

# **BLF – ein Labelformat für die maschinelle Sprachsynthese mit BOSS II**

*Stefan Breuer, Julia Abresch, Petra Wagner und Karlheinz Stöber  
breuer/abresch/wagner/stoeber@ikp.uni-bonn.de*

*Institut für Kommunikationsforschung und Phonetik  
Universität Bonn*

Für das am IKP in der Entwicklung befindliche Open-Source-Synthesystem BOSS II wurde ein eigenes Labelformat (BLF) erstellt. Die Besonderheit des Formates gegenüber früher genutzten Konventionen ist die Zusammenfassung von zeitlich nicht segmentierbaren Lautkombinationen zu jeweils einem Segment, was im Rahmen der Unit-Selection-basierten Synthesysteme Verbesserungen in der Segmentierungs- und Konkatenationsqualität bringen sollte.

## **1 Einleitung**

Für die neueren korpusbasierten Ansätze in der Sprachsynthese, wie das am IKP entwickelte Open Source-Synthesystem BOSS II (Klabbers et al., 2001), ist die Segmentierung und Etikettierung großer Mengen von Sprachsignalen erforderlich. Zu diesem Zweck müssen ein Symbolsatz, der die im Korpus vorkommenden Laute beschreibt, und Regeln zu dessen Anwendung definiert werden. Häufig wird als Basis für den Symbolsatz das ASCII-Transkriptionsalphabet SAMPA (Wells, 2000) verwendet, so auch im BMBF-Projekt Verbmobil (Kohler et al. 1995; Wahlster, 2000). Die Erfahrungen zeigen jedoch, dass die ohnehin schwierige Segmentierung des Signals durch die von der phonetischen Transkription (in SAMPA) implizierte Zuordnung von nur einem Laut zu jeweils einem Signalabschnitt in der Praxis große Probleme verursacht. Dies trifft insbesondere für Lautkombinationen zu, die koartikulatorisch stark miteinander verwoben sind. Auch die in Verbmobil zulässige Vielzahl an Varianten zur manuellen Annotierung phonetischer Variabilität führte zu Schwierigkeiten in Form von Inkonsistenzen bei der Vergabe der Etiketten. Das BOSS-Labelformat (BLF), das für BOSS II entworfen wurde, versucht diese Probleme zu minimieren. Auch hier dient als Basis der (X-)SAMPA-Symbolsatz, jedoch wurden einige phonetisch-akustisch motivierte Anpassungen in den Regeln zur Zuordnung der Symbole zum Sprachsignal vorgenommen, wie z.B. die Gruppierung bestimmter Lautfolgen zu einem Symbol. Dadurch wird die einzelne Verwendung schwer segmentierbarer Laute in der Synthese zu Gunsten der Qualität vermieden. Im Folgenden wird die Struktur des BLF, besonders hinsichtlich der Unterschiede zu bisher verwendeten Labelkonventionen, erläutert.

## **2 Das BOSS-Labelformat**

### **2.1 Das Dateiformat**

Ein BOSS II Label-File ist zeilenorientiert aufgebaut und orientiert sich stark an dem in Phondat II (Pompino-Marschall, 1992) verwendeten Format. Im Gegensatz zu Phondat II gibt es jedoch keinen Header, und die Konventionen zur Markierung von Auslassungen,

Einfügungen und Vertauschungen sind strenger. Jede Zeile, deren erstes Nichtblankzeichen ungleich »//« ist, enthält ein Lautlabel. Die erste Spalte einer Zeile gibt die Position des Lautes in Samples an. Die zweite Spalte beinhaltet das Lautlabel und ggf. zusätzliche Informationen. Beginnt eine Zeile mit einem doppeltem Schrägstrich (Slash, //), so gilt sie analog zur C++-Syntax als Kommentar und wird ignoriert. Die Folge der Lautlabel kodiert zusätzlich die Wort-, Morphem- und Silbengrenzen, die Position des lexikalischen Akzents, sowie Einfügungen, Vertauschungen und Auslassungen relativ zur kanonischen Transkription.

