Analysis of Voice-Onset Using Active Rays and Hidden Markov Models

O. Jesorsky, J. Denzler, E. Nöth, T. Wittenberg
Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg
Martensstr. 3, D–91058 Erlangen, Germany
email: orjesors@cip.informatik.uni-erlangen.de

orjesors@cip.informatik.uni-erlangen.de

The following paper will appear in the proceedings of
2^{nd} 'Round Table':

Advances in Quantitative Laryngoscopy using Motion-, Image- and Signal Analysis

Erlangen, Germany, July $18^{th} - 19^{th}$, 1997

Analysis of Voice-Onset Using Active Rays and Hidden Markov Models

O. Jesorsky, J. Denzler, E. Nöth, T. Wittenberg
Lehrstuhl für Mustererkennung (Informatik 5)
Universität Erlangen–Nürnberg
Martensstr. 3, D–91058 Erlangen, Germany
email: orjesors@cip.informatik.uni-erlangen.de

Abstract

In this paper we describe an approach for an automatic evaluation system for laryngoscopical image sequences. We describe how glottis segmentation in high speed camera sequences can be done by using active rays. We show that active ray contour segmentation provides a good base for useful feature calculation.

Afterwards we show how hidden Markov models (HMMs) can serve for a statistical evaluation of the given feature vector sequences. The main advantage of the HMM approach is the fact, that HMMs are able to supply both for classification and time segmentation of images of voice onsets showing functional disorders. There exist several algorithms for parameter estimation and the models are robust against single missegmentations.

Finally we describe how we can adapt the existing speech recognition system "ISADORA" for our purposes.

Keywords: glottography, active rays, hidden Markov models

1 Introduction

A big supposition for classification of functional disorders of the voice is knowledge about the vocal fold vibration. Because of the high base frequencies of this vibration (between 80 and 400 Hz) it is not possible to recognize motions in real-time. The most frequent way to cope with this problem is to work with laryngoscopes in combination with a stroboscopic light source and video recording [8]. But with this kind of examinations it is only possible to observate steady state vibrations.

High speed cameras, able to take up to 10000 frames per second provide a good possibility to analyze even non periodic vocal fold vibrations. In our work we use image sequences of a high speed camera system developed at the Fraunhofer Institute for Integrated Circuits in Erlangen in cooperation with the department of Phoniatrics of the University Erlangen-Nuremberg. This system is described in [8].

Given such images the next step is analysis of the video stroboscopic or high speed image sequences. Examination by an experienced doctor provides no quantitative results. Computer based examinations of the movement of the vocal folds are shown in [6]. An active contour based approach is used to detect position and shape of the vocal folds. Our approach is to describe position and movement of the vocal folds by segmentation of

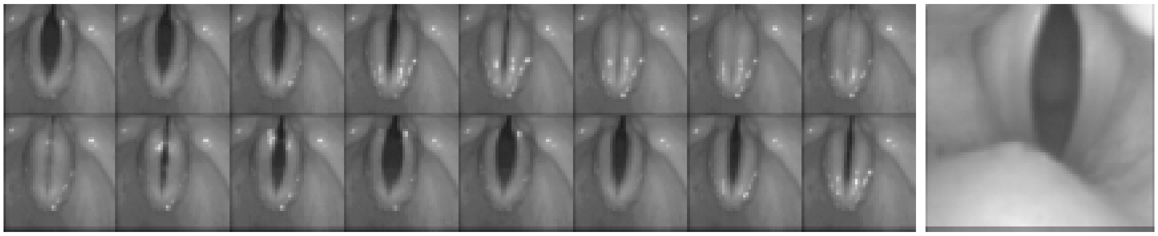

Figure 1: High speed glottography image sequence and single frame

the gap between them, the glottis. Detection is done by active rays [1], an active contour like approach to contour tracking. Glottis segmentation is also done at the department of Phoniatrics at University Erlangen-Nuremberg. But therea region growing algorithm is used to detect the glottis area [8].

The result of the segmentation by active rays is a 1D feature vector sequence. In our work those vectors are directly evaluated statistically by HMMs. There exist a lot of robust algorithms for parameter estimation and evaluation. An advantage for us is the fact that HMMs can be used for classification as well as for time segmentation of feature vector sequences. In speech processing HMMs have been used since the early 70's. Theory of these models is described in [5, 3, 7] and [4].

