

# Speech Based Classification (SBC) Demonstrator Guide

Bearbeiter: Felix Burkhardt

Letzte Bearbeitung: 5.5.2010

Dokument Version: 1.0

Software Version: 1.0

## *Inhalt*

Speech Based Classification (SBC) Demonstrator Guide.....	1
Inhalt .....	1
Überblick .....	1
Vorführszenario .....	2
Vorbereitung .....	2
Beispielablauf Emotionserkennung.....	2
Beispielablauf Alters/Geschlechtserkennung .....	3
Problembehebung .....	3
Anpassung an die Präsentationsumgebung .....	4
Grundlagen .....	5
Technische Grundlagen.....	5
Was ist Sprachbasierte Klassifikation? .....	5
Was kann es leisten? .....	5
Business Opportunities .....	6
Wie gut ist die Erkennung? .....	6
Konfiguration des Demonstrators.....	6

## *Überblick*

Der SBCDemonstrator analysiert die Stimme / Sprechweise des Vorführers und ordnet diese ENTWEDER einer Emotion ODER einer Alters/Geschlechtsklasse zu.

Es gibt zwei Emotionsklassen:

- „Nicht Ärgerlich“
- „Ärgerlich“.

Es gibt sieben Alters/Geschlechtsklassen:

- „Kind“ (bis 14),
- „Jugendliche“,
- „Jugendlicher“ (bis 25),
- „Erwachsene“,
- „Erwachsener“ (bis 55),
- „Seniorin“,
- „Senior“.

Entscheidend für die Klassifizierung sind alleine Lautstärke, Rhythmus, Melodie und Stimmklang, NICHT der Wortlaut.

Sie klicken auf das Mikrofon, sprechen eine kurze Äußerung und warten dann ab, bis das Erkennungsergebnis dargestellt wird. Bei der Emotionserkennung leuchte eine rote bzw. grüne Lampe, bei der Alterserkennung erscheint ein Prototyp der jeweiligen Altersklasse.

Das System ist an die Umgebungsgeräusche anpassbar. Falls die automatische Mikrofonabschaltung nicht funktioniert, klicken Sie einfach noch einmal auf das Mikrofon.







Die Emotionserkennung ist vom Vorführer trainierbar, aber NICHT die Alterserkennung.


## Vorführszenario

### Vorbereitung






- 1) Starten Sie den Demonstrator durch Klicken der Datei „startSBCDemo.bat“.
- 2) Passen Sie den Demonstrator an die Umgebung an (siehe nächster Abschnitt).

### Beispielablauf Emotionserkennung

1) Beginnen Sie mit Emotionserkennung, achten Sie darauf, dass in der unteren linken Ecke ein hellgraues „E“ dargestellt ist.			
2) Klicken Sie einmal auf das Mikrofon (es soll dann angestrahlt werden) und sagen Sie in ruhigem Tonfall „ <i>Ich sage jetzt mal irgendwas</i> “.			
3) Im Erfolgsfall geht das Mikrofon bei Stille von selber wieder aus und eine grüne Lampe leuchtet für zwei Sekunden.			
4) Sagen Sie jetzt in ärgerlichem Tonfall „ <i>Ich sage jetzt mal irgendwas</i> “, indem Sie z.B. laut und abgehackt sprechen.			
5) Die rote Lampe sollte leuchten.			
6) Um zu beweisen, dass es sich nicht um eine reine Lautstärkeschwelle handelt, sagen Sie jetzt in freudigem Tonfall „ <i>Ich sage jetzt mal irgendwas</i> “, indem Sie z.B. laut und jubelnd sprechen.			

