Erstellung einer Sprachdatenbank

sowie

eines Programms zu deren Analyse im Kontext einer Sprachsynthese mit spektralen Modellen

Tobias Platen

Gliederung

1. Einleitung

2. Untersuchung spektraler Modelle im Hinblick auf Eignung für Sprachsynthese

3. Erstellung der Sprachdatenbank

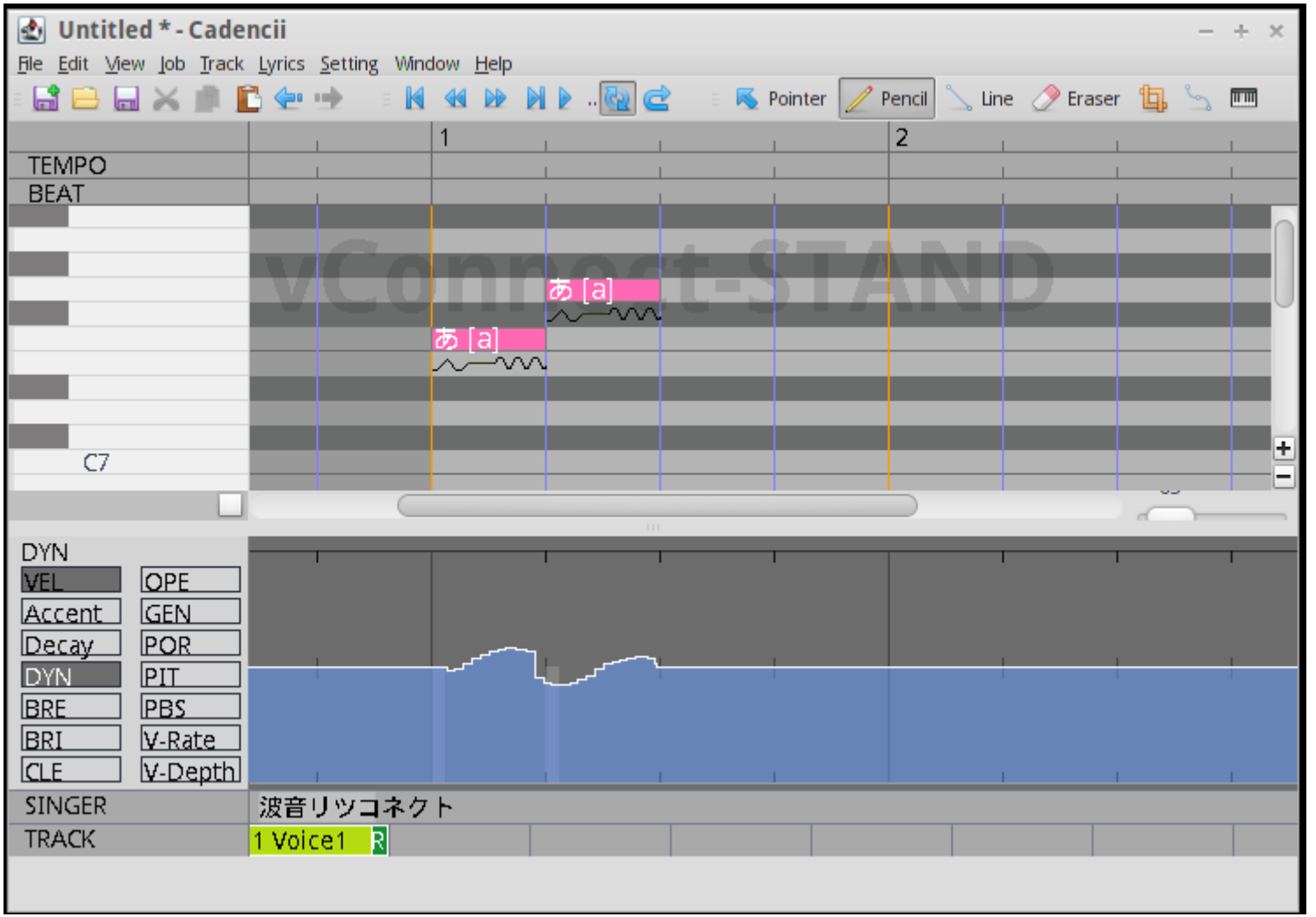
4. Ergebnisse

5. Ausblick

1. Einleitung

- **Motivation**
- **Ziele**
- **Wege und Methoden zum Ziel**

Cadencii


Untitled * - Cadencii
File Edit View Job Track Lyrics Setting Window Help
Pointer
Pencil
Line
Eraser
1
2
TEMPO
BEAT
あ [a]
あ [a]
C7
DYN
VEL
Accent
Decay
DYN
BRE
BRI
CLE
OPE
GEN
POR
PIT
PBS
V-Rate
V-Depth
SINGER
波音リツコネクト
TRACK
1 Voice1
R

Spectral Model Synthesis

Funktioniert durch H/S Zerlegung von Signalen

Vorteile des Modells:

- ermöglicht Transformationen der Klangfarbe
- effiziente Resynthese mit IFFT-Methode

Nachteile:

- verwirft Phaseninformation
- Referenzimplementierung fehlerhaft

Das WORLD Speech Toolkit

Funktioniert durch Zerlegung in drei Teile:

- DIO: fundamentale Frequenz
- STAR: spektrale Hüllkurve
- PLATINUM: Anregungssignal

Vorteile:

- Sehr gute Sprachqualität
- Freie Software unter der BSD-Lizenz

Verbesserung von eSpeak durch WORLD

- Integration in den speech-dispatcher
- Generierung der Prosody mit eSpeak
- Schreiben des Parsers für MBROLA pho-files
- Entnahme von Samples aus einer Datenbank
- Implementierung der Echtzeitsynthese
- Test der konkatenativen Sprachausgabe

Konkrete Schritte zur Erstellung der Sprachdatenbank

- Corpus-Design kompatibel zu MBROLA de2
- Rendern der Prompts
- Aufnahme der Sprachsamples
- Beschriften der Samples
- Spektrale Analyse
- Verpacken der Datenbank
- Produktivtest durch Endnutzer

Die Sekai Speech Tools

selbstgeschriebene grafische Programme:

