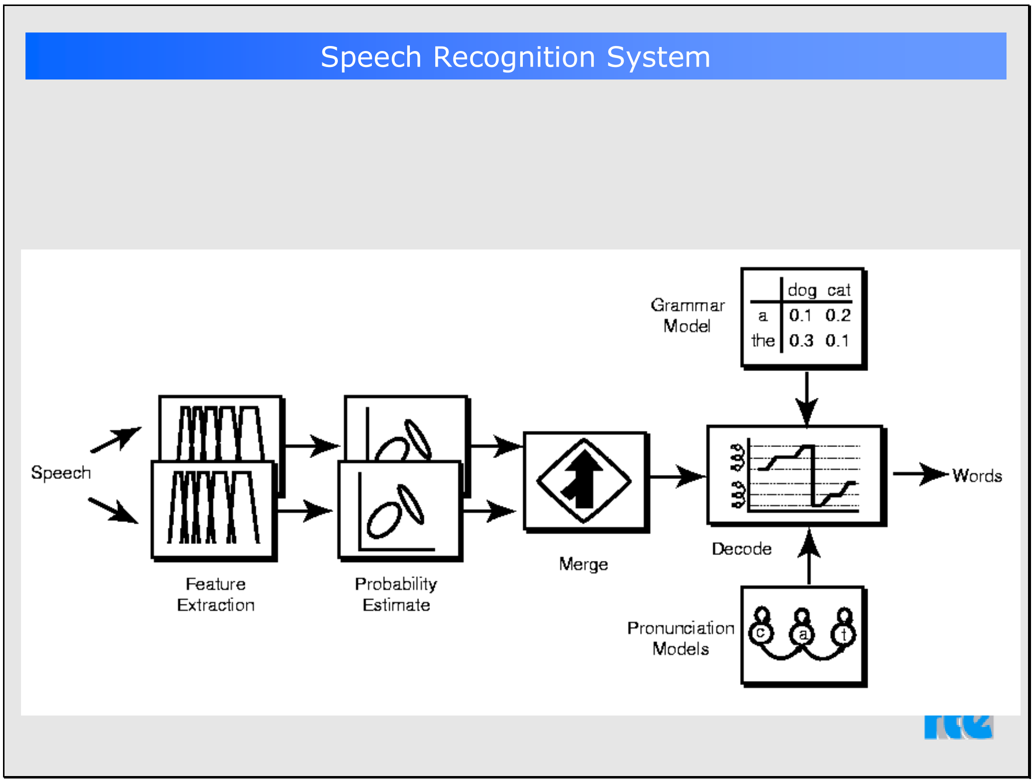
Verfahren der Spracherkennung: neue Möglichkeiten in der industriellen Anwendung der akustischen Qualitätssicherung

Christoph Lauer, RTE Akustik + Prüftechnik, Pfinztal

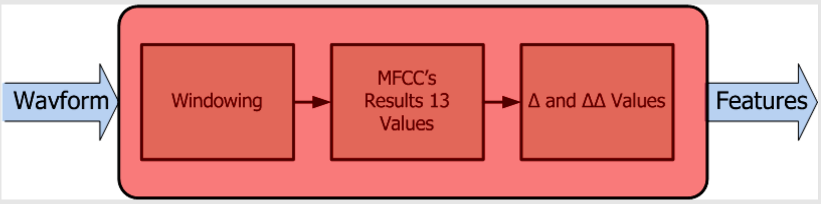
1 Einführung in die Spracherkennung

- ❒ Feature Extraction - Physik
- ❒ Training / Decoding - Stochastik
- ❒ Grammatik – Linguistik