## 2.2 Der Symbolsatz und seine Anwendung

Das hier vorgestellte Symbolinventar für die Segmentierung von Sprachsignalen für den Einsatz in BOSS ist im Hinblick auf die Bearbeitung mit einem automatischen Lautsegmentierer und die spätere Konkatenation der Einheiten optimiert. Das bedeutet, dass im Gegensatz zu SAMPA-D oder SAMPA-D-VMlex z.T. ganze Ketten von Lauten und häufige Morpheme bzw. unakzentuierte Silben, wie *-en* als **ein** Labelsegment definiert sind. Dies ist z.B. dann der Fall, wenn die Realisierungen dieser Laute koartikulatorisch stark miteinander verflochten sind. So werden z.B. Folgen von Halbvokalen/Liquiden und Vokalen, sowie /h/+Vokal als jeweils ein Segment definiert. Ein solches Vorgehen ist deshalb sinnvoll, weil die betreffenden Lautklassen kaum kontextunabhängige akustische Merkmale besitzen und sich auch selten zeitlich von den Folgevokalen abgrenzen lassen. In Abbildung 1 ist dies deutlich zu erkennen. Eine Segmentgrenze zwischen den Lauten /j/ und /e:/ lässt sich nicht zufriedenstellend festlegen, zumal die Artikulationsorte der beiden Laute auch noch sehr nah beieinander liegen.

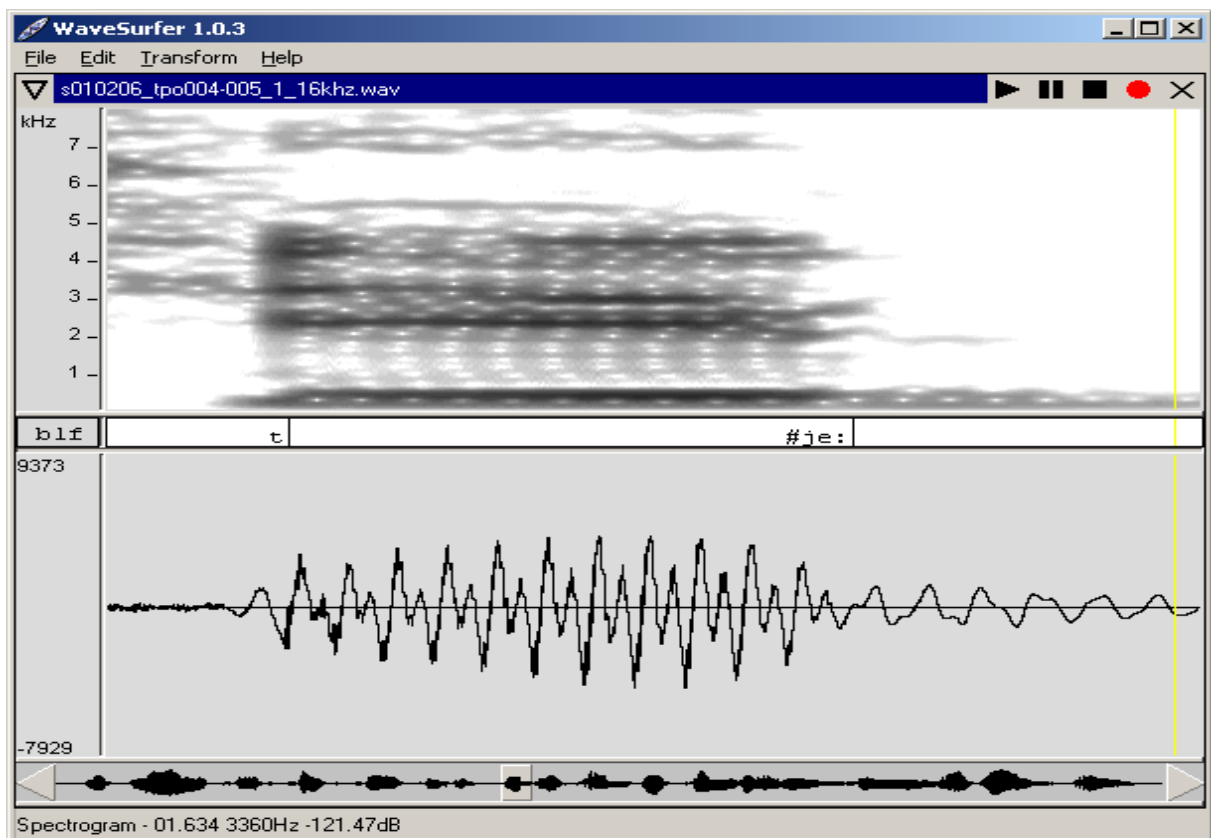


Abbildung 1: Sonogramm, Etikettierung und Oszillogramm der Lautfolge [tje:d] in der Äußerung "hat jedoch".

