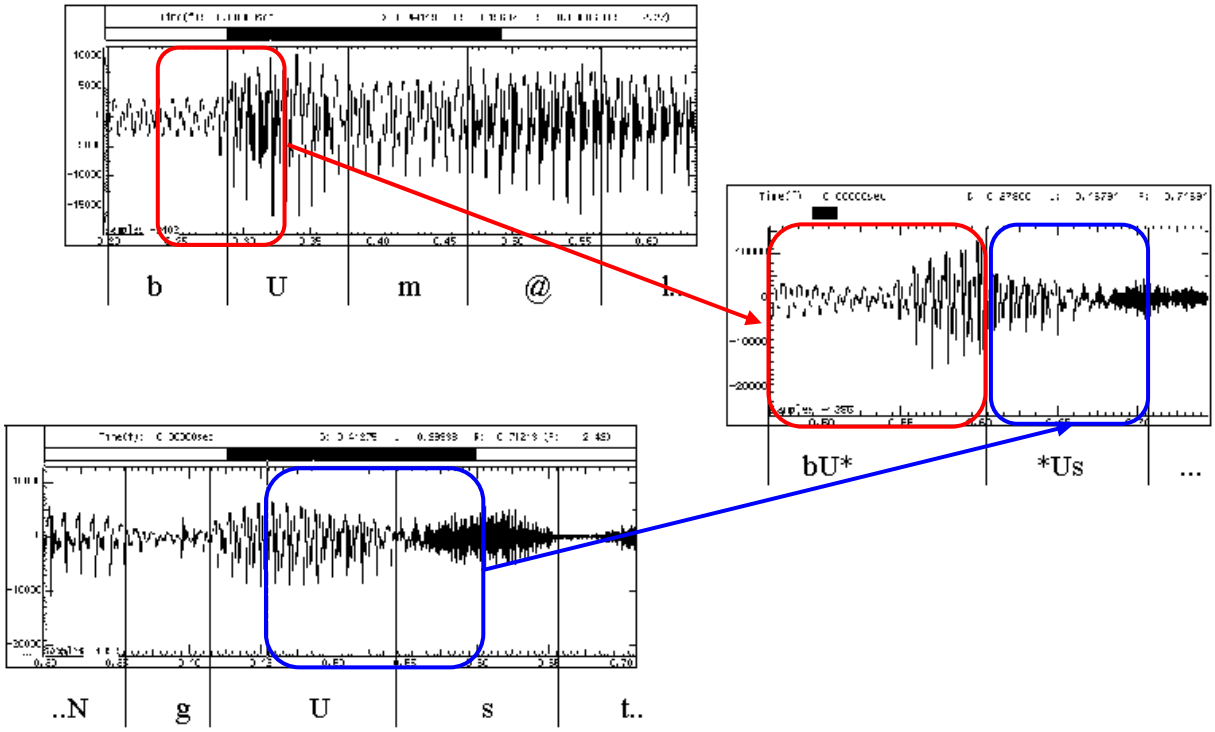

Abbildung 3.: Diphonesynthese: Verkettung der Bausteine /bU/ aus dem Quellwort „verbummeln“ und /Us/ aus „Languste“ zum Wort „Bus“. Quelle: [7].

2.2.3. Unit-Selection

Die Weiterentwicklung der Elementverbindung ist die Elementauswahl (in der Fachliteratur heißt dies Unit-Selection). In einem Unit-Selection Sprachsynthesesystem werden die Bausteine für das Ausgabesignal während der Laufzeit aus einem großen Sprachsignalkorpus ausgewählt [8]. Im Gegensatz zur klassischen konkatenativen Synthese (z.B. Diphonesynthese) können im Korpus mehrere Einheiten für jedes zu synthetisierende Element existieren, die idealerweise in verschiedenen prosodischen Varianten realisiert sind. Diese Bausteine können Phoneme, Silben, Wörter, Phrasen oder Sätze sein. Für die Synthese wird durch spezielle Suchalgorithmen eine Reihe möglichst großen Segmenten bestimmt. Da diese Reihe ohne oder mit wenig Signalverarbeitung ausgegeben wird, bleibt die Natürlichkeit der gesprochenen Sprache enthalten. Ein Unit-Selection System erfordert größere Speicher und mehr Rechnerkapazität als seine Vorläufer.