In our case we adapt the speech recognition system "ISADORA" [7], which has been been developed at the Chair for Pattern Recognition at University Erlangen-Nuremberg since 1989, for our purposes.

2 Images

The High-Speed-Glottography-System we are working with provides a full frame resolution of 128 by 128 pixels with a maximum speed of 1025 frames per second. Recording speed can be increased by reading only every second, fourth or eighth row from the camera's CCD–Chip.

In this case the resolution of the images is increased up to 128 by 128 pixels before the segmentation step by repeating existing lines, so that the following process is independent of the chosen camera resolution. For detailed information about the camera system see [8].

Figure 1 shows an image sequence and a single frame produced by such a system. The most interesting part for us, the glottis, corresponds to the dark area in the middle of the single picture framed by the vocal chords left and right. At the bottom of the image you can see the epiglottis which closes the respiration tract during swallowing.

3 Functional Voice Disorders

One task of phoniatrics is recognition and distinction of functional voice disorders (Disphonies). A field of special interest is differentiation between hypo- and hyper-functional disorders. The phase of beginning phonation, the voice onset, provides characteristic features for this two cases [2].

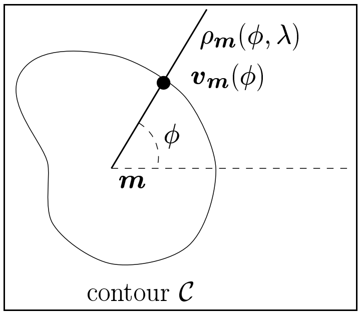
$\rho_{\boldsymbol{m}}(\phi, \lambda)$
$\boldsymbol{v}_{\boldsymbol{m}}(\phi)$
ϕ
$\boldsymbol{m}$
contour $\mathcal{C}$

Figure 2: Principle of one active ray.

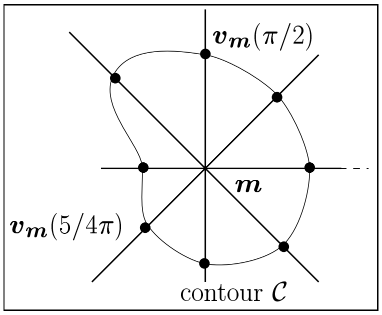
$\boldsymbol{v}_{\boldsymbol{m}}(\pi/2)$
$\boldsymbol{m}$
$\boldsymbol{v}_{\boldsymbol{m}}(5/4\pi)$
contour $\mathcal{C}$

Figure 3: Representation of a contour by active rays.

The voice onset can be departed into four different phases closing of the vocal folds, prephonetic glottis closure, beginning vibration and steady state vibration.

Duration of prephonetic closure and beginning vibration phases can indicate functional disorders [8].

In our work we want to build up a system which automatically evaluates high speed image sequences of vocal fold vibrations and is able to classify hypo- and hyper-functional disorders.

4 Active Rays

To detect the interesting sections in the image sequences we need information about the moving of the vocal folds. Therefore we are looking for points on the outer borders of the glottis. Active rays provide a good means to detect those points. This section is based on description of active rays in [1].

An active ray $\varrho_{\boldsymbol{m}}(\phi, \lambda)$ is defined on the image plane (x, y) as a 1D function depending on those gray values $f(x, y)$ of the image, which are on a straight line from the image point $\boldsymbol{m} = (x_m, y_m)^T$ in direction ϕ

$$\varrho_{\boldsymbol{m}}(\phi, \lambda) = f(x_m + \lambda \cos(\phi), y_m + \lambda \sin(\phi)), \quad 0 \leq \lambda \leq n_\phi, \tag{1}$$

where n_ϕ is given by the image size. The principle is clarified in Figure 2. The angle ϕ is measured counter clockwise.

