2018 European Control Conference (ECC)
June 12-15, 2018. Limassol, Cyprus

Stochastic Filter Techniques in Combination with Sliding Mode Differentiators as the Basis for a Reliable Neural Network-Based Recognition of Phonemes in Speech Signals

Andreas Rauh[1], Matthew Schmidt[1], Susann Tiede[2] and Cornelia Klenke[3]

Abstract— **The automatic estimation of the fundamental frequencies of phonemes is one of the important building blocks for the recognition of pronunciation disorders in spoken language. Based on the estimation of the so-called formant frequencies, it becomes possible to distinguish between voiced and unvoiced phonemes and to identify them reliably. Both the filter-based frequency estimation and the phoneme classification by neural networks are tasks that are investigated in the research project *SUSE* (A *S*oftware assistance system for *U*ncovering speech disorders by *S*tochastic *E*stimation techniques) bringing together the fields of signal processing as well as speech therapy and phonology. In this paper, a stochastic frequency estimation scheme based on the Unscented Kalman Filter is, firstly, extended by a sliding mode differentiator to enhance the accuracy of frequency estimation in naturally spoken language. Secondly, the estimation results are employed to train and implement a fundamental neural network classifier that can be used to distinguish automatically between different voiced and unvoiced phonemes. Classification results for an excerpt from a TV news broadcast conclude this paper.**

I. Introduction

The project *SUSE* aims at developing a software-based assistance system for speech therapists to detect disorders in the linguistic fields of pronunciation and grammar automatically. Besides disorders related to lexicon, these two fields are major areas with which speech therapists are concerned in their everyday work [1]. On the one hand, children, suffering from developmental speech acquisition disorders and, on the other hand, for example, elderly people or persons in rehabilitation after neurological diseases are typical groups of patients affected by pronunciation disorders. Especially if children are concerned, there is a large need to detect speech disorders at the earliest possible stage. The reason for this is that most information in primary education (not only in language-related subjects but, for example, also in natural sciences) is transferred by teachers in an oral or written manner despite numerous attempts to develop teaching techniques with an enhanced practical visualization.

The need for an assistance system that helps to detect language disorders at the earliest possible stage also becomes obvious by the fact that a recent investigation of first-grade pupils from various German primary schools [2] has shown a prevalence of speech disorders in the order of magnitude of approximately 50%. This large number is related to unreported cases of speech acquisition disorders in the range between 20 and 30% of all children attending primary school. Those children currently remain undetected by school entrance examinations (as well as tests during nursery school that are not unified wrt. standardized test procedures, quality, and contents) by medical doctors. This high number of unreported cases does not only increase the risk of failures with respect to knowledge transfer but also contradicts the rights of each pupil to be schooled according to her/his specific needs although this fact is widely acknowledged by governments world-wide and stated explicitly in the UN Convention on the Rights of Persons with Disabilities[1]. Therefore, the development of the assistance system *SUSE* aims at improving the currently unsatisfactory situation by extracting the most important aspects of a disorder and reporting it in condensed form to a qualified therapist.

In previous work, Extended Kalman Filter (EKF) [3] and Unscented Kalman Filter (UKF) [4] techniques were developed to estimate the formant frequencies included in speech signals in real time. These frequencies can be distinguished as follows: For voiced phonemes (like normal vowels), there exist a number of several sharp frequency components, while unvoiced sounds (like fricatives or whispered vowels) are characterized by several formants where each of them has a blurred frequency spectrum [5], [6]. State-of-the-art speech recognition techniques extract frequency information from the signal by a Discrete Fourier Transformation (DFT), partially in combination with a cepstrum transformation [7]. The advantage of EKF and UKF techniques is that they do not only provide expected values for the frequency estimates but also reveal information about the sharpness of the spectrum in terms of the corresponding standard deviations.

In this paper, continuous- and discrete-time state-space representations of speech signals are briefly reviewed in Sec. II as the basis for a subsequent observability analysis and filter design. In addition to [8], where banks of bandpass filters were employed to improve observability, Sec. II-C.3 presents a new approach making use of a sliding mode differentiator [9]. Optimal filter parameters (wrt. the numeric values of the process noise covariance) are determined for

[1]Chair of Mechatronics, University of Rostock, Justus-von-Liebig-Weg 6, D-18059 Rostock, Germany {Andreas.Rauh,Matthew.Schmidt}@uni-rostock.de
[2]Speech Therapist, Evangelisches Schulzentrum Demmin: Katharina von Bora, Waldstraße 20, D-17109 Demmin, Germany s.tiede.speechtherapy@gmail.com
[3]Speech Therapist, Lloydstraße 3, D-17192 Waren/ Müritz, Germany klenke.koerper-sprache@email.de

[1]http://www.un.org/disabilities/documents/convention/convoptprot-e.pdf, Article 24, 2(a)

specifically chosen test signals in Sec. III. The correspondingly parameterized UKF serves as the basis for frequency estimation in real speech signals for which phonemes are finally classified by a neural network trained with the help of the UKF estimates in Secs. IV and V. Conclusions and an outlook on future work are given in Sec. VI.