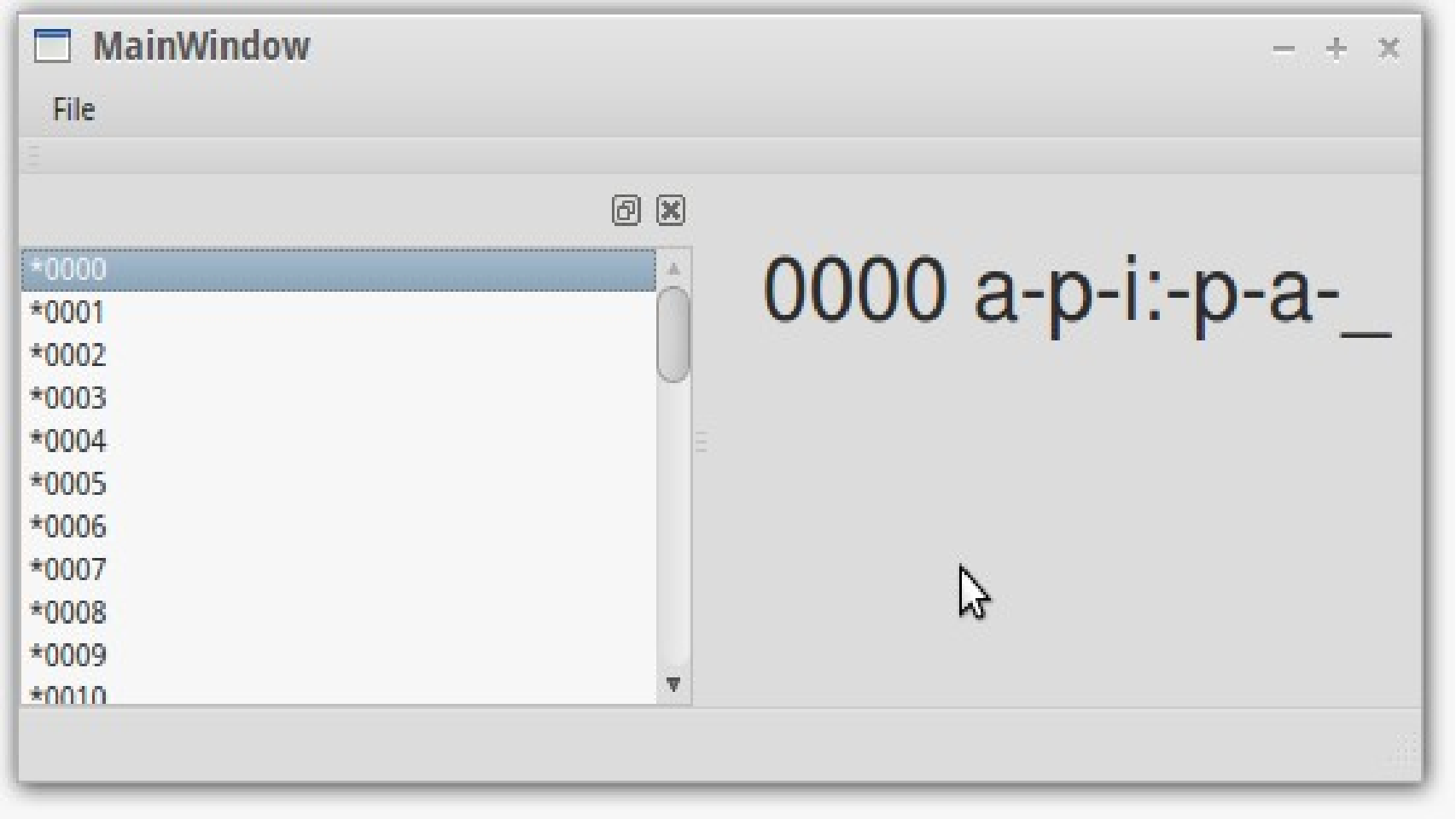
- Oremo2: Echtzeitrecording mit Jack
- OtoEdit: Visuelle Sample-Segmentierung
- SekaiViewer: Visualisierung der Sprachdaten

