



Produktbeschreibung

# PROSER-SDK 3.5

advanced technologies for information processing

ATIP GmbH  
Daimlerstraße 32  
60314 Frankfurt am Main  
<http://www.atip.de>

## PROSER

PROSER ist eine diphonbasierte Sprachsynthese (TTS) für Deutsch und Englisch mit hoher Sprachqualität. Die von PROSER erzeugte Sprache hat nichts mehr zu tun mit den bekannt blechernen "Computerstimmen" früherer Jahre. Durch den Einsatz von Methoden der künstlichen Intelligenz und der neuronalen Netze kombiniert mit konkatenativer Signalsynthese wird eine Sprachqualität erreicht, die bereits nahe an der Performance von Nachrichtensprechern liegt. Weiter gibt es komfortable Möglichkeiten, die Stimm- und Prosodiecharakteristik gezielt zu beeinflussen.

## Funktionsweise von PROSER

Die Informationsverarbeitung in PROSER gliedert sich in Symboltransformation, Prosodiegenerierung, Verkettung und akustische Synthese. Die Symboltransformation beinhaltet eine Umwandlung von Text in linguistisch-phonetische Basiselemente – auch als Transkription bezeichnet – die mit prosodischen Annotationen angereichert sind. Eine Stufe dieser Umwandlung besteht in der Textvorverarbeitung, bei der z.B. Abkürzungen, Zeitangaben etc. in die zu sprechenden Wortäquivalente aufgelöst werden. Danach erfolgt wortweise die Transkription in eine phonetische Repräsentation, der Annotationen in Form von Silbenstruktur und Silbenbetonung hinzugefügt sind.

Die Transkription erfolgt durch Kombination von Aussprachelexikon und neuronalem Netz. Hierdurch wird erreicht, dass zu jedem Wort eine Transkription geliefert wird und dennoch nutzerpräferierte Aussprachen realisiert werden können. Prosodiegenerierung bezeichnet die Berechnung des Betonungsprofils. Die Prosodieparameter werden mit einer neuartigen Methode der silbenbasierten Unit Selection in Kombination mit rekurrenten neuronalen Netzen bestimmt.

Aus der diskreten symbolischen Repräsentation entsteht durch Verkettung eine Parameterfolge zur Ansteuerung des akustischen Synthetisators, der aus dem patentierten Verfahren MBROLA besteht.

## Systemaspekte

PROSER basiert auf einem modularen Software-Konzept, das die einfache Portierbarkeit auf unterschiedlichste Plattformen unterstützt. Derzeit sind lauffähige Versionen, als Bibliothek (lib, so, dll) oder als ausführbares Programm, für die Betriebssysteme MS Win9x, NT, 2000, XP und Linux verfügbar. Kurze Programmlaufzeiten wurden durch optimierte Datenstrukturen und effiziente Implementierung erreicht. Ein mehrkanaliger Betrieb wird unterstützt, wobei auf einem PC mit PIII 500 MHz 80 voneinander unabhängige Kanäle gleichzeitig betrieben werden können.

Der Bedarf an Speicher ergibt sich aus dem, der für den Programmcode benötigt wird, der vom Lexikon beansprucht wird und der für mindestens eine Stimmdatei vorzuhalten ist. Während der Laufzeit wird für den ersten Sprachkanal ein Speicher von unter 10 MByte benötigt. Da es sich hier um shared memory handelt, steigt für jeden weiteren Sprachkanal der Speicherbedarf nur noch geringfügig an.

Der Speicherbedarf pro Stimme hängt von diversen Randbedingungen wie Sprache, Bandbreite, Geschlecht etc. ab. Beispielsweise wird für eine deutsche Frauenstimme mit 16 kHz Abtastrate ca. 8 MByte Speicher und mit 8 kHz Abtastrate (Telefonsprache) 4 MByte.

## Möglichkeiten der Prosodiesteuerung

Die Möglichkeiten, das Erscheinungsbild einer Stimme oder Sprechweise zu verändern, sind vielfältig. Die Auswahl einer Stimmdatenbank entscheidet zunächst über den grundsätzlichen Charakter der Stimme. Es können verschiedene weibliche oder männliche Stimmen ausgewählt werden. Hierbei ist es sogar möglich fremdsprachliche Stimmdatenbanken einzusetzen, um z.B. einen amerikanischen, türkischen oder französischen Akzent zu erzielen. Sprechmelodie und -tempo können sowohl in ihrer mittleren Lage als auch in ihrer Dynamik justiert werden. Bei Bedarf können auch sehr feine Justierungen, bis hin zu den Dauerparametern von Einzellauten vorgenommen werden. Die wichtigsten Parameter, z.B. für Sprechtempo und -melodie, Pausensetzung, Aussprache und Betonung, sind direkt mit in den Text eingebetteten Kontrollmarken in VoiceXML oder SAPI einstellbar.

## Beeinflussung der Aussprache

Die Aussprache von Wörtern ist definiert durch die zugehörige Lautfolge, die mit prosodischen Markierungen versehen wird. Erzeugt wird die zu einem Wort gehörende Lautfolge durch eine Kombination von Lexikon und neuronalem Netz. Das neuronale Netz liefert für jedes beliebige Wort eine phonetische Transkription. Diese kann unter Umständen nicht korrekt sein, oder es wird eine andere Aussprache gewünscht. Daher ist dem neuronalen Netz ein Lexikon vorgeschaltet, in dem explizit phonetische Transkription vorgegeben werden können.

Das Lexikon ist aufgliedert in einen nutzer- bzw. anwendungsspezifischen Teil (User-Lexikon) und einen Basisteil (System-Lexikon), die separat verwaltet werden. Die höchste Priorität für eine Transkription besitzt das User-Lexikon. Durch diese Priorisierung beim Erstellen einer Transkription ist sichergestellt, dass zum einen die vom Nutzer gewünschten Aussprachevarianten realisiert werden und zum anderen eine vollständige Abdeckung aller Wörter vorhanden ist.

Um die häufig in deutschen Texten auftretenden englischen Wörter möglichst korrekt auszusprechen, wurde eine englische Lexikonkomponente in PROSER aufgenommen. Ist diese Komponente aktiviert, werden englische Wörter in einer für das Deutsche angepassten Lautung transkribiert und synthetisiert.

## PROSER-SDK 3.5 bietet:

### Features

- annähernd natürliche Sprachqualität
- weibliche und männliche Stimmdatenbanken
- einfacher Zugriff auf phonetische und prosodische Parameter
- intelligente Textvorverarbeitung
- unbeschränkter Wortschatz
- definierbares Aussprachelexikon
- leichte Skalierbarkeit
- verschiedene Audioformate

### Benefits

- hohe Akzeptanz beim Endkunden
- personalisierbare Sprachausgabe
- gezielte Beeinflussung der Stimmcharakteristik
- akkurate Textumsetzung
- vollständige Sprachabdeckung durch Diphonbasis
- korrekte Aussprache von Eigennamen und Abkürzungen
- mehrkanaliger Betrieb bei geringen Systemressourcen
- an unterschiedliche Audiosituationen anpassbar