II. Modeling of Speech Signals

A. Continuous-Time Signal Model

As described in [8], [10], it is possible to approximate speech signals by a superposition of n harmonic oscillators with the amplitudes α_i and the phase shifts ϕ_i according to

$$y_{\mathrm{m}}(t) \approx y_{\mathrm{m},n}(t) = \sum_{i=1}^{n} \left(\alpha_i \cdot \cos\left(\omega_i \cdot t + \phi_i\right)\right) \quad . \tag{1}$$

In (1), $\omega_1 > 0$ is the basis frequency specified in $\mathrm{rad} \cdot \mathrm{s}^{-1}$ with the higher formant frequencies $\omega_2, \ldots, \omega_n$, $\omega_{i+1} > \omega_i$, $i \in \mathbb{N}$. Due to nonlinear vibrations of the vocal folds and the partially irregular, turbulent air flow expelled from the lungs during sound production, the frequencies in (1) are typically not integer multiples of ω_1. For each formant $i \in \{1, \ldots, n\}$, the fundamental signal model (1) can be transferred into a quasi-linear continuous-time state-space representation

$$\begin{aligned} &\dot{\mathbf{x}}_i(t) = \mathbb{A}_i\left(x_{3i}(t)\right) \cdot \mathbf{x}_i(t) \qquad \text{with} \\ &\mathbf{x}_i(t) = \begin{bmatrix} x_{3i-2}(t) & x_{3i-1}(t) & x_{3i}(t) \end{bmatrix}^T \;, \quad x_{3i} := \omega_i \;, \end{aligned} \tag{2}$$

the state-dependent system matrix

$$\mathbb{A}_i\left(x_{3i}(t)\right) = \left[\begin{array}{cc:c} 0 & 1 & 0 \\ -x_{3i}^2(t) & 0 & 0 \\ \hdashline 0 & 0 & 0 \end{array}\right] \;, \tag{3}$$

and the output equation

$$y_i(t) = \check{\mathbf{c}}_i^T \cdot \mathbf{x}_i(t) \quad \text{with} \quad \check{\mathbf{c}}_i^T = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \quad . \tag{4}$$

All submodels $i \in \{1, \ldots, n\}$ can then be combined into the joint set of state equations

$$\dot{\mathbf{x}}(t) = \mathbf{f}\left(\mathbf{x}(t)\right) = \mathbf{A}\left(\mathbf{x}(t)\right) \cdot \mathbf{x}(t) \quad \text{with} \quad \mathbf{x}(t) \in \mathbb{R}^{3n} \;, \tag{5}$$

the block diagonal system matrix

$$\mathbf{A}\left(\mathbf{x}\right) = \mathrm{blkdiag}\left\{\mathbb{A}_1\left(x_3\right), \mathbb{A}_2\left(x_6\right), \ldots, \mathbb{A}_n\left(x_{3n}\right)\right\} \;, \tag{6}$$

and the scalar overall output equation as a concatenation of all auxiliary vectors $\check{\mathbf{c}}_i^T$ defined in (4) according to

$$\begin{aligned} y_{\mathrm{m},n}(t) &= \sum_{i=1}^{n} x_{3i-2}(t) = \mathbf{c}^T \cdot \mathbf{x}(t) \quad \text{with} \\ \mathbf{c}^T &= \begin{bmatrix} \check{\mathbf{c}}_1^T & \check{\mathbf{c}}_2^T & \ldots & \check{\mathbf{c}}_n^T \end{bmatrix} \quad . \end{aligned} \tag{7}$$

B. Time Discretization of the System Model

For sufficiently small sampling times $T_{\mathrm{s}} = t_{k+1} - t_k = \frac{1}{f_{\mathrm{s}}} = \mathrm{const}$, the continuous-time model (5), (6) can be replaced by the discrete-time state-space representation [8]

$$\mathbf{x}_{k+1} = \exp\left(T_{\mathrm{s}} \cdot \mathbf{A}\left(\mathbf{x}_k\right)\right) \cdot \mathbf{x}_k =: \mathbf{A}_k^{\mathrm{d}} \cdot \mathbf{x}_k \;, \tag{8}$$

with the state vector $\mathbf{x}_k$ approximating the continuous-time states $\mathbf{x}(t_k)$ in (5) for $t = t_k$, the discretized output equation

$$y_k = y_{\mathrm{m},n,k} := y_{\mathrm{m},n}(t_k) = \mathbf{c}^T \cdot \mathbf{x}_k \;, \tag{9}$$

and the sequence $y_{\mathrm{m},k} := y_{\mathrm{m}}(t_k)$ of measured data.