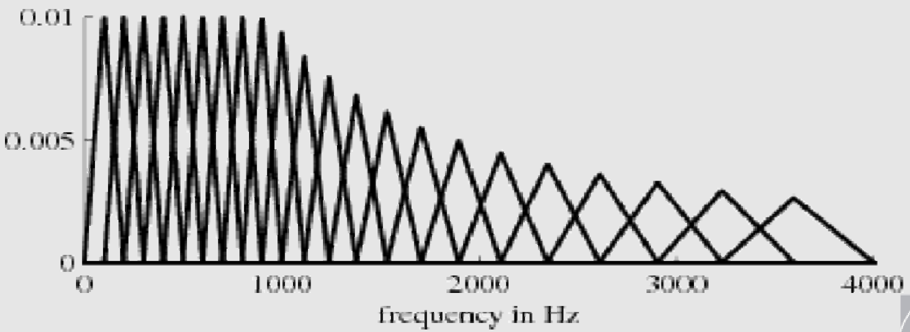
Feature Extraction

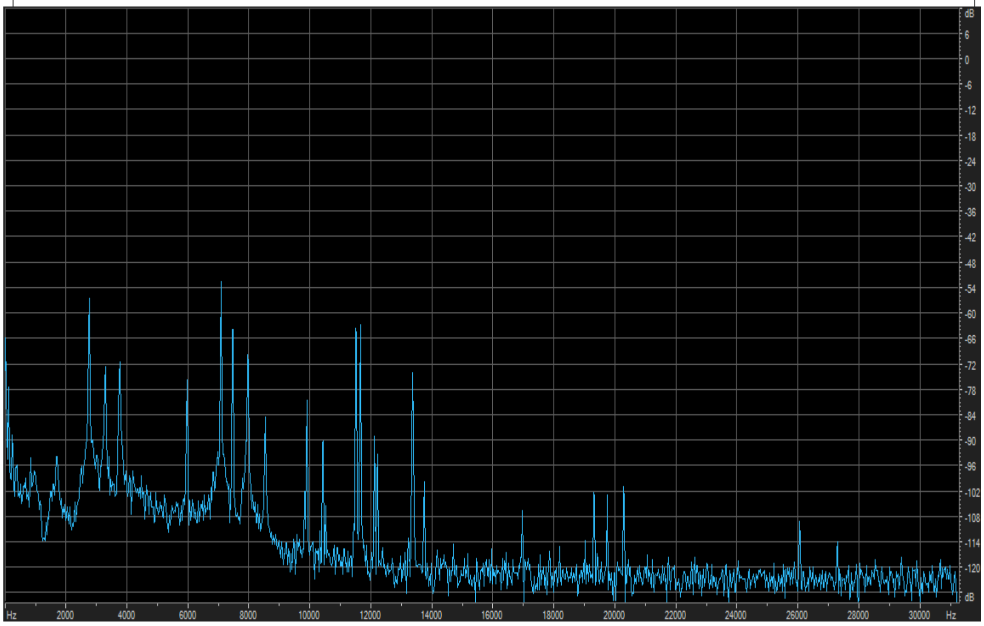
- Mel-Frequency-Cepstral-Coefficients (MFCC's)
- Typischerweise 12 Bänder plus Gesamtenergie = 13 Features alle 10ms
- macht mit Δ und Δ Δ = 39 Features



Feature Extraction

- Standard Feature Extraction bei der Sparacherkennung ist dem Frequenzverlauf des menschlichen Ohrs angepasst





