VERGLEICHENDE UNTERSUCHUNG ZUR NATÜRLICHKEIT VON SYNTHETISCHEN F_0-KONTUREN

Hansjörg Mixdorff und Dieter Mehnert

Institut für Technische Akustik, TU Dresden, Mommsenstr. 13 01062 Dresden

EINFÜHRUNG

Die Natürlichkeit von synthetischen F_0-Konturen kann sinnvoll nur auf dem Wege über Wahrnehmungsexperimente ermittelt werden. In diesem Zusammenhang ist ein Vergleich von kompletten TTS-Systemen nur bedingt nützlich, weil dabei von den Hörern gefordert werden muß, von den Unterschieden, die aus der segmentalen Qualität oder den 'Sprechereigenschaften' herrühren, abzusehen und nur auf die prosodische Qualität zu achten. Daher wurden in der vorliegenden Untersuchung drei verschiedene Verfahren zur Generierung von F_0-Konturen in dasselbe TTS-System eingebettet, zwei regelbasierte [1,2] und ein Verfahren basierend auf neuronalen Netzen [3]. In Abb. 1 ist für jedes Verfahren ein Beispiel dargestellt.

Tab. 1. Mittelwert und Standardabweichung der Natürlichkeitsbewertung für die einzelnen Verfahren im ersten Versuch [4]. Wertebereich der Natürlichkeitsskala: 0..4.

Verfahren	μ	σ
Mixdorff/Fujisaki	2.34	0.31
Hirschfeld	1.59	0.41
Neuronales Netz	1.45	0.53

1. VERSUCHSDESIGN UND VERWENDETE SPRACHPROBEN

Im ersten Teil des Experiments [4] waren insgesamt 14 Sätze unterschiedlicher Länge und Satztypen erstellt und in einem rechnergestützten Paarvergleich 20 Hörern (10 phonetisch geübt, 10 ungeübt) angeboten worden. Es zeigte sich eine Präferenz für die beiden regelbasierten Verfahren und speziell für das Verfahren von Mixdorff/Fujisaki (siehe Tabelle 1), die bei den 'Experten' besonders ausgeprägt war.

Der vorliegende Beitrag hat zum Ziel, das Ergebnis des ersten Versuchs einzuordnen, d.h. es soll die Natürlichkeit der synthetischen F_0-Konturen mit derjenigen von natürlichen F_0-Konturen verglichen werden. Ein direkter Vergleich mit natürlichen Lautäußerungen wurde verworfen, weil dieser die gleichzeitige Veränderung mehrerer Sprachsignaleigenschaften wie segmentale Qualität und Dauer mit sich gebracht hätte.

Daher wurden für insgesamt 8 Sätze aus [4] die F_0-Konturen aus natürlichen Lautäußerungen dieser Sätze auf synthetische Stimuli kopiert (siehe Beispiel Abb. 2, oben). Zusätzlich wurden Stimuli erzeugt, bei denen außer der kopierten F_0-Kontur die segmentalen Dauern weitestgehend denen in der natürlichen Lautäußerung angepaßt wurden (Abb. 2, Mitte). Ein Vergleich von synthetischen und natürlichen Konturen zeigt, daß letztere deutlich größere F_0-Auslenkungen und stärkere mikroprosodische Schwankungen aufweisen.

Zur Bestimmung der Kopierkonturen wurden zunächst in den natürlichen Lautäußerungen (Abb. 2, unten) die Lautgrenzen markiert. Dann wurden die F_0-Kontur in 10 ms-Intervallen extrahiert und die Zeitpunkte der F_0-Stützwerte auf die Lautlängen bezogen. Die resultierenden F_0-Werte wurden dann