Batch-processing der einzelnen Samples

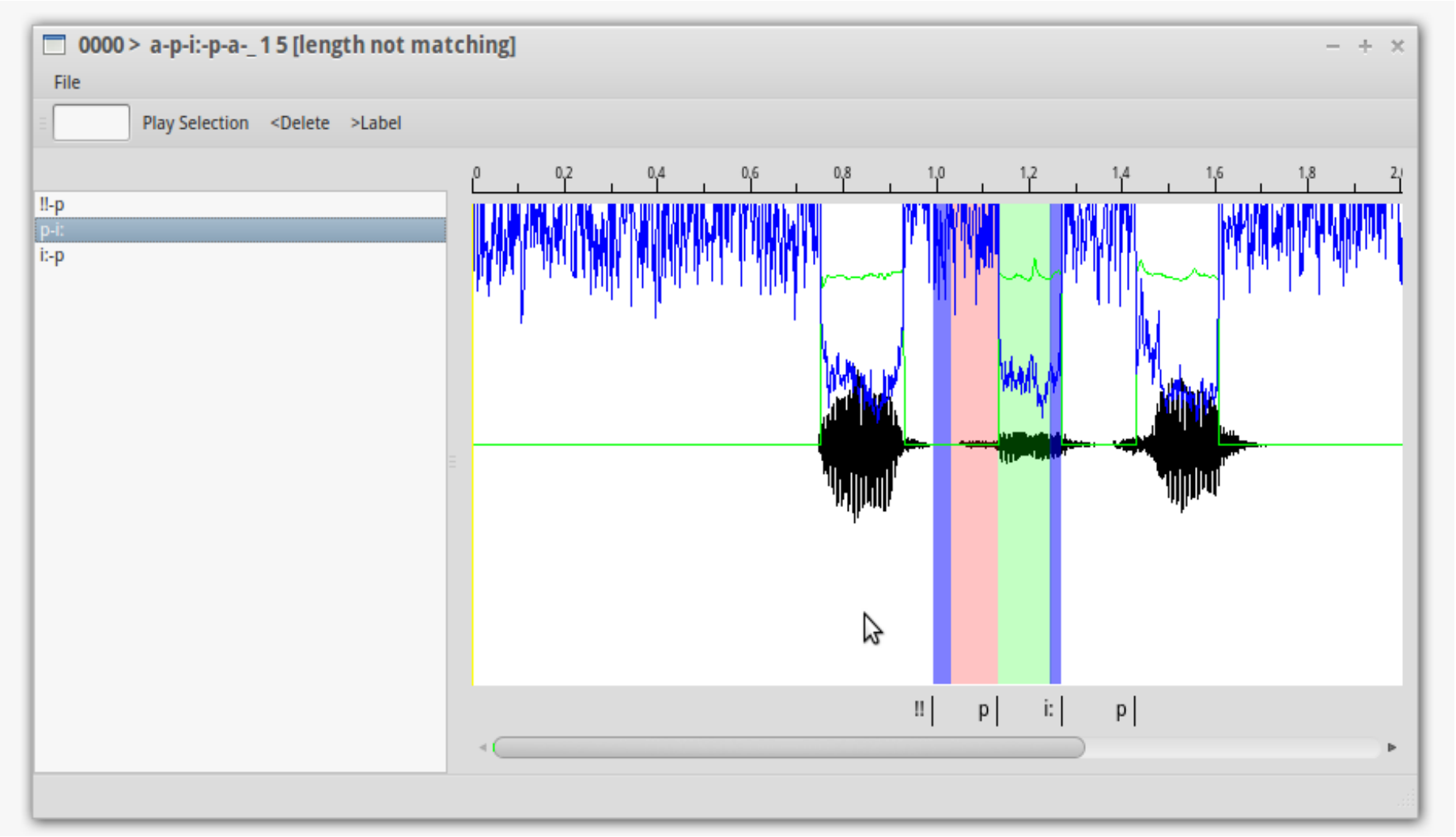
- temporäre Dateien sehr groß
- Vorbis und MFCC zur Kompression