C. Observability Analysis of the Quasi-Linear System Model

As a justification for the following extension of the output equations (7) and (9) by at least a second vector component, the original continuous-time model is analyzed wrt. observability [8]. For sufficiently small sampling times $T_{\mathrm{s}} = \frac{1}{f_{\mathrm{s}}}$, e.g. $f_{\mathrm{s}} = 44.1\,\mathrm{kHz}$ as in the presented application scenario, the discrete-time approximation is fully observable if this property can be verified for the continuous-time model.

1) Model with a Scalar Measurement: Interpreting the state dependency of $\mathbf{A}(\mathbf{x})$ as piecewise constant scheduling parameters, the observability matrix

$$\mathbf{Q}_{\mathrm{O}}^{\mathrm{lin}} = \begin{bmatrix} \mathbf{c}^T \\ \mathbf{c}^T\mathbf{A}(\mathbf{x}) \\ \mathbf{c}^T(\mathbf{A}(\mathbf{x}))^2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{1,1}^{\mathrm{lin}} & \mathbf{Q}_{1,2}^{\mathrm{lin}} & \ldots & \mathbf{Q}_{1,n}^{\mathrm{lin}} \\ \vdots & & & \vdots \\ \mathbf{Q}_{\tilde{n},1}^{\mathrm{lin}} & \mathbf{Q}_{\tilde{n},2}^{\mathrm{lin}} & \ldots & \mathbf{Q}_{\tilde{n},n}^{\mathrm{lin}} \end{bmatrix} \tag{10}$$

is obtained for all $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, \tilde{n}\}$ with

$$\mathbf{Q}_{j,i}^{\mathrm{lin}} = (-1)^{j-1} \begin{bmatrix} \left(x_{3i}^2\right)^{j-1} & 0 & 0 \\ 0 & \left(x_{3i}^2\right)^{j-1} & 0 \end{bmatrix} \quad .$$

In (10), the upward rounded parameter $\tilde{n} = \left\lceil \frac{3n}{2} \right\rceil$ denotes the maximum number of matrix blocks in each column of $\mathbf{Q}_{\mathrm{O}}^{\mathrm{lin}}$. Obviously, each third column of $\mathbf{Q}_{\mathrm{O}}^{\mathrm{lin}}$ is completely zero. Hence, if the system model is considered in a linearized form with the scalar output $y_{\mathrm{m},n}$, it is not fully state observable.

However, the situation changes if the nonlinear model is analyzed with respect to observability. In this case, it can be shown that the first $3n$ entries in the observability mapping

$$\mathbf{q}(\mathbf{x}) = \begin{bmatrix} \mathbf{q}_1(\mathbf{x}) \\ \vdots \\ \mathbf{q}_{\tilde{n}}(\mathbf{x}) \end{bmatrix} \quad \text{with } \mathbf{q}_j(\mathbf{x}) = \sum_{i=1}^{n} (-x_{3i})^{j-1} \cdot \begin{bmatrix} x_{3i-2} \\ x_{3i-1} \end{bmatrix} \;, \tag{11}$$

corresponding to the measured output and its derivatives up to the system order, can be solved locally for all state variables $\mathbf{x}$. To prove this fact, determine the Jacobian

$$\begin{aligned} \mathbf{Q}_{\mathrm{O}}^{\mathrm{nl}} &= \frac{\partial \mathbf{q}(\mathbf{x})}{\partial \mathbf{x}} = \begin{bmatrix} \mathbf{Q}_{1,1}^{\mathrm{nl}} & \mathbf{Q}_{1,2}^{\mathrm{nl}} & \ldots & \mathbf{Q}_{1,n}^{\mathrm{nl}} \\ \vdots & & & \vdots \\ \mathbf{Q}_{\tilde{n},1}^{\mathrm{nl}} & \mathbf{Q}_{\tilde{n},2}^{\mathrm{nl}} & \ldots & \mathbf{Q}_{\tilde{n},n}^{\mathrm{nl}} \end{bmatrix} \quad \text{with} \\ \mathbf{Q}_{j,i}^{\mathrm{nl}} &= (-1)^{j-1} \begin{bmatrix} x_{3i}^{2j-2} & 0 & (2j-2)x_{3i-2}x_{3i}^{2j-3} \\ 0 & x_{3i}^{2j-2} & (2j-2)x_{3i-1}x_{3i}^{2j-3} \end{bmatrix} \end{aligned} \tag{12}$$

as a representation of the local observability matrix of the nonlinear system; $\mathbf{Q}_{\mathrm{O}}^{\mathrm{nl}}$ has full column rank if the formant frequencies $\omega_i = x_{3i}$ are mutually unequal, which proves local observability of the nonlinear system model.