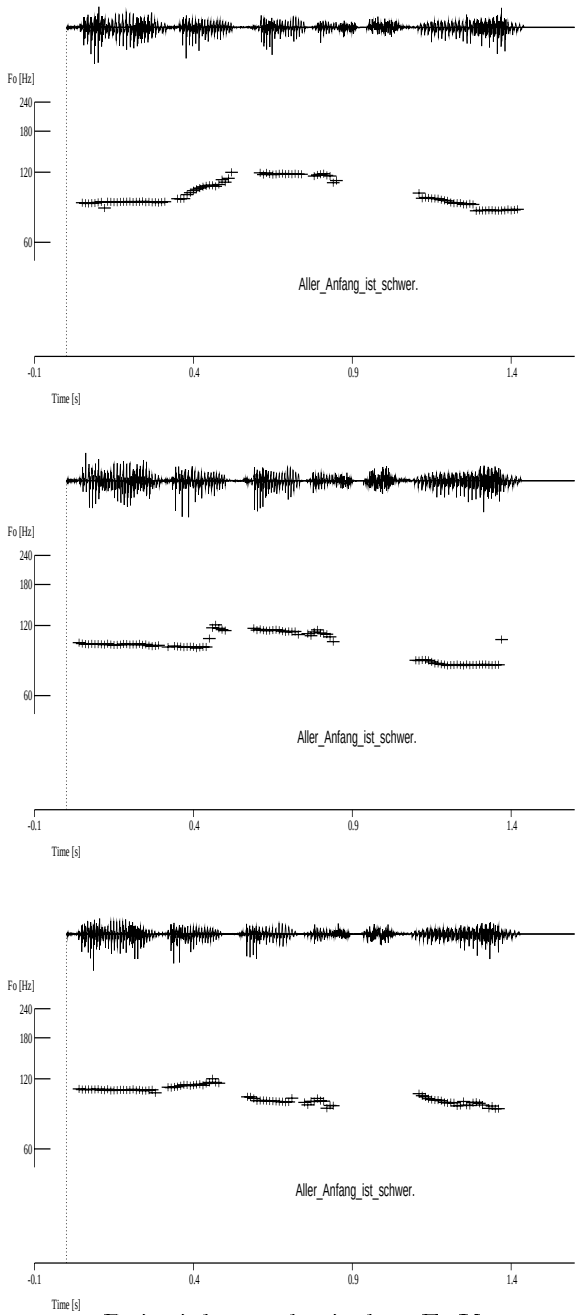

Abb. 1. Beispiele synthetischer F_0-Konturen für Satz 2 "Aller Anfang ist schwer." Von oben nach unten: Mixdorff/Fujisaki, Hirschfeld, Jokisch.

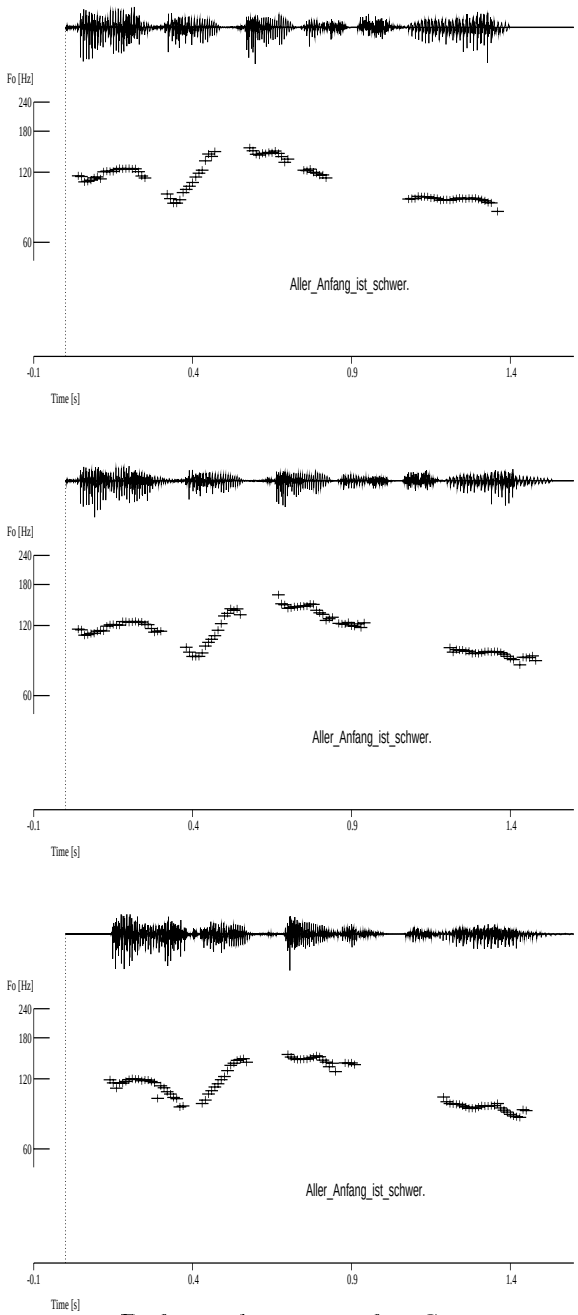

Abb. 2. Referenzkonturen für Satz 2. Von oben nach unten: Kopierte Originalkontur, kopierte Originalkontur und Originallängen, Originalkontur.

der synthetischen Lautäußerung aufgeprägt. Der Sprecher der natürlichen Beispiele hatte auch das bei der PSOLA-Synthese verwendete Diphon-Inventar produziert.

Im Gegensatz zum in [4] beschriebenen Versuch wurde eine modifizierte Variante des auf neuronalen Netzen beruhenden Verfahrens eingesetzt, die mit einer vergrößerten

Sprachdatenbasis trainiert worden war. Die resultierenden Sprachproben wurden wieder 20 Hörern paarweise verwürfelt angeboten, wobei diesmal die Wiedergabe von Tonband nach dem Muster *Probe A, 2.5 s Pause, Probe B, 5 s Pause, Probe A, 2.5 s Pause, Probe B, 10 s Pause* erfolgte.