2.3. Automatische Prosodiegenerierung

Im Unterabschnitt 2.2 wurden die zwei grundlegenden Arbeitsschritte der Vorlesautomaten erläutert: im ersten wird eine symbolische Abschrift des Texts gemacht, im zweiten Schritt wird auf Grund dieser symbolischen Transkription die Ausgabewellenform synthetisiert. Diese zwei Arbeitsschritte können auch vereinigt sein, da wir die intermediäre symbolische Information nicht immer brauchen. Es gibt verschiedene Modelle zur Prosodiegenerierung, die bei Sprachsynthesesystemen angewendet sind [9].

2.3.1. Regelbasierte Modelle

Die Prosodie kann mit Hilfe von Regeln modelliert werden. Der Text muss markiert werden: die Sätze können nach ihren Typen, die Wörter und die Silben nach dem Akzent kategorisiert werden. Die Laute kann das Verfahren nach der Position im Wort, nach der Tonfolge, oder nach dem Kontext sortieren. In regelbasierter Modellierung wird die Prosodie eines Satzes auf Grund des Texts und der Marken[1] durch Regeln abgebildet. Diese Regeln sind vom Menschen definiert.

Der Vorteil dieses Modelltyps ist die Berechenbarkeit; also die Qualität der produzierten Prosodie wird immer ähnlich. Das ist wichtig, weil man die Veränderung schlecht duldet, wenn man sich an eine gegebene Qualität schon gewöhnte. Ein Nachteil der regelbasierten Modelle ist, dass die Bildung der Regeln schwer ist, es braucht die Kenntnis von mehreren Wissenschaftsgebieten. In der Praxis wird das regelbasierte Verfahren zusammen mit einem anderen Modelltyp verwendet.

2.3.2. Phonologische Modelle

Die phonologischen Modelle erzeugen symbolische Information aus dem Input. ToBi[2] ist ein typisches Beispiel für diese. Die Intonation einer Äußerung ist eine Abfolge von hohen (high, H) oder tiefen (low, L) Tönen. Dieses System verwendet drei Typen von Akzenten: Pitchakzent, Phrasenakzent und Grenztöne. Es gibt kein Programm oder Verfahren, das die

---

[1] Marken sind Eigenschaften der Laute, wie z.B. die Stelle des Wortakzents.

[2] Tones and Break indices

ToBi Marken (z.B. high, low) automatisch abbilden könnte. Die ToBi Marken sind nur für Englisch geeignet, es gibt eine andere Variante, GToBi[1], das deutsche Intonation transkribieren kann.

2.3.3. Superpositionsmodelle

Die Superpositionsmodelle haben die Eigenschaft, dass sie verschiedene Realisierungen (Satz-, Wort- und Lautlänge) von Prosodiekomponenten addieren, mit anderen Worten: *superponieren*. Zum Beispiel im Profivox-System [6] beginnt die Modellierung der Grundfrequenz auf dem höchsten (suprasegmentalen) Niveau, zuerst wird die Satzmelodie bestimmt. Sie kann aufsteigend, fallend oder schwebend sein. Dann werden die Grundfrequenzkurven der Wörter und Silben definiert. Schließlich folgt das niedrigste, (segmentale) Niveau, auf dem die Mikrointonation festgelegt wird. Die drei Niveaus werden superponiert und ergeben die endgültige Melodie des Satzes.

Profivox, das Sprachsynthesesystem, das ich zu dieser Arbeit verwendet habe, verwendet ein regelbasiertes Superpositionsmodell.