Now, a contour point in direction ϕ regarding a given reference point $\boldsymbol{m}$ can be described by the parameter $\lambda(\phi) \geq 0$

$$\lambda(\phi) = \underset{\lambda}{\operatorname{argmin}} \left(- \left| \frac{\partial}{\partial \lambda} \varrho_{\boldsymbol{m}}(\phi, \lambda) \right|^2 \right), \quad 0 \leq \phi < 2\pi, \tag{2}$$

i.e., we are looking for points on the active ray with a maximum edge strength. The contour point $\boldsymbol{c}_{\boldsymbol{m}}(\phi)$ (see Figure 2) is then

$$\boldsymbol{c}_{\boldsymbol{m}}(\phi) = (x_m + \lambda(\phi) \cos(\phi), y_m + \lambda(\phi) \sin(\phi)), \quad 0 \leq \phi < 2\pi \tag{3}$$

A similar representation is used by the generalized Hough transform. In the discrete case the whole contour can be computed by defining a sampling step size $\triangle\phi$ for ϕ. This allows for different accuracy of the contour representation. An example for a representation of a contour is shown in Figure 3. The sampling step size $\triangle\phi$ is $\pi/4$.

Now, we have to discuss the choice of the reference point $\boldsymbol{m}$. In principle, every point within the object's contour is possible. But to have a unique point, which can be precalculated by a prediction step, the center of gravity of the contour extracted by the active ray is used in the following, i.e., the equation

$$\boldsymbol{m} = 1/2\pi \int_0^{2\pi} \boldsymbol{c}_{\boldsymbol{m}}(\phi)\, d\phi \tag{4}$$

should hold for the reference point $\boldsymbol{m}$. For convex contours $\boldsymbol{m}$ will also be the center of gravity of the object's contour. What happens, if the chosen reference point is not the center of gravity? Then, we can calculate a new reference point using the formula (4). After that, the new contour representation has to be calculated.

We have noted that the approach of active rays allows for multiple hypotheses. For this, we have to look for the i best solutions of equation (2), which means, that for each ray in direction ϕ we get a set $\Lambda(\phi)$

$$\Lambda(\phi) = \left\{ \lambda_k(\phi) | \lambda_k(\phi) = \underset{\lambda, \lambda \neq \lambda_l, l < k}{\operatorname{argmin}} \left(- \left| \frac{\partial}{\partial \lambda} \varrho_{\boldsymbol{m}}(\phi, \lambda) \right|^2 \right), 0 \leq k < i \right\} \tag{5}$$

of possible solutions for the contour instead of one single contour element.

We can build up a glottis segmentation system with the assumptions (according to [8]) that in its open state, the glottal area can be modelled as one uniform, dark segment, the darkest area in each single frame is assumed to be part of the glottal frame, the glottis can be seen in every single frame and the gray value of the darkest point in a frame showing a opened glottis is less than the gray value of the darkest point in a frame showing a closed glottis.

The system includes the following steps for every single frame. At first, detection of the area with minimum gray value is done. Based on this area, the decision whether the glottis is closed or not, i.e. whether the frame is to be processed or not is made. Then, if the frame is to be processed, the glottis contour is extracted by active rays. Fianlly the edge points are selected out of the calculated hypotheses.

For the computation of the area with minimum gray value we are moving a 3x5 window over the whole frame. The sum of the gray values serves as an energy function. The coordinates of the energy minimum are used as reference point m for the active rays.

The value of the energy minimum can serve as a basis for decision whether the glottis is closed or not. Frames showing a closed glottis need not to be processed anymore. The value is compared with the average minimum energy value of the last 20 frames. We assume glottis closure when the actual value is higher than the average value and vice versa (see assumptions).

This works well when there is a vibration of the vocal folds. When we start to process a new frame sequence, we have to precalculate the energy minimums of the first 20 frames, to get a reliable average value. If we have to process a frame, the coordinates of the minimum energy are chosen as the reference point m of the active rays.

We chose the image gradient $-\left|\frac{d}{d\lambda}\varrho_{\boldsymbol{m}}(\phi, \lambda)\right|^2$ as our energy function. With this function and the reference point m we calculate three hypothesis per ray, i.e. we choose the three strongest edge points on each ray in our set of hypotheses $\Lambda(\phi)$. Because there is a strong edge at the glottis border, we assume that this edge point is included in the set. We chose the edge point with minimum $\lambda(\phi)$, that means minimum distance from the reference point m as contour point.

We chose those distances $c_i = \lambda(\phi_i)$ as parts of our feature vector o when examining i rays.

Figure 4 shows how the feature vectors o are calculated out of the images by using active rays .