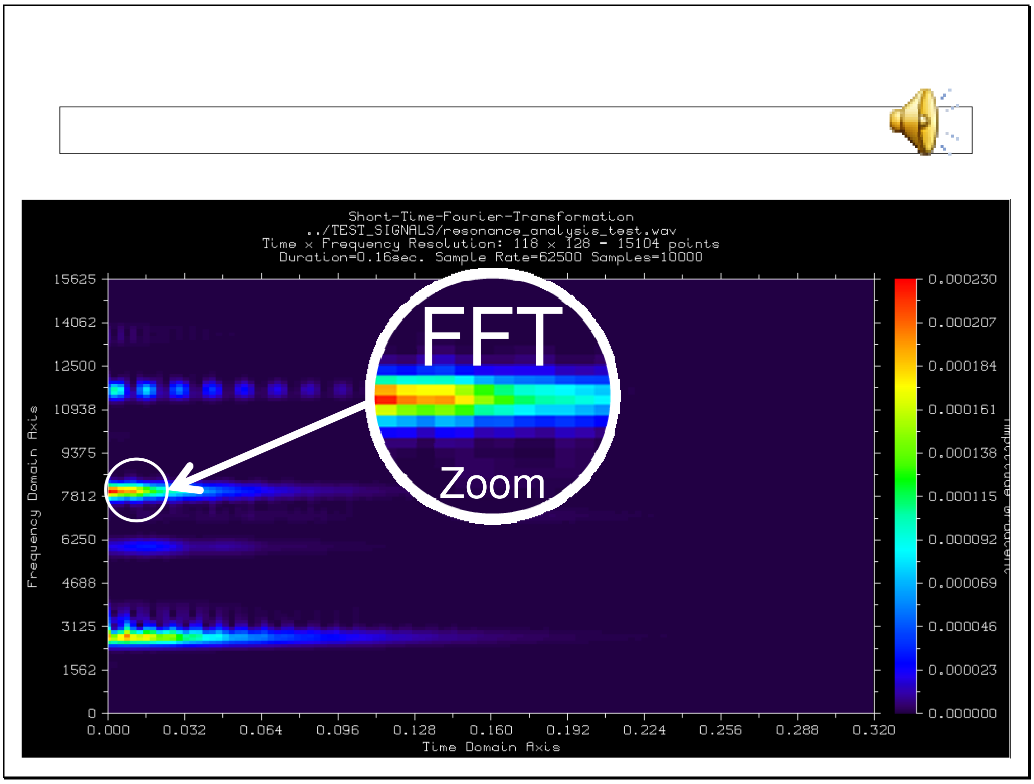
Feature Extraction

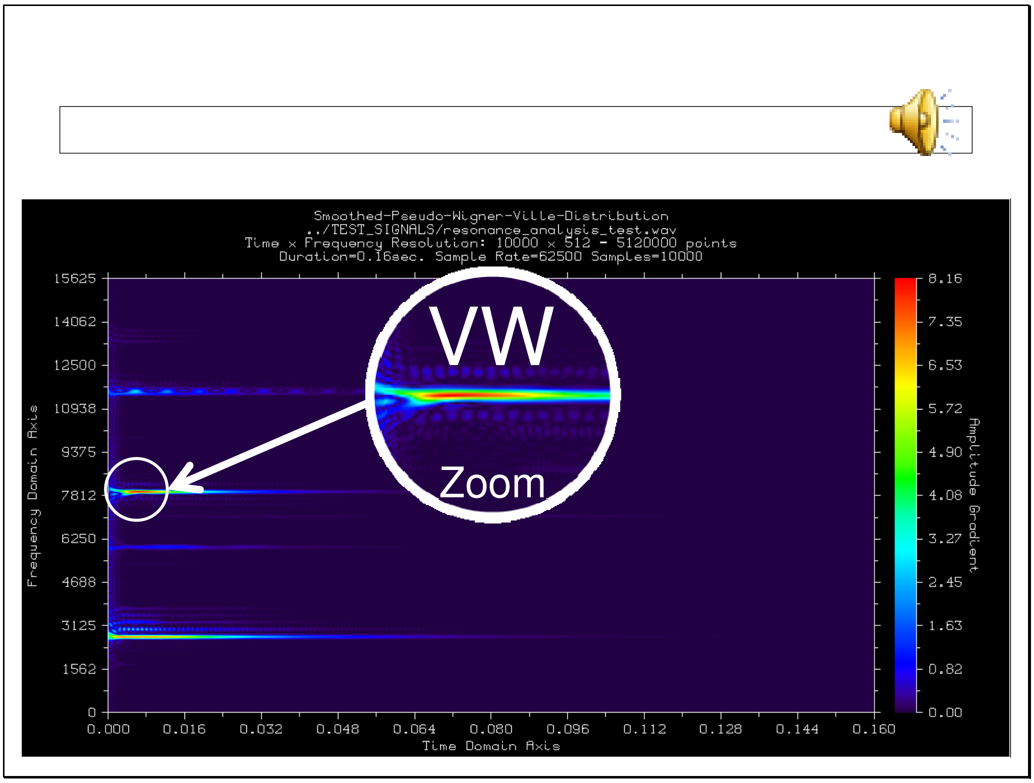
- Anpassung der Feature Extraction
- Akustische Unschärferelation:

$$\Delta f = \frac{1}{\Delta t}$$

  100 Hz Auflösung bei 10 ms !!! → Zu wenig für die Anwendung in der Resonanzanalyse

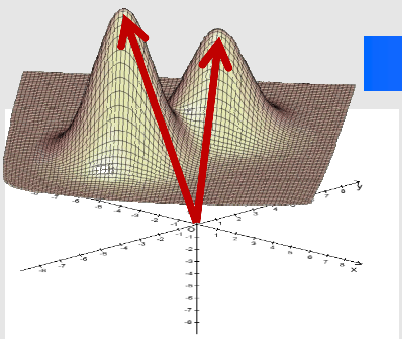
- →Entwicklung alternativer Verfahren.