3. Steigerung der Prosodie von Vorlesautomaten

Eine der Schlüsselfragen der Sprachtechnologie ist – wie bereits erwähnt - die Realisierung von entsprechender, gleichzeitig abwechslungsreicher Prosodie (Satzmelodie, Betonung, Sprechrhythmus). Heute ist es nicht mehr genug, mit einem Regelsystem zu einem festgelegten Input-Text immer die gleiche Prosodie zu fügen. Das Ziel ist, der menschlichen Stimme nahe zu kommen.

Unterabschnitt 3.1 stellt meine Methode zur Prosodiekopierung dar. Im Unterabschnitt 3.2 findet man die Bedingungen und Ergebnisse eines Tests.

[1] German Tones and Break indeces

3.1. Meine Methode

Beim Studium der Fachliteratur traf ich keine solche Verfahren, die dem Problem der unnatürlichen Prosodie abhelfen konnte, deshalb begann ich mit der Steigerung der Prosodie von TTS Systemen zu experimentieren. Variabilität kann man auf vielerlei Art interpretieren: z.B. mit der Veränderung von Satzmelodie, Betonung, Sprechrhythmus oder Lautstärke. In dieser Arbeit beschäftige ich mich nur mit der Veränderung von Grundfrequenzkurven, und ließ die andere Parameter unverändert.

3.1.1. Sprachkorpus

In dieser Arbeit wurde eine Teilmenge von einem großen Sprachkorpus benutzt, die mir von BME-TMIT[1] zur Verfügung gestellt wurde. Die Teildatenbank enthält 200 ungarische Sätze im Thema Wettervorhersage, die von einer professionellen Sprecherin ausgesprochen wurde. Diese Aussagesätze bestehen aus einer prosodischen Einheit[2], um die Satzmelodien einfacher untersuchen zu können. In der Datenbank ist jeder Satz als Tonfile vorhanden, textlich transkribiert, und auch phonetisch mit Elementgrenzen transkribiert (Beginn und Ende der Laute und Wörter in der Aufnahme). Die Stimmlautperioden (also Grundfrequenzkurve) des Satzes sind auch vorhanden.

3.1.2. Untersuchung von Sätzen

Die Sätze des Sprachkorpus wurden verglichen, um bedeutende Charakteristiken festzustellen. Erstens suchte ich nach Sätzen mit ähnlicher Länge, danach nach solchen mit ähnlicher Satzmelodie. Ich hatte die These, dass durch Melodiekopierung von ähnlichen Sätzen prosodische Variabilität verwirklicht werden kann.

Die ähnliche Länge konnte zeitliche Ähnlichkeit bedeuten, aber bei Betrachtung der Silbenstruktur der Sätze näherte sich auch die Satzmelodie an. Ich gewann die Silbenzahlen der 200 Sätze aus den phonetischen Transkriptionen.

---

[1] Budapesti Műszaki és Gazdaságtudományi Egyetem, Távközlési és Médiainformatikai Tanszék

[2] Prosodiche Einheit ist ein Teil des Satzes, der ohne Pause ausgesprochen wird.

Serienzahl des Satzes	Silben-zahl	Wort-anzahl	Silbenzahl der Wörter
#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

Tabelle 1.: Die Gruppierung der Sätze auf Grund der Silbenzahl

Die Tabelle 1. bezieht sich auf den Wettervorhersage-Sprachkorpus. Es ist ersichtlich, dass in der Datenbank es einige Sätze mit identischer Silbenzahl gibt, aber keine mit vollständig identischer Silbenstruktur (Übereinstimmung der Silbenzahlen der Wörter). Das *#2031-#2380* Satzpaar wurde zu weiteren Experimenten ausgewählt, weil die ersten zwei Wörter der Sätze die gleiche Silbenzahl haben, deshalb kann man voraussetzen, dass die Wortakzente auch an ähnlicher Stelle sind.