7) Die Lampe sollte grün leuchten.	
------------------------------------	--

### Beispielablauf Alters/Geschlechtserkennung

1) Schalten Sie auf Alterserkennung um, indem Sie in der unteren linken Ecke auf das „E“ klicken (es sollte dann zu „A“ werden).	
2) Sagen Sie in ruhigem Tonfall „Ich sage jetzt mal irgendwas“.	
3) Ein Prototyp Ihrer Alters/Geschlechtsgruppe sollte erscheinen.	
4) Sprechen Sie jetzt mit verstellter Stimme (z.B. als Mann sehr hoch oder als Frau sehr tief).	
5) Eine andere Alters/Geschlechtsklasse sollte erscheinen.	

### Problembehebung

Das Mikrofon bleibt nicht aktiv	Wenn das Mikrofon sofort wieder ausgeht, verringern Sie den Geräuschpegelwert (Anpassung) oder erhöhen die Mikrofon Wartezeit,
Das Mikrofon schaltet sich nicht aus.	Wenn es nach dem Sprechen weiter an bleibt, erhöhen Sie den Geräuschpegelwert (oder klicken einfach jedes Mal zum beenden auf das Mikrofon).
Es findet keine / keine richtige Erkennung statt.	Kontrollieren Sie das eingegangene Audiosignal durch Klicken mit RECHTS auf das Mikrofon.

	Wenn die Lautsprecher funktionieren, sollte die zuletzt aufgenommene Äußerung zu hören sein.
Meine Emotion wird nicht erkannt.	Üben Sie zwei akustisch deutlich verschiedene Äußerungen. Nehmen Sie diese dann einzeln auf und klicken jeweils nach dem Leuchten der Lampe links oben, wenn Sie Ärger zeigen wollten oder rechts oben, wenn Sie Neutral zeigen wollten. Nach etwa 2, 3 Beispielen pro Zustand, klicken Sie mit der rechten Maustaste im oberen Bereich um den Konfigurator zu öffnen, und dann auf „train model“.
Meine Altersgruppe wird nicht erkannt.	Die Alterserkennung ist leider nicht trainierbar, Kontrollieren Sie das eingehende Audiosignal durch Klicken mit RECHTS auf das Mikrofon.

### *Anpassung an die Präsentationsumgebung*

Durch Rechtsklick im oberen Bereich öffnen Sie den Konfigurator, der in Abbildung 1 dargestellt ist..