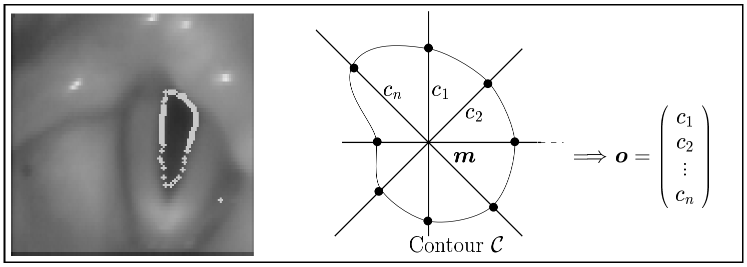
c_n
c_1
c_2
$\boldsymbol{m}$
Contour $\mathcal{C}$
$\Longrightarrow \boldsymbol{o} = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{pmatrix}$

Figure 4: Extraction of feature vectors using active rays

5 Hidden Markov Models

In the section before we have shown how to get feature vectors o out of each single frame. Now we have to find a system to evaluate feature vector sequences $O = o_1 \cdots o_i \cdots o_n$ computed out of an image sequence. A statistical solution is provided by the HMMs.

In this section we will describe the theory of the HMMs, how we can use them for classification and time segmentation of laryngoscopical image sequences and how we can adapt the existing speech recognition system "ISADORA" for our purposes.

Please note that both, active rays and HMMs use the mathematical symbol λ in their descriptions. So please don't mix up the lambdas in this section with the lambdas in previous sections.

5.1 Hidden Markov Models - Theory

The behavior of a HMM λ can be described by a finite automaton with states $S = \{S_1, S_2, \cdots, S_N\}$ and transition probabilities $a_{ij} = P(s_t = S_j | s_{t-1} = S_i)$ with $a_{ij} \geq 0$ and $\sum_{j=1}^{N} a_{ij} = 1$ for $1 \leq i \leq N$. Those transition probabilities can be embraced in the matrix $A = (a_{ij})$. The vector $\Pi = (\pi_i)$ with $\pi_i = P(s_1 = S_i), \qquad 1 \leq i \leq N$ includes the starting probabilities.

To simulate time dependent processes we limit the transition probabilities by $a_{ij} = 0, \forall i > j$. This secures that a state which has been left can not be reached again. Those models are called "Left–Right–Models". In addition we force the process to start in state S_1 by setting $\pi = (1, 0, \cdots, 0)^T$. Because of the restrictions state S_N never can be left due to $a_{NN} = 1$.

Each time t the automaton takes a new state (even if it's the same state as at time $t-1$) it produces a symbol o_t out of the finite alphabet $O = \{O_1, O_2, \cdots, O_K\}$ with production probabilities $b_{jk} = b_j(O_k) = P(o_t = O_k | s_t = S_j)$ for $1 \leq j \leq N, 1 \leq k \leq K$ with $b_{jk} \geq 0$ and $\sum_{k=1}^{K} b_{jk} = 1$. Production probabilities can be combined in the matrix $B = (b_{jk})$.

A HMM λ can be described completely by the parameters A, B and Π. Now there are three interesting questions concerning HMMs:

1. How big is the *production probability* $P(O|\lambda)$ that a HMM $\lambda(A, B, \Pi)$ produces the given symbol sequence $O = o_1 o_2 \cdots o_T$?

2. Which is the *most probable state sequence* $S = s_1 s_2 \cdots s_T$ for given HMM $\lambda(A, B, \Pi)$ and symbol sequence $O = o_1 o_2 \cdots o_T$?

3. How to estimate the *parameters* of a HMM $\lambda(A, B, \Pi)$ out of a given symbol sequence $O = o_1 o_2 \cdots o_T$?

For each of this problems exist one or more useful solutions. The production probability can easily be calculated by the so called *Forward-Backward-Algorithm*. The most probable state sequence is obtained by the *Viterbi-Algorithm* and for parameter estimation there exist two efficient methods: the *Baum-Welsh-Training* and *Viterbi-Training*.

Descriptions of the algorithms above can be found in [5, 3, 4] and [7].

5.2 Classification and Time Segmentation

For the purpose of time segmentation, we can build up elemental HMMs for each interesting phase of our vocal fold vibrations. That means we are modeling glottis closing, prephonatoric closure, beginning vibration and steady state vibration as elemental or atomar models.