Um die Grundfrequenzkurven der zwei Sätze besser beobachten zu können, ist es wichtig, die F_0-Kurven in der Zeit zu verschieben, damit die gleiche Teile der zwei Sätze in der Abbildung aneinander liegen. Dazu brauchte ich den ersten Satz als Referenz, und der zweite Satz wurde auf Grund der Silbengrenzen auseinander gezogen.

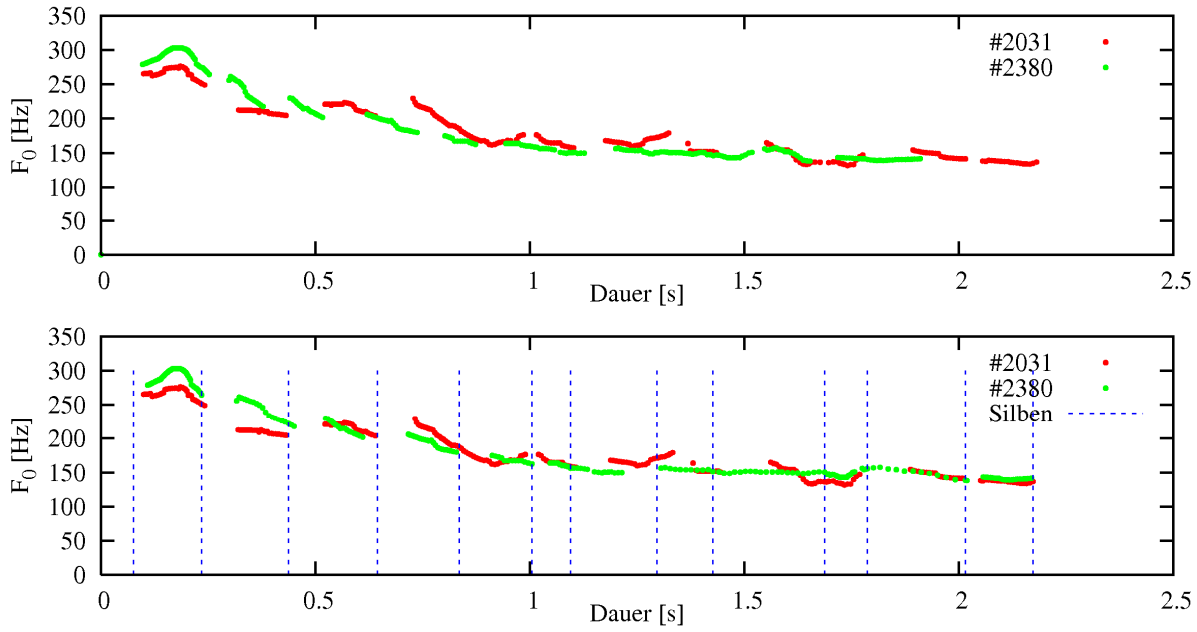

Abbildung 4.: Die Grundfrequenzkurven von Sätzen *#2031* und *#2380* ohne (oben) und mit (unten) Zeit-Anpassung

Auf der Abbildung 4. kann man die Satzmelodien der zwei ausgewählten Sätze im Vergleich sehen. Im oberen Teil sind die Sätze nicht synchron: obwohl ihre Silbenzahlen stimmen, sind die Längen nicht ganz gleich. Im unteren Teil der Abbildung kann man das Ergebnis der Zeit-Anpassung sehen. Da die Silbengrenzen der Sätze synchronisiert wurden, passen sich die F_0-Kurven besser an.

Nachdem ich dieses Satzpaar analysiert hatte, suchte ich weitere ähnliche Sätze im Sprachkorpus für ausführlichere Untersuchungen.

3.1.3. Melodiekopierung

In den bisherigen Sprachsynthesesystemen gehört zu einem Text immer die gleiche Prosodie. In meinem Versuch wollte ich die Variabilität der Melodie verwirklichen.