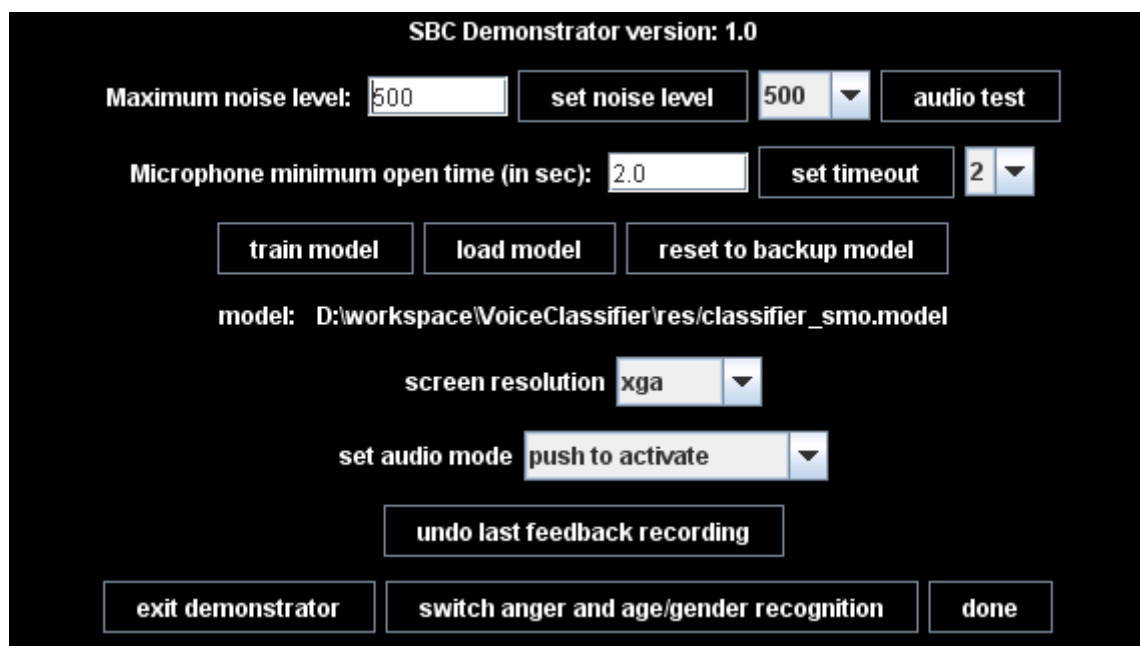


Abbildung 1: Screenshot des Konfigurators

Es folgt eine zeilenweise Aufzählung der Einstellmöglichkeiten.

- **Noise Level:** Setzen Sie hier den maximal zu erwartenden Geräuschpegel für die automatische Stilleerkennung. Entweder direkt durch Eintippen, durch Selektion der vorgegebenen Möglichkeiten oder automatisch („audio test“) durch 5 Sekunden Einschalten des Mikrofons. **DIESER WERT SOLLTE EUF JEDEN FALL ÜBERPRÜFT WERDEN.** Sie Können ihn einfach testen indem Sie de Audiomodus „permanent recording,, stellen und dann den Wert verändern.
- **Microphone Open Time:** Setzen Sie hier dir minimale Zeit, die das Mikrofon bei automatischer Stilleerkennung geöffnet bleibt. Entweder direkt durch Eintippen, der vorgegebenen Möglichkeiten.
- **Modell Training:** Der SBC Demonstrator basiert technisch gesehen auf dem Vergleich Ihrer Eingaben mit einem statistischen Modell welches aus den Audiomeerkmalen berechnet wurde.

Sie haben hier die Möglichkeit, dieses Modell nachzutrainieren, d.h. ihre aufgenommenen Äußerungen in das Modell aufzunehmen. Dies funktioniert nur für die Emotionserkennung und nur, wenn Sie zuvor mindestens eine Äußerung durch Klicken auf eine der beiden Lampen gespeichert haben. Sie können alternativ ein bereits vorhandenes Modell laden, bzw. zum Auslieferungszustand zurück wechseln („reset to background model“ = Panikknopf). DER PANIKKNOPF LÖSCHT ALLE GESPEICHERTEN AUFNAHMEN. Das jeweils aktive Modell wird angezeigt.

- **Screen Resolution:** Stellen Sie hier die passende Monitorgröße für Vollbilddarstellung ein.
- **Set Audio Mode:** Es gibt drei Modi der Aufnahmesteuerung:
  - o **Push to talk:** Aufnahme während Gedrückt-Halten des Mikrofons (Walkie Talkie), danach startet die Bewertung.
  - o **Push to activate:** Einmaliges klicken des Mikrofons aktiviert die Aufnahme, die automatische Stilleerkennung ODER ein zweiter Klick auf das Mikrofon beenden die Aufnahme und starten die Bewertung.
  - o **Permanent Recording:** Wie **Push to activate**, aber eine neue Aufnahme startet automatisch nach erfolgreicher Bewertung oder Timeout, d.h. der Aufnahmemodus wird erst durch Klicken auf das Mikrofon beendet. Einmaliges Klicken aktiviert diesen Modus.
- **Undo last feedback recording:** Klicken Sie hier, um die zuletzt gespeicherte Aufnahme zu löschen (und somit vom Training auszuschließen).

## Grundlagen

### Technische Grundlagen

Der Demonstrator basiert, technisch gesehen, auf der Extraktion von akustischen Merkmalen und darauf basierender statistischer Klassifikation.

Zur Merkmalsextraktion wird derzeit das open-source Paket OpenSMILE der Technischen Universität München verwendet.

<http://www.openaudio.eu/>

Als Merkmale werden jeweils für die Gruppen Intensität, Grundfrequenz, Spektrum und weitere eine Reihe von Äußerungsglobalen (= ein Wert pro Äußerung) Werten berechnet. Beispiele sind Mittelwert, Steigung, Anzahl der stimmhaften Perioden etc.