2) Model with a Vector-Valued Measurement (Extension by Linear Bandpass Filter Banks): Considering the observability analysis above, it is possible to enhance the accuracy of EKF and UKF approaches for the frequency estimation in speech signals if a vector-valued output is considered instead of a scalar one. Besides the extension of the output equation by a bank of parallel bandpass filters ($j \in \{1, \ldots, M\}$), it is also possible to include derivatives of the speech signal directly in an extended estimation scheme.

If bandpass filters are considered (for more details, see [8]), their cut-off frequencies have to be specified within the relevant frequency range $\omega \in [0\ ;\ \omega_n]$, $\omega_n < \pi f_\mathrm{s}$ which is bounded from above according to Shannon's sampling theorem. A description of these filters by using linear discrete-time, asymptotically stable state-space representations

$$\boldsymbol{\xi}_{j,k+1} = \boldsymbol{\Xi}_j \cdot \boldsymbol{\xi}_{j,k} + \boldsymbol{\zeta}_j \cdot \psi_k \ , \quad \boldsymbol{\gamma}_{j,k} = \boldsymbol{\Gamma}_j \cdot \boldsymbol{\xi}_{j,k} \tag{13}$$

with the measured signal as the filter input $\psi_k = y_{\mathrm{m},k}$ leads to the possibility to define an extended system model

$$\begin{bmatrix} \mathbf{x}_{k+1} \\ \boldsymbol{\xi}_{1,k+1} \\ \vdots \\ \boldsymbol{\xi}_{M,k+1} \end{bmatrix} = \begin{bmatrix} \mathbf{A}_k^\mathrm{d} & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{\zeta}_1 \mathbf{c}^T & \boldsymbol{\Xi}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \boldsymbol{\zeta}_2 \mathbf{c}^T & \mathbf{0} & \boldsymbol{\Xi}_2 & \dots & \mathbf{0} \\ \vdots & & & \ddots & \vdots \\ \boldsymbol{\zeta}_M \mathbf{c}^T & \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Xi}_M \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\xi}_{1,k} \\ \vdots \\ \boldsymbol{\xi}_{M,k} \end{bmatrix} \ . \tag{14}$$

This extended model can be summarized in compact form by the equation $\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{A}}_k^\mathrm{d} \cdot \tilde{\mathbf{x}}_k$, where the filter outputs $\boldsymbol{\gamma}_k = \begin{bmatrix} \boldsymbol{\gamma}_{1,k}^T & \dots & \boldsymbol{\gamma}_{M,k}^T \end{bmatrix}^T$ are appended to the actually measured signal $y_{\mathrm{m},k}$ during both filter design and implementation.

This linear bandpass filter approach has two drawbacks:

(a) limitations of the gain factors to prevent the amplification of measurement noise and

(b) the inevitable introduction of phase shifts (time delays) by linear filters which always have (for sufficiently high frequencies) global low-pass dynamics.

3) Model with a Vector-Valued Measurement (Extension by Sliding Mode Differentiators): Both drawbacks (a) and (b) can be eliminated if sliding mode differentiation is used to extend the scalar system output $y_{\mathrm{m},k} = y_\mathrm{m}(t_k)$ by its first time derivative $\dot{y}_{\mathrm{m},k} = \dot{y}_\mathrm{m}(t_k)$. For that purpose, first- or higher-order differentiators can be applied in general, cf. [9]. Due to the better noise suppression properties of a second-order differentiation scheme, it is applied in the following.

As a computationally inexpensive solution for the derivative estimation, the second-order sliding mode differentiator [9] given by the continuous-time state equations

$$\begin{aligned} \dot{\chi}_0 &= \upsilon_0 \ , \quad \dot{\chi}_1 = \upsilon_1 \\ \dot{\chi}_2 &= -1.1 \cdot L \cdot \mathrm{sign}\,(\chi_2 - \upsilon_1) \end{aligned} \tag{15}$$

with the two variable-structure expressions

$$\begin{aligned} \upsilon_0 &= -2 \cdot L^{\frac{1}{3}} \cdot |\chi_0 - y_\mathrm{m}|^{\frac{2}{3}} \cdot \mathrm{sign}\,(\chi_0 - y_\mathrm{m}) + \chi_1 \\ \upsilon_1 &= -1.5 \cdot L^{\frac{1}{2}} \cdot |\chi_1 - \upsilon_0|^{\frac{2}{3}} \cdot \mathrm{sign}\,(\chi_1 - \upsilon_0) + \chi_2 \end{aligned} \tag{16}$$

is used in this paper. If $L > 0$ is chosen according to [9] as a Lipschitz constant overapproximating the absolute value of the largest time derivative of the signal to be differentiated, the derivative of the speech signal is approximated by

$$\upsilon_0(t_k) \approx \dot{y}_\mathrm{m}(t_k) \ , \tag{17}$$

which serves as a further output

$$\dot{y}_i = \check{\mathbf{c}}_i^T \cdot \mathbf{x}_i \quad \text{with} \quad \check{\mathbf{c}}_i^T = \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} \tag{18}$$