Um Satzmelodie zu einem Eingabetext abzubilden, braucht man einen Satz von dem Sprachkorpus. Ich hatte die These, dass wenn ich bei Sprachsynthese die Satzmelodie von einen ähnlichen Satz zu dem Eingabetext abbilde (also die Melodie kopiere), bekomme ich natürlich lautende Sprache.

Zur Melodiekopierung verwendete ich das ungarische Profivox-System, das von Olaszy et al. entwickelt wurde [6]. Dieser Vorlesautomat kann mit mehrerlei Eingaben verwendet werden: wenn man nur einen Text als Input gibt, bestimmt das TTS die

Grundfrequenzwerte, Lautdauer und Intensität. Durch ein intermediäres File können die Eigenschaften des synthetisierten Satzes determiniert werden. Dieses File enthält eine so genannte Intonationsmatrix, in der die Kodes der Laute, und die Grundfrequenz-, Längen-, Intensitätswerte definiert sind.

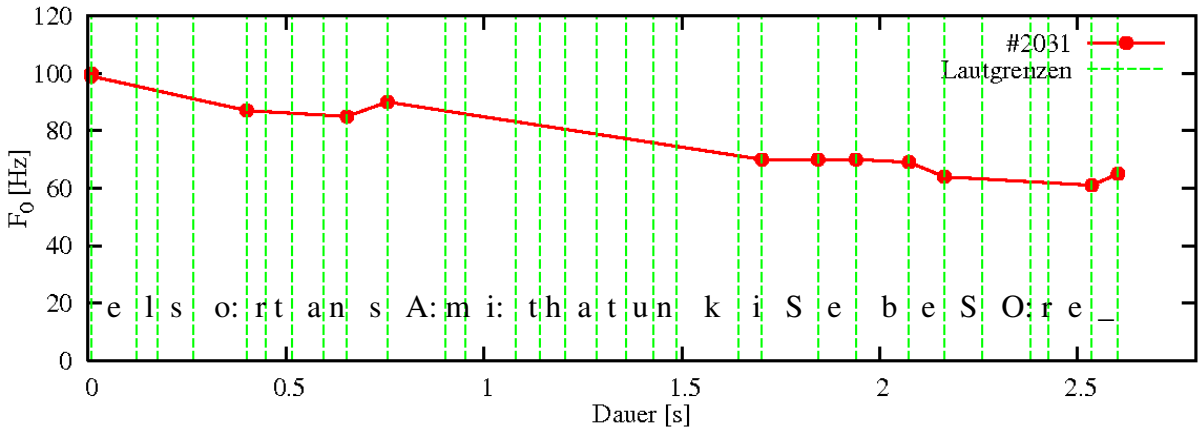

Abbildung 5.: Einstellung der Grundfrequenz in Profivox Intonationsmatrix: höchstens eine Bruchstelle pro Laut

Im Profivox-System kann man zu jedem Laut einen Grundfrequenzwert in Prozent angeben, wie auf der Abbildung 5. zu sehen ist. Das fixiert, wie hoch oder tief er im Vergleich zum Grundton (z.B. 110 Hz) sein soll. Die Stelle des Bruchs im Laut kann auch in Prozent angegeben werden. Danach produziert das TTS System die volle Satzmelodie mit linearer Interpolation auf Grund der gegebenen Bruchstellen.

Erstens versuchte ich den Satz *#3373* mit der Satzmelodie von Satz *#3056* zu synthetisieren. Ich produzierte die Dauer- und Intensitätswerte durch die Regeln von Profivox, und die F_0-Bruchstellen wurden manuell eingestellt. Ich versuchte die Satzmelodie *#3056* so genau wie möglich zu kopieren. Es wurde auch ein Programm geschrieben, das diese Kopierung automatisch ausführen konnte.