Short-Time-Fourier-Transformation
../TEST_SIGNALS/resonance_analysis_test.wav
Time x Frequency Resolution: 118 x 128 = 15104 points
Duration=0.16sec. Sample Rate=62500 Samples=10000
FFT
Zoom
Frequency Domain Axis
Time Domain Axis


Smoothed-Pseudo-Wigner-Ville-Distribution
../TEST_SIGNALS/resonance_analysis_test.wav
Time x Frequency Resolution: 10000 x 512 = 5120000 points
Duration=0.16sec. Sample Rate=62500 Samples=10000
VW
Zoom
Frequency Domain Axis
Time Domain Axis
Amplitude Gradient

Feature Extraction

- Die verbesserte Zeit-Frequenz-Analyse mit der Wigner-Ville-Analyse kann nun gezielt zur Feature-Extraction in der Resonanzanalyse genutzt werden.
- Geringe Frequenzabweichungen in kleinen Bändern können nun wahrgenommen werden.

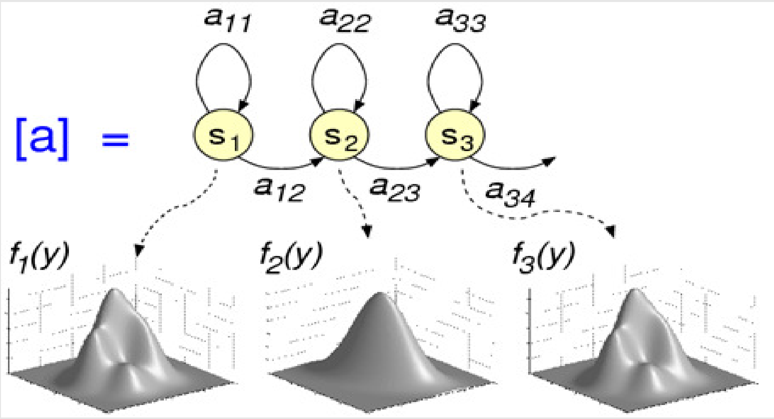


Klassifikation

- Zwei Wörter, IO/NIO werden auf N-dimensionale Gaußglocken projiziert.
- Die Parameter der Gaußglocke (Mean und Varianz) werden beim Training ermittelt.
- Bei der Erkennung (Decoding) wird geschaut, in welche Klasse (Wort) der Feature-Vector zeigt.

Klassifikation

- Die Klassifizierung des Feature-Vectors in Zustandsräume geschieht nun für jedes Fenster mit sog. **Hidden Markov Modellen**.



Klassifikation

- Diese Art der Klassifikation wird **GMM/HMM**-Klassifikation genannt.
- HMM Training mit dem Viterbi-Algorithm/Baum-Welch.
- GMM Training Forward-Backward Algorithm.
- Wegen des einfachen Sprachmodels mit nur zwei Wörtern wird in unserem Fall kein linguistisches Model benötigt

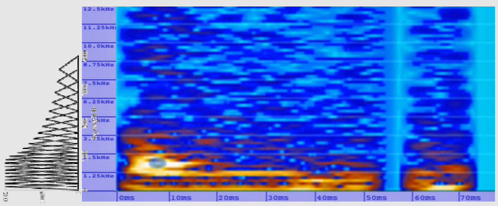
Implementierung

- Erster Test mit der Standard MFCC Feature Extraction.
- Decoder/Trainer implementiert in ANSI C/C++.
- Akustische Modelle 3-State HMM‘s

2 Anwendung in der Akustischen Geräusch- / Resonanzanalyse

Anwendung

- Problem: Erkennung von hohlem Gemüse an der Impulsantwort.
- Herkömmliche Verfahren erfolglos wegen zu wenig Klassifizierungsdifferenz
- Spektrum im Bereich der menschlichen Sprache, deshalb kann die Standard MFCC-Feature-Extraction verwendet werden.



3 Anwendungsergebnisse

Anwendung

Non Destructive Acoustical Classification of woody and hollow Asparagus with Methods derived from the Speech Technology

Christoph Lauer
RTE Acoustical Measurement and Testing Technology - Pfinztal
www.rte.de c.lauer@rte.de

September 11, 2008

Abstract

Non-destructive classification of woody or hollow asparagus with acoustical methods is still an unsolved problem, because the elastic modulus of fruit and vegetables is not large enough for the extraction of sufficient resonancy eigenfrequences, like it has been a well-known practice in the metal-working industry for decades. Coming from speech technology, we try to use methods well known in Natural-Language-Processing like Hidden-Markov-Models (HMM's) and Gaussian-Mixture-Models (GMM's) for the automatic classification of fruit and vegetables. In this paper, we describe our implementation and present our first lab results.