Insgesamt werden über 1000 Werte pro Äußerung ausgewertet.

Als Klassifikator findet die open-source library WEKA der Universität von Waikato in Neuseeland Verwendung.

<http://www.cs.waikato.ac.nz/~ml/weka/>

Der Verwendete Klassifikator basiert auf dem „Support Vektor Machine“ Prinzip. Danach bildet jede zu unterscheidende Klasse einen Teil im N-dimensionalen Merkmalsraum (N ist die Anzahl der Merkmale). Da das SVM Verfahren nur binär arbeitet, muss für jede Klasse die Wahrscheinlichkeit gegen alle anderen Klassen berechnet werden.

### Was ist Sprachbasierte Klassifikation?

Sprachbasierte Klassifikation ordnet Anrufer bestimmten Sprechergruppen zu.

Beispiele sind Alters-, Geschlechts-, Sprachen- oder Emotionserkennung, aber grundsätzlich kann jede menschliche Eigenschaft, die auf Sprachsignal Auswirkungen hat, klassifiziert werden.

### Was kann es leisten?

- **Marktforschung:** Wie viele Frauen interessieren sich für Produkt X. Wieviele Kunden waren nach der Preissteigerung verärgert?
- **Vorqualifizierung:** Welche meiner Kunden sprechen Englisch?

- **Dialogadaption:** Angepasste Hilfe-prompts für ältere Kunden. Stark verärgerte Kunden vor dem Auflegen mit Agenten verbinden.
- **Gaming:** Liebt mich mein Freund? Wie hoch ist mein "stimmliches" Alter?

## Business Opportunities

- Zufriedenere Kunden durch Dialogadaption.
- Steigerung der Verkaufszahlen durch gezielte Angebote.
- Optimierung der Produktwerbung durch Kundenanalysen.
- Innovative neue Angebote im Gaming Sektor

## Wie gut ist die Erkennung?

Die Exaktheit der Erkennung hängt stark von der Anwendung ab.

Einige Faustregeln:

- Das menschliche Unterscheidungsvermögen bildet eine gute Abschätzung: Maschinen haben sehr ähnliche Ergebnisse.
- Je weniger Klassen, desto höher die Genauigkeit: Männer von Frauen zu trennen ist einfacher als das Alter auf's Jahr genau.
- Je mehr Sprachmaterial, desto besser: ganze Dialoge sind einfacher als ein "Hallo?" von einer Sekunde Dauer.
- Je stärker sich die Eigenschaft in der Stimme ausdrückt, desto besser: Fremdsprachen sind leichter zu erkennen als das Alter. Beispiele:
  - o Geschlechtererkennung bei längeren samples: nahezu 100%
  - o Ärgererkennung fast ohne Fehlalarme: cirka 40% Verbesserung.

## Konfiguration des Demonstrators

Der Demonstrator ist ein Java-Programm mit angeschlossenen nativen Hilfsanwendungen zur Analyse der Audiosignale und Klassifikation der Merkmale.

Das Hauptprogramm ist ein Java Archiv namens „VoiceClassifier.jar“

Im Ordner „tools“ liegen die Hilfsprogramme.

Im Ordner „res“ liegen Ressourcen wie z.B. Bilddateien.

Der Demonstrator wird in seinen Voreinstellungen durch eine Textdatei konfiguriert.

Sie heißt „sbcDemo.properties“ und liegt im Ordner „res“.