Abbildung 6. veranschaulicht, wie diese Methode funktioniert. Ausgehend von dem Eingabetext (*„Felhősödés estétől várható.”*) wird ein ähnlicher Satz (mit Ähnlichkeitsmaß definiert im Teil 3.1.2) in der Datenbank gesucht. Dieser Satz ist auf der Abbildung der Satz *#3056* (*„Egyeseket fejfájás gyötörhet.”*). Die Melodie des Satzes *#3056* wird verwendet, um die Satzmelodie des Eingabetextes zu definieren. Also die obere rote Grundfrequenzkurve auf der Abbildung wird an den Eingabetext kopiert. Das Profivox-System erzeugt durch diese symbolische Grundfrequenzkurve die Ausgangstonfolge.

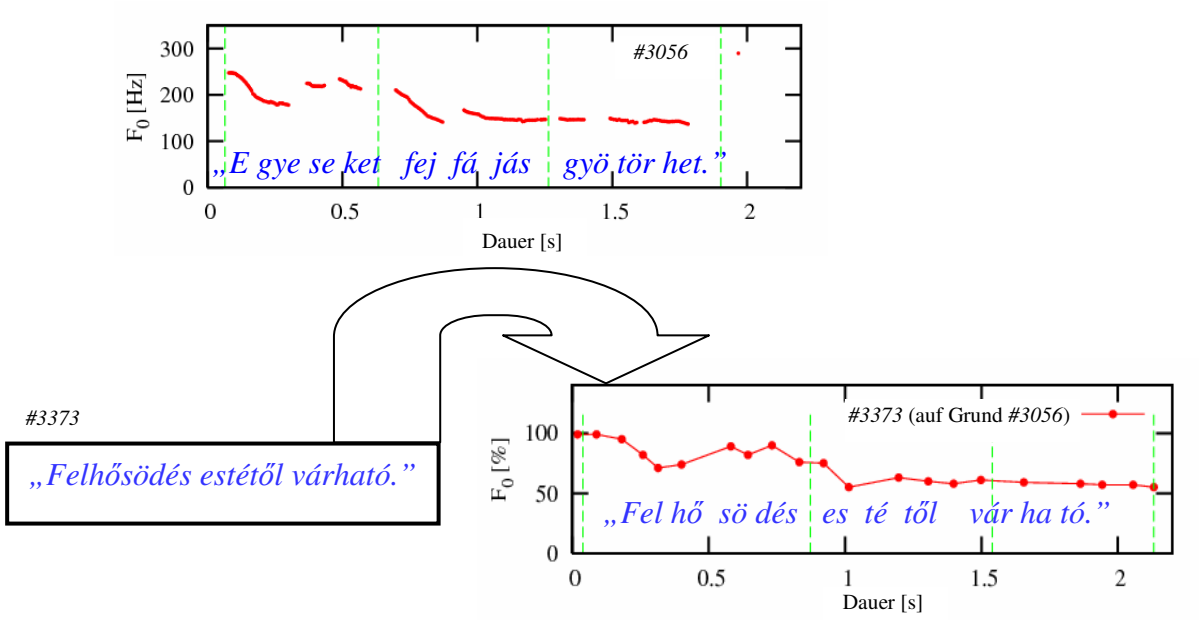

Abbildung 6.: Die Satzmelodie wird auf einen anderen Satz kopiert.

Wenn diese Methode gut funktioniert, können zu einem Eingabetext mehrere Sätze mit verschiedenen Prosodievarianten produziert werden. Um die Lautung meines Programms zu evaluieren, synthetisierte ich einige Sätze auf die Art wie nachfolgend genauer beschrieben.

3.2. Experiment

Ich erzeugte 28 Satzpaare um die Qualität meiner Methode zu kontrollieren. Die Satzpaare bestanden aus zwei Prosodievarianten (eine regelbasierte und eine mit Melodiekopierung) derselben Sätze. Mit dem Experiment wollte ich feststellen, ob die verschiedenen Varianten akzeptabel sind und ob der Unterschied zwischen ihnen bemerkbar ist.