Die Probanden hatten wie im ersten Versuch die Aufgabe, entweder die Natürlichkeit von Probe A als besser oder schlechter oder beide Proben als gleich gut bzw. schlecht zu bewerten.

2. ERGEBNISSE DES HÖRVERSUCHS

In Tabelle 2 sind die Ergebnisse der Natürlichkeitsbewertung nach Sätzen aufgeschlüsselt dargestellt. Bei der Bewertung wurde dem Verfahren, das im Paarvergleich besser abschnitt, ein Punkt gegeben, bei einer 'gleich'-Bewertung beiden Verfahren ein halber und bei einer 'schlechter'- Bewertung keiner. Da jedes Verfahren jeweils mit vier anderen verglichen wurde, liegt die maximal erreichbare Punktzahl bei 4. Es wird deutlich, daß die Rangfolge der Verfahren in Abhängigkeit vom Satz stark schwankt.

Im Mittel (siehe Tabelle 3) wurden die Proben mit Kopiekontur und natürlichen Längen am besten bewertet und liegen deutlich vor jenen mit Kopiekontur und synthetischen Längen. Ähnlich gut wie erstere schneidet das Verfahren Mixdorff/Fujisaki ab. Im Gegensatz zum Ergebnis im ersten Versuch liegt der modifizierte Ansatz mit neuronalen Netzen jetzt vor dem regelbasierten von Hirschfeld und wird für Satz 3 sogar als am natürlichsten empfunden. Die Natürlichkeitsbewertung für dieses Verfahren schwankt auch nicht mehr stärker als bei den regelbasierten Verfahren, wie an der Standardabweichung im Vergleich mit Tabelle 1 abzulesen ist.

Abb. 3 gibt einen Überblick über die Gesamtbewertung der einzelnen Verfahren. In den Tortendiagrammen sind die Entscheidungen in den insgesamt 640 (20 Vps*4 Paare pro Satz*8 Sätze) Paarvergleichen pro Verfahren (entspr. 100%) wiedergegeben. Den Diagrammen ist z.B. zu entnehmen, daß das neuronale Netz bei 28.1% der Paarvergleiche als besser, bei 22.8% als gleich gut und bei 49.1% als schlechter eingestuft wurde. Es fällt auf, daß bei allen Verfahren relativ ähnlich bei etwa einem Viertel der Entscheidungen einer Gleichbewertung der Vorzug gegeben wurde. Der Anteil liegt übrigens bei den ungeübten Hörern mit einem Mittel von 25.1% nur leicht über dem der geübten mit 23.5%.

Tab. 3. Mittelwert und Standardabweichung der Natürlichkeitsbewertung für die einzelnen Verfahren.

Verfahren	μ	σ
Kopiekontur, natürliche Dauern	2.48	0.45
Mixdorff/Fujisaki	2.35	0.52
Kopiekontur, synthetische Dauern	2.08	0.49
Neuronales Netz (modifiziert)	1.58	0.50
Hirschfeld	1.48	0.54

Die Tatsache, daß die Proben mit kopierter F_0-Kontur nicht durchgehend als besser bewertet wurden, ist möglicherweise damit zu erklären, daß die größeren F_0-Auslenkungen sich bei einigen Beispielen ungünstig auf die segmentale Qualität beim PSOLA-Verfahren auswirken, da dieses nur eingeschränkt F_0-Variationen (ca. +/-30%) zuläßt.

3. DISKUSSION UND AUSBLICK

Im vorliegenden Beitrag wurde der Versuch unternommen, einen Vergleich zwischen der Natürlichkeit synthetischer und natürlicher (kopierter) F_0-Konturen anzustellen. Es zeigt sich, daß der geringe Umfang des untersuchten Materials nicht ausreicht, um eine abschließende Beurteilung zu ermöglichen. Es wird jedoch deutlich, daß die Annäherung an natürliche Lautdauern eine deutliche Verbesserung bewirkt, was in Hinblick auf die Sprachsynthese bedeutet, daß eine unzulängliche Dauersteuerung der mit einem 'guten' F_0-Verfahren erreichbaren Natürlichkeit Grenzen setzt. Als Abschluß der Reihe von perzeptiven Versuchen zur prosodi-

Tab. 2. Rangfolge und Rating für die F_0 Synthese-Verfahren von Hirschfeld (HI), Mixdorff/Fujisaki (FU) und neuronale Netze (NN) im Vergleich zu den Referenzverfahren mit kopierter natürlicher Kontur ohne (MA) und mit (MD) natürlichen Dauern in Abhängigkeit vom Satz. Wertebereich für den Natürlichkeitsscore: 0.0 - 4.0.