Based on these models we can define a general model by melting the elemental models together. This is done by introducing a transmission probability $a_{N_1 1_2}$ when melting elemental models λ_1 and λ_2.

Parameter estimation can be done also by training the elemental models with hand-segmentated feature vector sequences as by training the complete model with complete feature vector sequences. Training with segmentated data works quite better but you have to spend time on labeling the vector sequences.

The most probable state sequence supplied by the Viterbi-Algorithm indicates the time behavior of the interesting phases.

For classification of functional voice disorders we can define an elemental HMM for every interesting class, e.g. hyper and hypo functional disorders. These models are trained with complete sequences and classification decision is based on the production probability for a new sequence of each of the interesting models. We decide on the class with the highest production probability.

Another possibility is to combine both approaches by building up a larger alphabet of elemental models. For each of the models, which are significant for the estimated classes the afore mentioned models are replaced by class–specific new ones. In our case we for example replace the model of beginning vibration by a model of beginning vibration for hyper-functional disorders and a another model of beginning vibration for hypo-functional disorders. Models representing phases in which we dont expect different behavior like the model of steady–state vibration can be left in our alphabet as common models.

We then build up class specific general models out of elemental models. Training can also be done with segmentated and unsegmentated data. The advantage of such kind of models is, that classification and time segmentation can be combined: after deciding for the model with biggest production probability (classification) for a new feature vector sequence we use the Viterbi-Algorithm to calculate the most probable state sequence (time segmentation).

5.3 ISADORA

The speech recognition system ISADORA [7] provides all possibilities for building up elemental and combined HMMs, training and evaluation. So we can use this system to test out and build up all described features in this section.

6 Experiments and Results

First experiments for glottis tracking and time segmentation with active rays and HMMs have been conducted. We used 43 complete sequences showing simulated hyper- and hypo–functional disorders and tried to segment the boundaries of glottis closing and closure.

Therefore we tracked the glottis during the whole sequences by applying 90 active rays, calculated 3 hypotheses per ray and evaluated by hand. As a result we tracked 33 (76.7%) sequences good and 7 (16.3%) wrong. The remaining 3 (7.0%) sequences were also tracked correctly except for a short period of time at the beginning of the sequence. A result of glottis tracking can be seen in Figure 5.

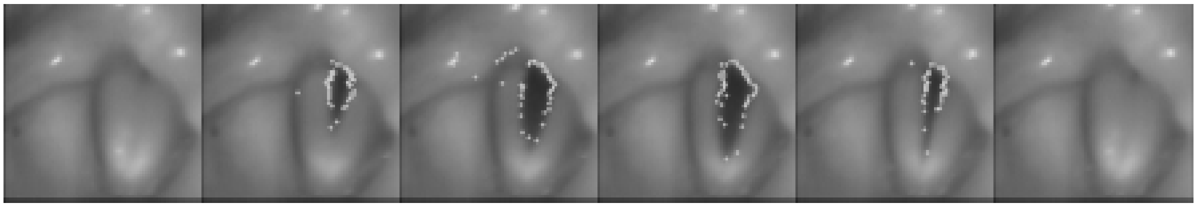

Figure 5: Results for tracking the glottis with active rays

To evaluate the time segmentation, we chose 14 out of the 90 rays and applied a normalization step to reduce influence of the glottis size and the angle of the main axis.

For each of the 36 above mentioned sequences we trained an own HMM network with the remaining data, calculated the interesting time boundaries and compared them against hand segmented values. We received an average deviation of 42.4 ms. In a few cases the correct boundaries have not been detected. Diagrams in Figure 6 show deviation of automatically calculated from manually segmentated boundaries in 20 ms intervals on the x-axis and the number of boundaries found in the corresponding interval on the y-axis.