Test der Konkatenation und Kompression

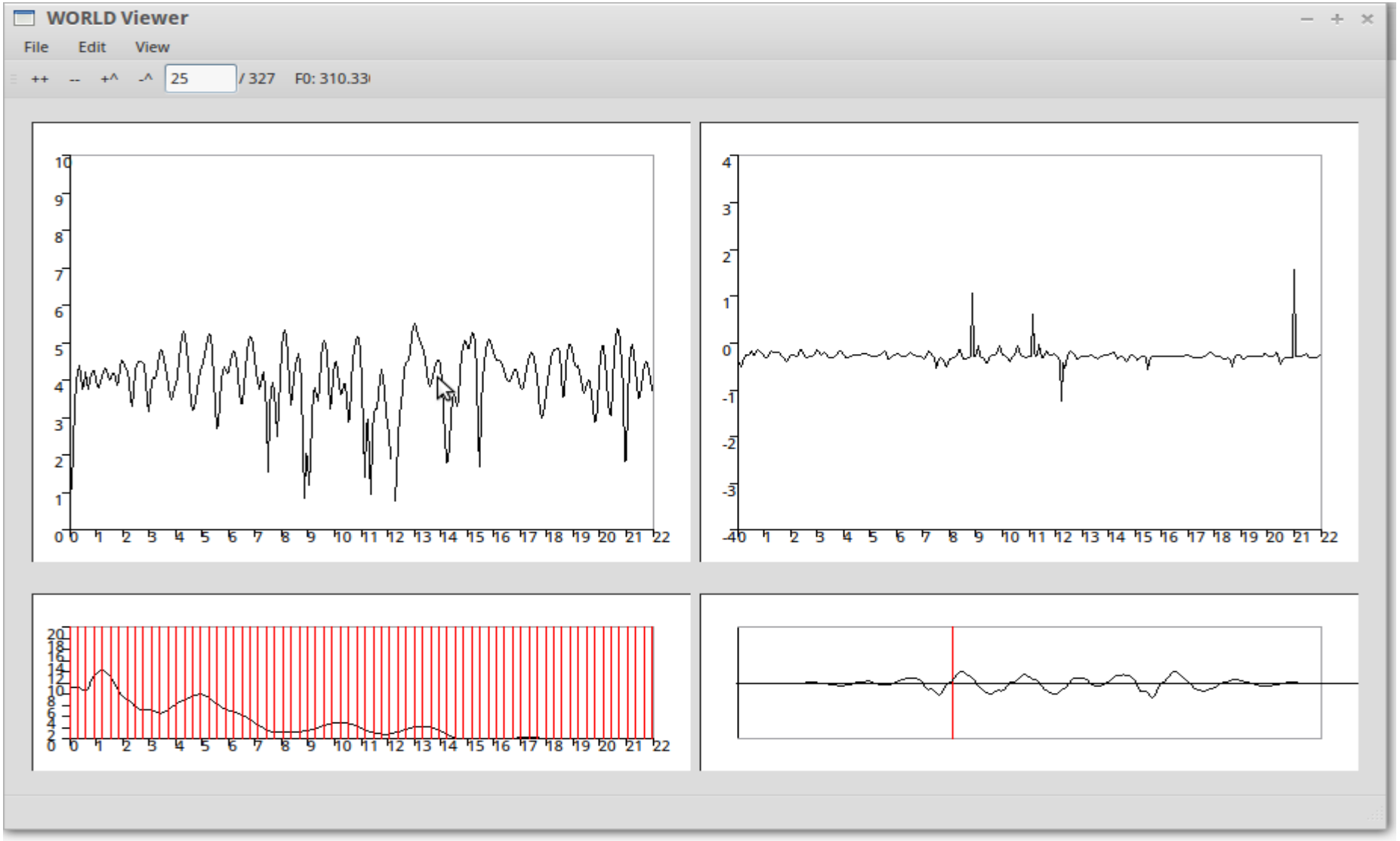
Oremo2


MainWindow
File
*0000
*0001
*0002
*0003
*0004
*0005
*0006
*0007
*0008
*0009
*0010
0000 a-p-i:-p-a-_

OtoEdit


0000 > a-p-i:-p-a-_ 1 5 [length not matching]
File
Play Selection
<Delete
>Label
!!-p
p-i:
i:-p
0
0,2
0,4
0,6
0,8
1,0
1,2
1,4
1,6
1,8
!!
p
i:
p

SekaiViewer



Ergebnisse

- Vorteile von Vocoder und TD-PSOLA nutzbar
- automatische Bestimmung der GCIs gelöst
- Nutzung von spektralen Stimmenmodellen sinnvoll
- realistische Formantbewegungen jetzt möglich
- Tests zeigen Eignung der Algorithmen
- besserer Klang auch gegenüber Festival

Ausblick

nötig:

- weitere Tests der Sprachdatenbank

sinnvoll:

- Refactoring von eSpeak
- Implementation von EpR und IFFT-Methode
- Ergänzung weiterer Phoneme
- Spracherkennung mit HMMs und DTW
- Integration der HTS-Sprachausgabe
- Bessere Modellierung des Anregungssignals
- Verbesserung der Echtzeitsynthese

Fazit

Vielen Dank für Ihre Aufmerksamkeit

Noch Fragen?