Wird in diesen Fällen eine Segmentierung erzwungen, wie es z.B. auch im Verbmobil-Projekt Konvention war, kommt es zu Qualitätseinbußen bei der späteren Konkatenation im Rahmen der Synthese. Dies betrifft insbesondere Laute, die sich auch in (phonologisch) gleichen Kontexten von Realisierung zu Realisierung stark unterschiedlich ausprägen. In rein diphonbasierten konkatenativen Systemen, die mit einem festen Inventar von Einheiten arbeiten, stellt dies kein Problem dar, weil pro Kontext nur eine Realisierung vorliegt. Erst mit den neueren, auf Unit Selection basierenden Synthesystemen wird die Problematik der Segmentierung an diesen Punkten relevant.

Tabelle 1 listet diejenigen Laute auf, die in BLF im Anlaut mit einem folgenden Vokal oder Diphthong zu einem Segment verschmelzen. Bis auf /l/, das auch in der Koda vorkommt, werden alle Symbole ausschließlich in solchen Verbindungen genutzt. Die betroffenen Lautklassen sind die Approximanten des Deutschen, wobei hier auch das häufig als Approximant [v] realisierte /v/ und die in englischen und französischen Lehnwörtern auftauchenden [w] und [ʍ] dazugerechnet werden, die Liquide inklusive aller im Anlaut vorkommenden Realisierungsvarianten von /r/, sowie die stimmhaften und stimmlosen Varianten von /h/, dessen akustische Ausprägung vollständig vom vokalischen Kontext abhängig ist.

| IPA                | BLF | (X-)SAMPA                  | Beispiel                       |
|--------------------|-----|----------------------------|--------------------------------|
| j / j̥ / ç / j̧    | j   | j / j\ / c / <i            | Jahr / Ach ja                  |
| l                  | l   | l                          | Lauge / Wald                   |
| ʀ / R / ʁ / ʁ̥ (ʀ) | r   | R / R\ / R\_o / R\_o_0 (X) | Raute / Raute / Waren / treten |
| h / h̥             | h   | h / h\                     | Hand                           |
| v / v̥             | v   | v / p                      | wann                           |
| ʍ                  | H   | H / <u                     | Etui                           |
| w / ʋ              | w   | w / <o                     | Reservoir                      |
| ʔ                  | ʔ   | ʔ (Q)                      | _Apfel                         |

**Tabelle 1: Lautsymbole, die im BLF nur im Verbund mit Folgevokalen ein Segment bilden können.**

Zu den Diphthongen, die sich mit den obigen Symbolen verbinden, werden hier auch die Vokale mit folgendem vokalisiertem /r/ bzw. hinteren Zentralvokal [ɐ] gezählt, wobei eine im Silbenauslaut vorkommende Realisierung von /r/ in BLF immer als vokalisiert transkribiert wird. Ein [ɐ] (SAMPA: [ɛ]) wird also stets mit dem Vorgängervokal und eventuell vorangehenden Lauten der in Tabelle 1 angegebenen Klassen zu einem Segment verbunden. Ein solches Segment endet aber stets an der Silbengrenze, so dass die Triphthonge [aɪ.ɛ] (wie in "Eier"), [aʊ.ɛ] (wie in Mauer) usw. jeweils zwei Segmente darstellen.

Einen weiteren Fall, in dem die einzelnen Laute zeitlich nur schwer separierbar sind, bilden die Schwa-Konsonant-Folgen in unbetonten Endsilben (orthographisch: -en, -em, -el). In der Regel werden in der Spontansprache diese Folgen zu silbischen Konsonanten reduziert. In Lesesprache kann der Neutralvokal aber durchaus voll realisiert werden, oder es sind zumindest geringfügig vokalische Anteile im Signal enthalten. Letztere aber als Schwa zu etikettieren und gesondert in der Konkatenation als solche einzusetzen macht wenig Sinn. Weder das [ə] noch der folgende Konsonant lassen sich dann in einer Weise separieren, die einen sinnvollen Einsatz der Einzellaute in der Synthese ermöglicht. Daher wird in BLF auch gar nicht erst versucht, diese Segmentierung vorzunehmen. Die in Tabelle 2 aufgelisteten

Symbolketten stehen jeweils für die silbisch realisierten und die mit vokalischen Anteilen realisierten Endsilben (für [ŋ] ist letzteres auszuschließen). Ein wider Erwarten voll realisierter Neutralvokal müsste hingegen als Einfügung markiert werden (s. Markierung von Abweichungen).

| IPA     | BLF | (X-)SAMPA | Beispiel                 |
|---------|-----|-----------|--------------------------|
| əl / ɫ  | @l  | l=        | Rubel                    |
| ən / ŋ  | @n  | @n, n=    | raten / raten            |
| əm / m̩ | @m  | @m, m=    | jedem / Raben            |
| ŋ       | @N  | N         | ragen (bei [ə]-Elision!) |