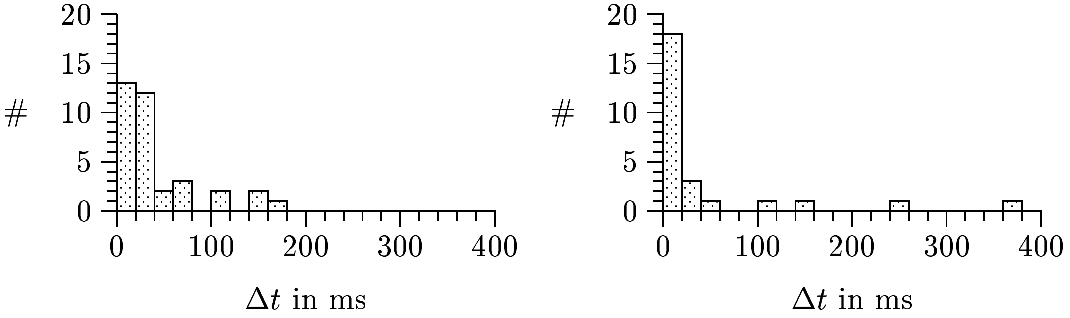

Figure 6: deviation of automatically segmentated from manually segmentated boundaries for the end of glottis closing (left) and end of glottis closure (right).

7 Conclusion and Future Work

In our contribution we have presented a new approach to glottis tracking by using active rays. Experiments have proven that this trial is well suited for an accurate glottis contour extraction.

In addition we introduced hidden markov models as a means to evaluate feature vector sequences provided by active rays. Major advantage of those models is that time segmentation and classification can easily be combined in one system. Existing systems can be adapted for our purposes without big problems.

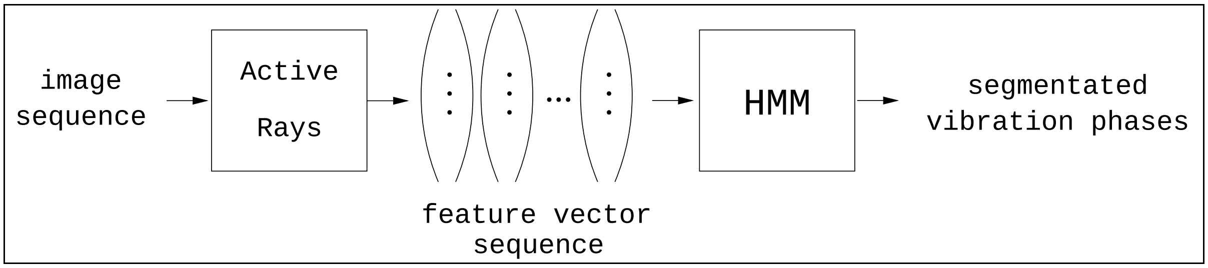

Figure 7: Overview about the system

An overview of the complete system can be seen in Figure 7. In our current work we are testing out the HMMs with results provided by glottis tracking using active rays. If our system shows good classification and time segmentation of hyper- and hypo-functional disorders we will extend classification to classes like paralyses of the vocal folds.

References

1. J. Denzler and H. Niemann. Active rays: A new approach to contour tracking. *International Journal of Computing and Information Technology*, 4(1):9–16, 1996.
2. U. Eysholdt, U. Pröschel, and M. Tigges. Direct evaluation of high speed recordings of vocal fold vibrations. In *The 3rd International Symposium on Phonosurgery, Kyoto, Japan*, 1994.
3. X. D. Huang, Y. Ariki, and M. A. Jack. *Hidden Markov Models for Speech Recognition.* Number 7 in Information Technology Series. Edinburgh University Press, Edinburgh, 1990.
4. T. Kuhn. *Die Erkennungsphase in einem Dialogsystem*, volume 80 of *Dissertationen zur künstlichen Intelligenz.* infix, St. Augustin, 1995.
5. L. Rabiner and B. Juang. An introduction to hidden markov models. *ASSP Magazine*, 1(3):4–16, 1986.
6. A.K. Saadah and N.P. Galatsanos. Deformation analysis of the vibrational patterns of the vocal folds. In *Bildverarbeitung für die Medizin, Algorithmen, Systeme, Anwendungen, Proceedings des Aachener Workshops am 8. und 9. November 1996*, volume 6 of *CEUR Workshop Proceedings.* Lehmann, T and Scholl, I and Spitzer, K, 1996.
7. E.G. Schukat-Talamazzini. *Automatische Spracherkennung.* Vieweg-Verlag, Braunschweig, 1995.
8. T. Wittenberg and U. Eyshold. Estimation of vocal fold vibrations using image segmentation. In G. Sagerer, G. Posch, and F. Kummert, editors, *Mustererkennung 1995 . Verstehen akustischer und visueller Informationen*, Informatik aktuell. Springer-Verlag, Berlin, 1995.