to be appended to the output Eqs. (7) and (9) according to

$$\begin{bmatrix} y_{\mathrm{m},n} \\ \dot{y}_{\mathrm{m},n} \end{bmatrix} = \mathbf{C} \cdot \mathbf{x} \ \text{ with } \ \mathbf{C} = \begin{bmatrix} \mathbf{c}_1^T & \mathbf{c}_2^T & \dots & \mathbf{c}_n^T \\ \check{\mathbf{c}}_1^T & \check{\mathbf{c}}_2^T & \dots & \check{\mathbf{c}}_n^T \end{bmatrix} \ . \tag{19}$$

D. Consideration of Process and Measurement Noise

1) The Scalar Output Case: To implement EKF and UKF approaches for the quasi-linear discrete-time model (8), (9), the additive stochastic process noise $\mathbf{w}_k \in \mathbb{R}^{3n}$ (summarizing the effect of all modeling errors concerning (5) and (8)) and the uncorrelated additive measurement noise $v_k \in \mathbb{R}$ are introduced in the state-space representation according to

$$\mathbf{x}_{k+1} = \mathbf{A}_k^\mathrm{d} \cdot \mathbf{x}_k + \mathbf{w}_k \quad \text{and} \tag{20}$$

$$y_k = \mathbf{c}^T \cdot \mathbf{x}_k + v_k \ . \tag{21}$$

In (20), both noise terms $\mathbf{w}_k$ and v_k are normally distributed with vanishing mean values $\boldsymbol{\mu}_{w,k} = \mathbf{0}$ and $\mu_{v,k} = 0$ as well as with positive definite (co-)variances $\mathbb{C}_{w,k}$ and $\mathbb{C}_{v,k}$.

2) The Vector-Valued Case (Extension by Linear Bandpass Filter Banks): If a bank of linear bandpass filters is introduced according to Ssec. II-C.2 to obtain a vector-valued system output, the state equations including an additive term $\tilde{\mathbf{w}}_k$ for the process noise turn into $\tilde{\mathbf{x}}_{k+1} = \tilde{\mathbf{A}}_k^\mathrm{d} \cdot \tilde{\mathbf{x}}_k + \tilde{\mathbf{w}}_k$ with the corresponding covariance matrix

$$\mathbb{C}_{\tilde{w},k} = \begin{bmatrix} \mathbb{C}_{w,k} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\zeta}_1 \mathbb{C}_{v,k} \boldsymbol{\zeta}_1^T & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\zeta}_M \mathbb{C}_{v,k} \boldsymbol{\zeta}_M^T \end{bmatrix} \ . \tag{22}$$

Analogously, the extended output model

$$\hat{\boldsymbol{\gamma}}_k = \begin{bmatrix} \mathbf{c}^T & \mathbf{0}^T & \dots & \mathbf{0}^T \\ \mathbf{0} & \boldsymbol{\Gamma}_1 & \dots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Gamma}_M \end{bmatrix} \cdot \begin{bmatrix} \mathbf{x}_k \\ \boldsymbol{\xi}_{1,k} \\ \vdots \\ \boldsymbol{\xi}_{M,k} \end{bmatrix} + \tilde{\mathbf{v}}_k = \tilde{\mathbf{C}} \cdot \tilde{\mathbf{x}}_k + \tilde{\mathbf{v}}_k \tag{23}$$

involves the measurement noise covariance

$$\mathbb{C}_{\tilde{v},k} = \begin{bmatrix} \mathbb{C}_{v,k} & \mathbf{0}^T \\ \mathbf{0} & \epsilon \cdot \mathbf{I} \end{bmatrix} \ , \tag{24}$$

where $\mathbf{I}$ is an identity matrix of appropriate dimension and ϵ a small non-negative constant preserving the full rank of the matrix to the inverted in the innovation steps of both the EKF and UKF [8].

3) The Vector-Valued Case (Extension by Sliding Mode Differentiators): For the extension of the system output by a sliding mode differentiator, the state equations (20) and the included process noise remain unchanged. However, the output equation (21) has to be replaced by an extension of (19) with stochastic measurement noise according to

$$\begin{bmatrix} y_{\mathrm{m},n} \\ \dot{y}_{\mathrm{m},n} \end{bmatrix} = \mathbf{C} \cdot \mathbf{x} + \mathbf{v}'_k \ \text{ with } \ \mathbb{C}_{v',k} = \begin{bmatrix} \mathbb{C}_{v,k} & 0^T \\ 0 & \bar{\omega}^2 \cdot \mathbb{C}_{v,k} \end{bmatrix} \ , \tag{25}$$

where $\bar{\omega}$ is the characteristic ratio between the amplitudes of the differentiated and the original speech signal.