**Tabelle 2: Etikettierung der Endsilben Schwa + n/m/l**

Die Folge @N in Tabelle 2 darf für die kanonische Transkription im Wörterbuch der Synthese nicht eingesetzt werden. Das BLF-Symbolinventar ist zweigeteilt: ein Teil für die kanonischen Wörterbucheinträge, also die breiten Transkriptionen der Sollaussprache und für die engen Transkriptionen im Rahmen der manuellen Etikettierung und Segmentierung der Sprachsignale des Synthesekorpus, der andere Teil steht nur für die enge Transkription zur Verfügung. Die Symbole für die hier nicht aufgelisteten Konsonanten und Vokale entsprechen den SAMPA-, bzw. X-SAMPA-Konventionen, aus Platzgründen wird hier auf eine Abbildung verzichtet. Das vollständige Inventar für die Transkription des Deutschen kann in der Dokumentation zu BOSS II (Stöber et al., 2000) nachgeschlagen werden.

### 2.3 Markierung von Abweichungen

Um die im Korpus auftretenden Abweichungen von der kanonischen Form zu markieren, werden, in Anlehnung an die Phondat-Konvention, die in Tabelle 3 aufgeführten Zeichen verwendet. Die Markierung von Auslassungen (Elisionen) wird mittels nachgestelltem "-" vorgenommen, Einfügungen (Epenthesen) mit "+" und Vertauschungen (also die Ersetzung eines Lautes oder einer Lautkette durch andere) durch Anhängen von "-" und den Symbolen für den neuen Lautwert. Die Anwendung dieser Konventionen wird jedoch strikter gehandhabt als es in Phondat der Fall war. So werden phonetische Detailbeschreibungen, wie sie für die Annotation von Korpora für linguistische Untersuchungen durchaus sinnvoll sind, in BLF weitgehend vermieden. Abweichungen in den Vokalqualitäten werden z.B. nur bei deutlicher Ausprägung markiert, etwa wenn ein anderes Vokalphonem realisiert wurde, oder wenn eine starke Reduktion stattgefunden hat, die die Etikettierung als Neutralvokal rechtfertigt. Denn gerade bei Vokalen ist die Korrespondenz zwischen den Urteilen verschiedener menschlicher Etikettierer häufig schlecht (siehe: Eisen et al. 1992).

Auch die Menge an verfügbaren Diakritika zur Beschreibung abweichender Lautqualitäten ist stark begrenzt. Welche Diakritika Verwendung finden, ist ebenfalls aus Tabelle 3 ersichtlich. Allgemein ist die Toleranzschwelle für Abweichungen von der kanonischen Form niedrig angesetzt. Bereits während der Aufnahmen für das Korpus soll stark auf eine orthophone Realisierung geachtet werden. Wenn Abweichungen außerhalb eines akzeptablen Rahmens liegen, wird das gesamte betroffene Wort mit dem Etikett "#\$j" versehen. So markierte Einträge können von einer Sprachsynthese erkannt und ignoriert werden.

| Sonderzeichen           | Bedeutung   | Beispiel  |
|-------------------------|---|---|
| - (Minus / Bindestrich) | markiert Auslassung oder Vertauschung eines Lautes (Elision / Assimilation) | Du hast<br>d"u:#"hast-<br>Er hat mir<br>"?e:6#"hat-p#"i:6 |
| + (Plus)                | markiert Einfügung eines Lautes (Epenthese)                                 | Samstag<br>z"amp+s;t%a:k                                  |
| #\$j                    | markiert einen für die Synthese inakzeptablen Signalabschnitt               | Störgeräusche, nicht adäquat realisierte Wörter           |
| #\$p                    | markiert eine Pause   | Stille und Sprechpausen (gefüllte und stille)             |
| _v                      | stimmhafte Anteile bei einem stimmlosen Phonem                              |   |
| _0                      | stimmlose Realisierung eines stimmhaften Phonems                            |   |
| _~                      | Nasalisierung   |   |
| _\<br>_\'               | Friktion bei nicht frikativischen Phonemen                                  |   |