Nr.	Text	1	2	3	4	5
1	"Bereitwillig gab er Auskunft."	MD(2.63)	FU(2.50)	NN(2.25)	MA(1.60)	HI(1.02)
2	"Aller Anfang ist schwer."	MD(2.75)	MA(2.05)	FU(2.03)	HI(1.98)	NN(0.95)
3	"Die Begründung ist stichhaltig."	NN(2.25)	FU(2.23)	MD(2.08)	MA(1.93)	HI(1.53)
4	"In dem Lehrbuch sind..."	MD(2.95)	MA(2.42)	FU(2.00)	NN(1.48)	HI(1.15)
5	"Üben und immer wieder üben..."	FU(2.93)	MD(2.30)	HI(2.03)	MA(1.73)	NN(1.02)
6	"Das Gespräch zeigte..."	FU(2.88)	MD(2.25)	MA(2.13)	HI(1.43)	NN(1.33)
7	"Wenn wir die Maschine..."	FU(2.80)	HI(2.10)	MD(1.93)	MA(1.60)	NN(1.58)
8	"Es regnete soviel..."	MD(3.10)	MA(3.05)	NN(1.80)	FU(1.45)	HI(0.60)

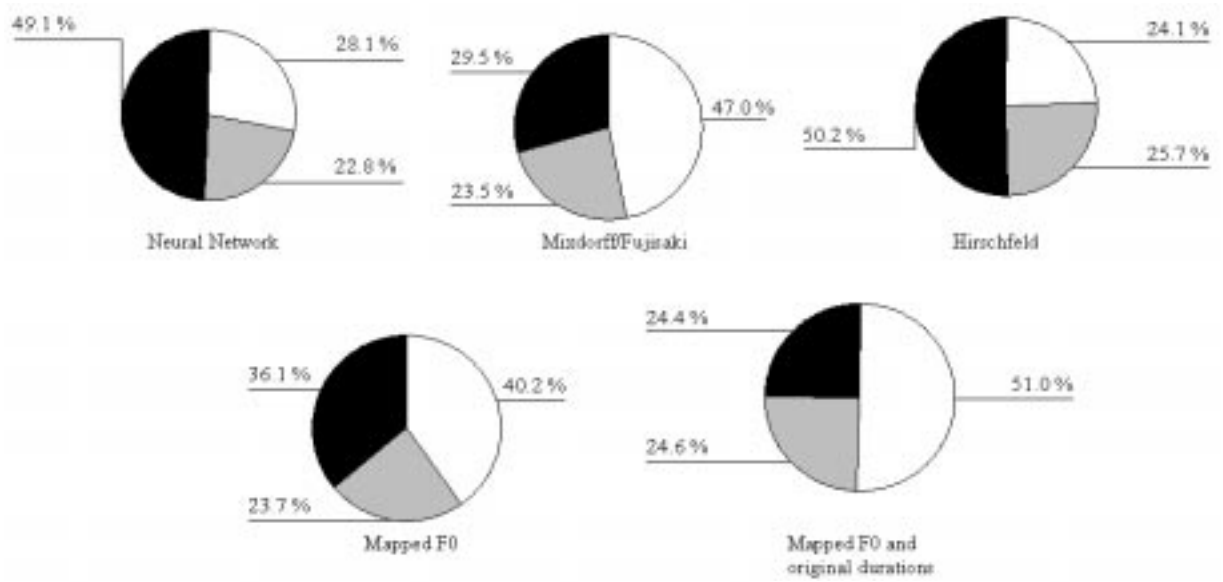

Abb. 3. Tortendiagramme mit der Natürlichkeitsbewertung je Verfahren. Weiß: Im Paarvergleich als besser eingestuft, grau: als gleich gut eingestuft, schwarz: als schlechter eingestuft.

schen Qualität von TTS-Systemen ist ein Experiment geplant, bei dem komplette Systeme mit ähnlicher segmentaler Qualität verglichen werden sollen.

LITERATUR

[1] Hirschfeld, D. (1996): The Dresden Text-to-Speech System, 6th Czech-German Workshop on Speech Processing, Prague, Sept. 1996, pp. 22-24.

[2] Mixdorff,H., Fujisaki, H. (1995): A Scheme for a Model-based Synthesis by Rule of F0 Contours of German Utterances. *Proceedings of the '95 Eurospeech*, Madrid, Spanien, Bd. 3, pp. 1823-1826.

[3] Jokisch, O., Pescheck, M. (1997): Neuronale Prosodiegenerierung - Einfluß der Trainingsdaten, In *Fortschritte der Akustik 1998*, Zürich, Schweiz.

[4] Mixdorff,H., Mehnert, D. (1998): Perceptual Evaluation of Three Different Approaches for Generating F_0 contours in TTS. In *Fortschritte der Akustik 1998*, Zürich, Schweiz.