Es folgt ein Auszug dieser Datei:

```
# screen resolution: vga 640x480, svga 800x600, xga 1024x768, sxga 1200x1024, wsxga+
1680*1050, uxga 1600x1200
resolution=xga
# extension of audio files
audioFormat=wav
# more debug (log level in log config file)
garrulous=true
# title string of configurator
titleString=. Fachl. Ansprechpartner Felix Burkhardt, DT-Labs, Tel: 49-15116710189
# audio sample rate
sampleRate=16000
# Threshold for automated silence detection to stop recording automatically.
# Raise for noisy environments.
silenceThreshold=500
# Timeout for automated silence detection in sample vales.
# I.e. if sample rate=16000 a value of 8000 means half a second.
speechTimeout=8000
# Timeout for automated silence detection in sample vales.
# This is the value before the user starts speaking.
initialTimeout=32000
# time to wait till result disappears in milliseconds
waitTime=2000
# If true, microphone has to be pressed to talk (walkie-talkie)
# If false, microphone has to be clicked to start recording.
pushToTalk=false
# if true, microphone is permanently open (after click)
# and will analyze speech between pauses (as defined by speechTimeout)
```

```

permanentRecording=false
    # if true, positive or negative feedback
    # for last recording can be given
feedback=true
    # if retraining is possible
training=true
    # time in seconds to listen for maximal audio value
    # while calibrating microphone
calibrationTime=5
    # whether feature extraction before model building starts from scratch
    # or extracts only files that have no prediction yet.
additiveTraining=true
    # label to category mapping: pairs of category descriptors
    # assigned to minimum labels, e.g. 1,N;2,NA means category "N"
    # is assigned for values between 1>=x<2.
    # MUST be in ascending order!
categories=2,N;3,A
    # classifier type, smo, j48 or naiveBayes
classifier=smo

    # identifier for label-lines in annotation file
labelIdentifier=LABELS:
    # identifier for transcription-lines in annotation file
transcriptIdentifier=TRANSCRIPTION:
    # identifier for prediction-lines in annotation file
predictionIdentifier=PREDICTION:
    # extension for audio-accompanying annotation file
labelFileExtension=txt

#### pathes
ressourceDir=res
recordingDir=recordings/
testAudioFile=tmp/test.wav
logConfig=./res/logConfigDemonstrator.xml
tmpFeatFile=tmp/tmpFeat.txt
tmpTxtFile=tmp/tmp.txt
tmpWavFile=tmp/tmp.wav
trainFile=res/train.arff
arffBackupFile=res/angerTrain.arff
testFile=res/test.arff
modelFile=res/classifier.model
modelBackupFile=res/classifierAnger_smo.model
recordings=recordings/
openEarCommand=./tools/SMILExtract.exe
openEarConfig=res/IS10.anger_funcs.conf

# gui labels
gui.title.start=SBC Demonstrator version:
gui.noiseLevel.label=Maximum noise level:
gui.noiseLevel.button=set noise level
gui.audioTest.button=audio test
gui.timeout.label=Microphone minimum open time (in sec):
gui.timeout.button=set timeout
gui.screenResolution.label=screen resolution
gui.switchMode.button=switch anger and age/gender recognition
gui.done.button=done
gui.trainModel.button=train model
gui.loadModel.button=load model
gui.resetModel.button=reset to backup model
gui.undoLastStorage.button=undo last feedback recording
gui.changeAudioMode.label=set audio mode
gui.changeAudioMode.pushToTalk=push to talk
gui.changeAudioMode.pushToActivate=push to activate
gui.changeAudioMode.permanentRecording=permanent recording
gui.exit.button=exit demonstrator
# images
angerRecognition=true
agenderRecognition=false
angerImage=res/images/e.gif
agenderImage=res/images/a.gif
greenImage=res/images/black_04_green.jpg
redImage=res/images/black_03_red.jpg
startImage=res/images/black_01_start.jpg
activeImage=res/images/black_02_spot.jpg
okImage=res/images/ok.gif
cImage=res/images/gui_boy.jpg
yfImage=res/images/gui_woman_young.jpg
ymImage=res/images/gui_man_juvenile.jpg
afImage=res/images/gui_woman_middle_age.jpg
amImage=res/images/gui_man_middle_age.jpg
sfImage=res/images/gui_grandma.jpg
smImage=res/images/gui_man_older.jpg

```