3.2.1. Test-Bedingungen

Der Test wurde per Internet gemacht, durch ein Testsystem entwickelt von BME-TMIT. Die Testpersonen, die unsere Homepage besuchten, mussten erstens einen kurzen Text lesen und hören. Einige Daten über die „Hörer und Hörerinnen“ (Name, Lebensalter und Geschlecht) wurden gespeichert, danach begann das Hören von 28 Satzpaaren, was ungefähr 10 Minuten dauerte. Nach jedem gehörten Satzpaar sollten die Tester entscheiden, ob der

erste, der zweite Satz natürlicher war, oder man zwischen den beiden keinen Unterschied feststellen konnte. Der Test wurde zwischen 18. und 25. Oktober 2006 von 23 Leuten gemacht. Sie waren zwischen 20 und 30 Jahren alt, alle mit unversehrtem Gehör und ungarischer Muttersprache.

3.2.2. Ergebnisse

Die Ergebnisse des Testes sind auf der Abbildung 7. zu sehen. Man sieht, dass in einem Teil der Fälle ich die Melodiekopierung erfolgreich durchführen konnte. Bei fast der Hälfte aller Sätze wurde meine Methode präferiert, die regelbasierte Profivox-Satzmelodie wurde als ähnlich evaluiert, und bei 14% konnten sich die „Hörer und Hörerinnen“ nicht zwischen den zwei Varianten entscheiden.

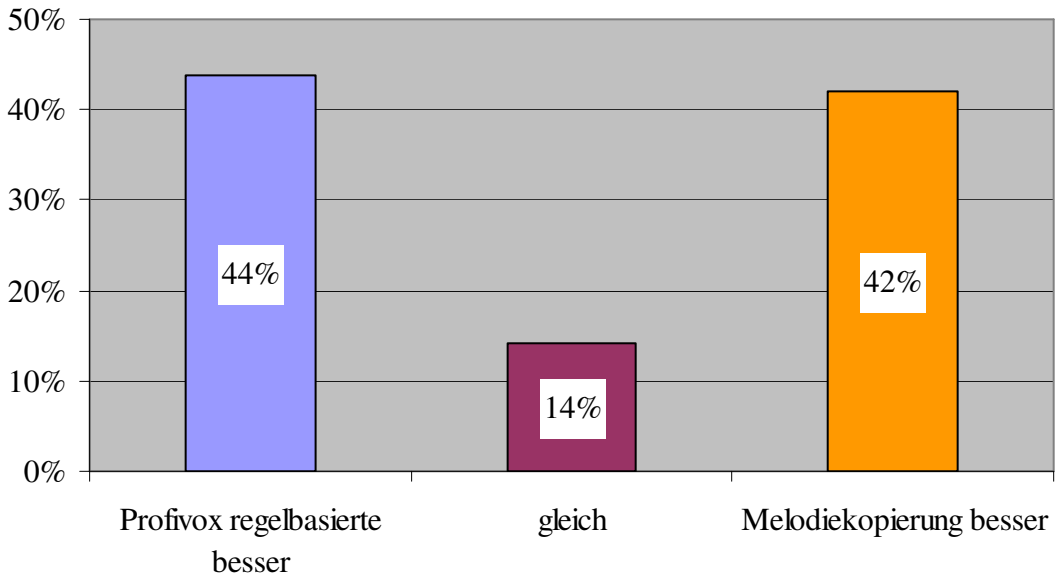

Abbildung 7.: Das Ergebnis des Experimentes

4. Konklusion, Zusammenfassung

Das Ziel meiner Arbeit war, die Natürlichkeit der synthetisierten Sprache zu steigern. Natürlichkeit bedeutete hier die Variabilität der Prosodie. Die Arbeit wurde mit einem Überblick über die Fachliteratur begonnen. Zuerst wurden die wichtigsten Komponenten der Prosodie erklärt, danach wurden die verschiedenen Generationen von Vorlesautomaten vorgestellt. Schließlich wurden einige Prosodiemodelle dargestellt. Das Experiment begann