**Tabelle 3: Symbole zur Markierung von Abweichungen**

## 2.4 Erweiterbarkeit des Symbolinventars

Bei der Definition des deutschen Symbolinventars wurde darauf geachtet, dass die für englische und französische Lehnwörter wichtigsten Laute zur Verfügung stehen. Um eine Erweiterbarkeit des Inventars und multilinguale Synthese zu ermöglichen, steht in BLF die Länderkennung zur Verfügung. Dabei handelt es sich um das in runde Klammern gesetzte Länderkürzel entsprechend dem jeweiligen Toplevel-Domainnamen im Internet, das an das zugehörige Lautsymbol angehängt wird. Für verschiedene Sprachen oder Dialekte innerhalb eines Landes kann außerdem noch ein Zahlencode angehängt werden. Beispiel: leicht eingedeutschte Aussprache von engl. „brother“: br(uk)aD(uk)6

## 3 Diskussion und Ausblick

Durch die Zusammenfassung von mehreren Lauten zu einem Segment wird die forcierte Segmentierung zeitlich nicht voneinander abzugrenzender Phonemrealisierungen vermieden, was eine Qualitätssteigerung in der automatischen Lautsegmentierung und in der Unit-Selection-basierten Synthese bewirken sollte. Aus der verbesserten automatischen Segmentierbarkeit ergibt sich wiederum ein Nutzen für die Synthese, weil die Anzahl falsch gesetzter Segmentgrenzen auch an nicht direkt betroffenen Positionen im Signal sinkt. Dies ist deshalb der Fall, weil eine nicht korrekt gesetzte Segmentgrenze auch die Setzung der anderen Grenzen in einem automatisch segmentierten Signal negativ beeinflusst.

Die Spezifikation in ihrem momentanen Stadium berücksichtigt aus mehreren Gründen nur Lautfolgen von Konsonanten im Silbenanlaut und Vokalen im Silbenn Kern, obwohl auch an anderen Positionen nicht zufriedenstellend segmentierbare Lautkombinationen auftreten. Wort- und silbenübergreifende Folgen sind zum einen aus implementierungstechnischen Gründen nicht möglich, zum anderen würde durch die Kombinatorik die Anzahl der möglichen Folgen und damit Labelsymbole übermäßig ansteigen. Betrachtet man allein die

Kombination der streng nach deutscher Phonologie und Phonotaktik möglichen silbeninternen Lautfolgen, kommt man auf 210 Symbole. Auch wenn in deutschen Wörtern vielleicht nicht alle Kombinationen vorkommen, so werden diese Folgen spätestens bei der Transkription von Eigennamen relevant. Die bereits jetzt in der Spezifikation enthaltenen Symbole sollten ausreichen, um die größten Schwierigkeiten zu umgehen. Von den betrachteten Anfangskonsonanten kommt ohnehin nur das /l/ in der Silbenkoda vor. Eine Aufnahme des /l/ kommt für spätere Versionen in Betracht, allerdings würde der Umfang des Symbolinventars dadurch um 369 Einheiten expandieren. Die jetzt schon integrierten Anfangskonsonanten sind für die Verständlichkeit und die Qualität der synthetischen Sprache ohnehin wichtiger als die in silbenfinaler Position auftretenden Laute. Ebenfalls aus Gründen der Kombinatorik bilden die Plosive bisher keinen Verbund mit Folgevokalen. Es wird sich zeigen, ob eine solche Zusammenfassung notwendig ist.

## Literatur

Eisen, B.; Tillmann H.G. (1992): "Consistency of Judgements in Manual Labelling of Phonetic Segments: The Distinction between Clear and Unclear Cases", Proceedings of ICSLP 1992, Banff.

Klabbers, E.; Stöber, K.; Veldhuis, R.; Wagner, P.; Breuer, S. (2001): "Speech synthesis development made easy: The Bonn Open Synthesis System", EUROSPEECH 2001, Aalborg, Denmark.

Kohler, K.; Pätzold, M.; Simpson, A. (1995): "From scenario to segment. The controlled elicitation, transcription, segmentation and labelling of spontaneous speech." AIPUK 29, Kiel.

Pompino-Marschall, B. (1992): "PHONDAT. Verbundvorhaben zum Aufbau einer Sprachsignaldatenbank für gesprochenes Deutsch", Forschungsbericht des IPSK München (FIPKM) 30.

Stöber, K.; Breuer, S.; Wagner, P.; Abresch, J. (2000): *Dokumentation zum Bonn Open Synthesis System (BOSS) II*. Unveröffentlichtes Dokument, IKP Bonn.

Wahlster, W., Hrsg. (2000): *Verbmobil: Foundations of Speech-to-Speech Translation*, Symbolic Computation, Berlin: Springer.

Wells, J. C. (2000): *SAMPA - computer readable phonetic alphabet*.  
<http://www.phon.ucl.ac.uk/home/sampa/home.htm>