4) Filter Algorithms: EKF and UKF Implementation: Using the discrete-time system models and the normally distributed noise processes described above, both EKF and UKF algorithms can be implemented. For a detailed description of both filters as well as for a comparison between their estimation accuracy, the reader is referred to [8]. In that

paper, it was shown that the UKF estimate is significantly better than the EKF since it captures the nonlinearity of the state equations (square values of formant frequencies are included in the state-dependent system matrix) in a significantly better form and, hence, fulfills the observability requirement of handling the nonlinear system model directly. This latter feature is due to the fact that the UKF is based on multiple evaluations of the nonlinear state equations at several so-called sigma points which serve as a numerical approximation of the expected values and covariances in the prediction step that is executed between two subsequent points of time at which measured data become available.

III. Optimization of the Filter Parameters

A. Generation of Test Signals

The parameterization of the scalar measurement variance $\mathbb{C}_{v,k} > 0$ is straightforward by an analysis of the recorded speech signal. This parameter captures the influence of both random disturbances in the recorded signal and quantization noise in terms of a Gaussian approximation. In contrast, determining suitable values for $\mathbb{C}_{w,k} \in \mathbb{R}^{3n\times 3n}$ can only be performed heuristically if no further information is available. Heuristic approaches, used for example in [8], lead to a choice of this matrix in diagonal form where the absolute entries are chosen such that they cover typical variation rates of all state variables between two subsequent sampling points t_k and t_{k+1} in terms of an approximation of the corresponding variances. However, this heuristic selection procedure involves some trial-and-error parameter finding. It can be replaced by the optimization procedure described in the following subsection, if test signals are defined which represent transitions between piecewise defined signals consisting of a superposition of various harmonic components with different amplitudes. To make these test signals as close as possible to a real speech signal, frequencies and amplitudes of the first two/three formants of voiced phonemes are included in these test signals as summarized in Tabs. I and II. Symbols in IPA transcription in the first rows of these tables represent phonemes that are similar to the considered test values. Between each of the given constant values, phases are introduced in which a linear interpolation between the subsequent frequency and amplitude values is performed over a time span that corresponds to the typical transition times between two subsequent sounds in spoken language. In addition, the duration of piecewise constant signal phases is adjusted to the typical length of some relevant phonemes. Moreover, the range specification a–b in Tab. II means that the corresponding parameter varies linearly between the bounds a and b during the specified duration.

B. Optimality Criterion and Numerical Optimization Results

The systematic parameterization of the diagonal elements of the process noise covariance matrices is performed in such a way that the performance criterion

$$J = \sum_{i=1}^{n} J_i \quad \text{with} \tag{26}$$

TABLE I

Test signal for $n = 2$.

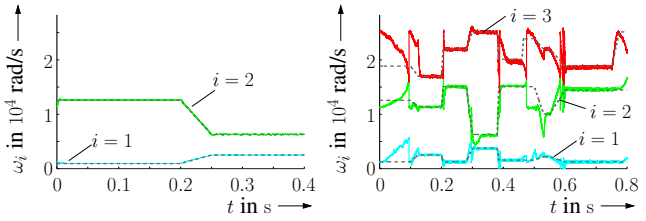
	[eː]	[ɔ]
frequency ω_1 in rad/s	942	2513
frequency ω_2 in rad/s	12566	6283
amplitude α_1	0.18	0.09
amplitude α_2	0.085	0.035
duration in s	0.2	0.15

$$J_i = \sqrt{\frac{1}{N}\cdot\sum_{k=1}^{N}\left(\omega_{i,k}-\hat{\omega}_{i,k}\right)^2} + \sqrt{\frac{\bar{C}}{N}\cdot\sum_{k=1}^{N}\left(\alpha_{i,k}-\hat{\alpha}_{i,k}\right)^2} + \Xi_i \tag{27}$$

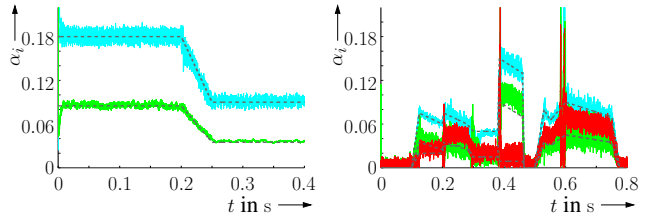
for $i \in \{1,\ldots,n\}$ is minimized by means of the Nelder-Mead simplex method. The numerical optimization has been initialized with values according to [8] representing typical variation rates of the state variables to be estimated by the stochastic filter approach. In (27), $\bar{C}$ is a scaling factor ensuring similar weighting of deviations in the estimated frequencies and amplitudes; Ξ_i represents a non-zero penalty term that is added to the average deviations of the estimated frequencies $\hat{\omega}_{i,k} = \mu^{\mathrm{e}}_{x,3i,k}$ and amplitudes from their corresponding true values if unphysical (e.g. negative) frequencies are estimated by the UKF with a given order n. With the help of the estimated state vector $\boldsymbol{\mu}^{\mathrm{e}}_{x,k}$ in the UKF's innovation step, the signal amplitudes can be determined as

$$\hat{\alpha}_i = \sqrt{\left(\mu^{\mathrm{e}}_{x,3i-2,k}\right)^2 + \left(\frac{\mu^{\mathrm{e}}_{x,3i-1,k}}{\mu^{\mathrm{e}}_{x,3i,k}}\right)^2}\,. \tag{28}$$