mit der präzisen Definierung meines Zieles. Ich besorgte einen Sprachkorpus, in dem ich nach ähnlichen Sätzen suchte. Einige wurden zu Experimenten ausgewählt. Dann wurde ein Test ausgeführt, dessen Ergebnisse auch ausgewertet wurden.

5. Möglichkeiten zur Weiterentwicklung

Durch meine Methode kann man natürliche synthetisierte Sprache produzieren. In dieser Arbeit beschäftigte ich mich nur mit der Satzmelodie, einer der drei Komponenten der Prosodie. In Zukunft wäre es interessant, auch die Eigenschaften der Lautdauern zu untersuchen, und die regelbasierte Dauermodellierung von Profivox zu verändern.

Ein solches Text-To-Speech-System kann bei zahlreichen praktischen Anwendungen Verwendung finden, wie z.B. SMS-, E-Mail-, Buchvorlesautomat. Abwechslungsreiche Satzmelodie ist hauptsächlich bei der Vorlesung von längeren Texten vorteilhaft, weil dabei die Monotonie des TTS-Systems störend wäre.

6. Danksagung

Vielen Dank für die Hilfe, Ratschläge und nützliche Bemerkungen von meinem Konsulenten, Clemens Prinz.

7. Fachliteratur

[1] Németh Géza, Olaszy Gábor, Vorlesungsmaterial von „Sprach Information Systems“ Lehrfach, Abschnitt 1., Technische und Wirtschaftliche Universität Budapest, S. 17-18., 2005., *http://speechlab.tmit.bme.hu/postnuke/modules.php?op=modload&name=Downloads&file=index&req=getit&lid=91, 27. September 2007.*

[2] Olaszy Gábor, Kovács Magdolna, Nikléczy Péter, Gósy Mária, „Magyar nyelvi beszédtechnológiai alapismeretek. (600 oldal CD-ROM-on)”, Red.: Olaszy Gábor, Nikol Verlag, Budapest, 2002*., http://alpha.tmit.bme.hu/pub/beszinf/start.html, 30. September 2007* .

[3] Fék Márk, Pesti Péter, Németh Géza, Zainkó Csaba, „Generációváltás a beszédszintézisben”, *Híradástechnika*, Vol. LXI., Nr. 3, S. 21–30., 2006.

[4] Wikipedia: Sprachsynthese, *http://de.wikipedia.org/wiki/Sprachsynthese, 27. September 2007.*

[5] Bőhm Tamás, „Számítógépes program formánsok szemléltetésére”, *http://alpha.tmit.bme.hu/~tbohm/formant/Formant.html, 30. Oktober 2007.*

[6] 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, Nr. 3/4, S. 201-216., 2000.

[7] „Sprachsynthese”, Technische Universität Dresden, Institut für Akustik und Sprachkommunikation, *http://www.ias.et.tu-dresden.de/sprache, 21. Oktober 2007.*

[8] Julia Abresch, Stefan Breuer, „Unit-Selection-Sprachsynthese für die Telefonauskunft”, IKP-Arbeitsbericht NF 09, 2004., *http://www.ifk.uni-bonn.de/forschung/abteilung-sprache-und-kommunikation/ikp-arbeitsberichte-neue-folge/ikpab-nf09.pdf, 21. Oktober 2007.*

[9] Bernd Möbius, Vorlesungsmaterial von „Sprachsynthese I.“ Lehrfach, Universität Stuttgart, 2007., *http://www.ims.uni-stuttgart.de/lehre/teaching/2007-SS/Sprachsynthese-I/index.html, 22. Oktober 2007.*