1 Introduction

Extraction of the acoustic resonance eigenfrequencies has been a well-known practice in the product testing environment at the end of the manufacturing processes for decades. A reasonable acoustic impulse response extraction is only possible for a Device Under Test *(following named DUT)* with a high elastic-modulus, where the resonance frequency can simply be extracted by mechanical stimulation with a solid-borne sound sensor or an air microphone. The whole impulse response will be transformed into the frequency-domain, and fluctuations of the own-frequencies seen in the transfer function can be used to classify the broken parts. All phase and frequency-localisation information are lost during the extraction of the magnitude spectrum. In most cases the DUT's consist of metal, ceramic, glass or another composite material which has a very high elastic-modulus of more than 50 kN/mm^2 which results in a very recognizable characteristic set of resonance eigenfrequencies. In our case we have the problem that fruits and vegetables have wides span of geometric variations, and the impulse response has no typical characteristic like a metal or ceramic DUT because of the lack of an elastic-modulus. So we decided to look into time domain variations, and use a hybrid of Hidden-Markkov-Models and Gaussian-Mixed-Models as classification base for a windowed set of input features, similar to the models used to classify phonemes in speech recognition.

2 Implementation

Our impulse response classificator implementation is based on Hidden Markov Models. The extracted model parameters from the training can then be used to perform further analysis of the impulse responses and allow us to classify between the hollow and not-hollow asparagus. We will now give a short description of our lab setup.

2.1 Physical Feature Extraction

Because the frequency range of the transfer-function matches the mel-scale very well *(see Figure 1)*, we decided to use standard Mel-Frequency-Cepstral-Coefficients (MFCC's) as front-end for the training and the decoding of the GMM/HMM hybrid. The front-end transforms the impulse response into a set of features, to be used for training and decoding of the acoustic model. After a standard FIR pre-emphasis filter is applied to the input signal, a window will be cut and multiplicated with a Hamming function. This window will be transformed into the frequency domain, and the mel spectrum of this power spectrum is computed by multiplying the power spectrum by each of the 40 triangular mel weighting filters. Finally the mel cepstrum is extracted by applying a discrete cosine transformation to the natural logarithm of the mel

1

Asparagus	Condition	N	Corr.	Rel.
Asparagus 13	NOK	95	90	94.44%
Asparagus 17	NOK	96	78	81.25%
Asparagus 19	NOK	150	143	95.33%
Asparagus 21	NOK	96	76	73.68%
Asparagus 30	NOK	96	80	80.00%
Asparagus 37	NOK	96	92	95.65%
Asparagus 41	NOK	96	88	91.66%
Asparagus 52	NOK	84	51	60.71%
Asparagus 64	NOK	96	50	52.08%
Asparagus 65	NOK	96	30	31.25%
Asparagus 66	NOK	96	93	96.87%
Asparagus 67	NOK	96	77	80.21%
Asparagus 68	NOK	96	64	66.66%
Asparagus 69	NOK	96	30	68.74%
Asparagus 70	NOK	96	93	96.88%
Σ	15	1481	1075	71.58%

Asparagus	Condition	N	Corr.	Rel.
Asparagus 7	OK	96	90	93.75%
Asparagus 8	OK	96	58	60.41%
Asparagus 9	OK	96	75	78.12%
Asparagus 14	OK	96	79	82.29%
Asparagus 15	OK	96	77	80.21%
Asparagus 16	OK	96	84	87.50%
Asparagus 20	OK	96	91	94.80%
Asparagus 22	OK	96	95	99.95%
Asparagus 24	OK	96	83	86.45%
Asparagus 26	OK	96	66	68.75%
Asparagus 31	OK	96	90	93.75%
Asparagus 32	OK	96	91	94.79%
Asparagus 33	OK	96	36	37.50%
Asparagus 35	OK	96	60	62.25%
Σ	14	1344	1075	79.85%

Erkennungsrate > 90%

Anwendung

- Problem: Erkennen von kratzenden Schreibfedern
- Klassische Spektralanalyse erfolglos.
- Spektrum in ähnlichem Bereich wie die menschliche Sprache → Standard MFCC Feature-Extraktion.
- Erkennungsrate > 90%