If the optimized process noise covariances are applied to the test signals defined by Tabs. I and II, accurate estimates can be obtained as shown in Fig. 1. These filter parameters represent the settings that are also used in the following section for the frequency estimation in a real speech signal.



(a) Frequency estimates for $n = 2$. (b) Frequency estimates for $n = 3$.



(c) Amplitude estimates for $n = 2$. (d) Amplitude estimates for $n = 3$.

Fig. 1. Estimation of signal frequencies and amplitudes according to the test signals in Tabs. I and II. Phases of significant deviations between the true frequency values (dashed lines) and the estimates (solid lines) in Fig. 1(b) correspond to phases of silence with vanishing signal amplitudes.

TABLE II
TEST SIGNAL FOR $n = 3$.

	[ ]	[eː]	[iː]	[ɔ]	[iː]	[ ]	[eː]	[ɜ]	[ ]
freq. ω_1 in rad/s	1257	2513	1257	3770	1571	1571	2199	1257	1257
freq. ω_2 in rad/s	12566	11310	15080	6283	15080	15080	10053	14451	14451
freq. ω_3 in rad/s	18850	16965	21991	25133	20106	23876	21991	18850	25133
amplitude α_1	0	0.075–0.06	0.075–0.06	0.05	0.15–0.13	0	0.08–0.065	0.09–0.07	0
amplitude α_2	0	0.035–0.025	0.035–0.025	0.015	0.09–0.07	0	0.035–0.03	0.045–0.035	0
amplitude α_3	0	0.02	0.03	0.015	0.009	0	0.035–0.03	0.045–0.035	0
duration in s	0.1	0.075	0.075	0.08	0.075	0.035	0.02	0.15	0.025

IV. NEURAL NETWORK PATTERN RECOGNITION FOR THE CLASSIFICATION OF VOICED AND UNVOICED PHONEMES

In this section, a neural network-based pattern recognition approach is described for the classification of voiced and unvoiced phonemes, where the input data for the neural network are determined by the UKF without and with an extension by the sliding mode differentiator.

For the generation of training, validation, and testing data, it is assumed that temporal phoneme boundaries are available. In this paper, they are determined manually by listening to the speech signal, which corresponds to a $91\,\mathrm{s}$ excerpt from a German TV news broadcast (pure speech of the anchorman without background music). In future work, the manual segmentation will be replaced by automatically determined phoneme boundaries for which the segmentation procedure published in [11] will be further developed.

The **input data set 1 (IDS1)** consists of the estimated state vectors $\boldsymbol{\mu}^{\mathrm{e}}_{x,k}$ which are averaged over the phoneme duration, the averaged estimated variances $\mathbb{C}^{\mathrm{e}}_{x,i,i,k}$ for each $i \in \{1,\ldots,3n\}$, and the phoneme duration. In addition, the **input data set 2 (IDS2)** further contains the minimum and maximum values of the variances for each phoneme and an upper state boundary according to $\max_k\{\mu^{\mathrm{e}}_{x,i,k}\} + 3\sqrt{\max_k\{\mathbb{C}^{\mathrm{e}}_{x,i,i,k}\}}$, where the maximization takes place over all time instants k within each segmented phoneme.

For classification purposes, a two-layer feedforward network containing a single hidden sigmoid layer with $\mathcal{M}$ neurons and a softmax output layer is employed. Training is performed by backpropagation based on a scaled conjugate gradient method. Note that the classification accuracy can be improved by increasing the number of neurons in the hidden layer. However, a reasonable choice of the number of neurons should take into account the error rates with respect to both training and validation data (the latter expressing generalization capabilities to unknown inputs during training) and to independent test data. Training and validation of the neural network is restricted to those 36 phonemes $\mathbb{I} = \{$[aː], [eː], [iː], [oː], [a], [ɐ], [ə], [ɛ], [ɪ], [ɔ], [ʊ], [ʏ], [a͡ɪ], [a͡ʊ], [ɛ͡ɐ], [b], [d], [g], [p], [t], [k], [m], [n], [ŋ], [ɳ], [f], [v], [h], [z], [s], [ʃ], [ç], [x], [l], [r], [ts]$\}$ of the German language that appear at least 10 times in the above-mentioned excerpt from the news broadcast. The complete set of training, validation, and test data consists of 1069 phonemes.

V. ESTIMATION RESULTS

A. Filter-Based Frequency Estimation

For $n = 3$, Fig. 2 gives a graphical representation of the frequency values estimated by the UKF with an extension of the system output by the sliding mode differentiator for a $10\,\mathrm{s}$ excerpt of the test signal. A more detailed analysis of UKF outputs with different orders n, however, without using the sliding mode differentiator was published in [8] together with a comparison to a DFT-based frequency estimation.

B. Classification of Phonemes

The UKF estimates for $n = 2$ as well as $n = 3$ according to the previous subsection are now employed for the neural network-based classification of individual phonemes. Tab. III gives a summary of the training and validation results of the neural network according to Sec. IV with different numbers $\mathcal{M}$ of neurons in the hidden layer. To make the results insensitive against the actual choice of training and validation data from the 1069 available phonemes, each network configuration was trained 20 times with randomly chosen inputs (in each case 70% of the data set) as well as validated and tested against two independent sets of each 15% randomly chosen values. The results in Tab. III represent the averaged percentages of correctly classified phonemes from the complete set $\mathbb{I}$ as well as the maximum values of correctly classified vowels with a distinction between short and long variants $\mathbb{V}_1 = \{$[aː], [eː], [iː], [oː], [a], [ɛ], [ɪ], [ɔ], [ʊ]$\}$ as well as without their distinction according to $\mathbb{V}_2 = \{$([aː],[a]), ([eː],[ɛ]), ([iː],[ɪ]), ([oː],[ɔ]), ([ʊ])$\}$. It can be noticed that both increasing the number of neurons in the hidden layer as well as increasing the filter order leads to improved estimation accuracies. Moreover, especially for $n = 3$, the use of the sliding mode differentiator ($n_y = 2$) leads to better results than a pure usage of the scalar system output ($n_y = 1$). This can be seen by comparing the corresponding percentages for identical values of n and $\mathcal{M}$, where the best results are marked in boldface in Tab. III. Finally, Fig. 3 shows the confusion matrix (given in percentages) for the classification of all phonemes from the set $\mathbb{I}$, where $\tilde{\mathbb{I}}$ denotes those entries of the set $\mathbb{I}$ to which the actual phonemes are mapped. It can be seen that the accuracy of classification is very similar for each considered sound representing the large percentage of correct classifications (located on the diagonal of Fig. 3) according to the last row of Tab. III.

TABLE III

COMPARISON OF DIFFERENT NEURAL NETWORK PHONEME CLASSIFIERS BY THE PERCENTAGE OF CORRECT CLASSIFICATIONS.

			training data (all phonemes $\mathbb{I}$)		validation data (all phonemes $\mathbb{I}$)		test data (all phonemes $\mathbb{I}$)		vowels $\mathbb{V}_1$: w/ distinction betw. short/long		vowels $\mathbb{V}_2$: w/o distinction betw. short/long	
			IDS1	IDS2	IDS1	IDS2	IDS1	IDS2	IDS1	IDS2	IDS1	IDS2
$n=2$	$n_y=1$	$\mathcal{M}=100$	**59.78%**	**67.16%**	42.43%	**49.10%**	**42.20%**	**47.98%**	41.27%	**46.19%**	**71.21%**	**80.96%**
		$\mathcal{M}=250$	**63.19%**	**69.73%**	**44.21%**	**50.22%**	43.08%	**51.27%**	44.29%	**49.12%**	**76.80%**	**79.71%**
		$\mathcal{M}=1000$	**69.11%**	**79.52%**	**45.90%**	**50.06%**	**45.17%**	**53.46%**	**52.96%**	**56.93%**	**87.34%**	**91.67%**
	$n_y=2$	$\mathcal{M}=100$	56.90%	65.30%	**42.61%**	45.59%	39.99%	45.64%	**46.05%**	40.85%	71.04%	71.64%
		$\mathcal{M}=250$	60.91%	67.65%	41.42%	47.35%	**44.72%**	46.72%	**46.82%**	42.88%	76.01%	75.36%
		$\mathcal{M}=1000$	68.78%	73.07%	43.26%	47.12%	42.37%	46.84%	47.62%	41.35%	79.64%	74.61%
$n=3$	$n_y=1$	$\mathcal{M}=100$	**66.63%**	**81.00%**	49.26%	58.30%	49.48%	58.18%	**48.20%**	46.75%	**73.57%**	75.81%
		$\mathcal{M}=250$	**70.21%**	82.32%	50.61%	59.37%	50.41%	58.33%	**49.22%**	**57.03%**	**86.29%**	**92.14%**
		$\mathcal{M}=1000$	**78.52%**	**86.17%**	53.29%	57.78%	**52.53%**	56.90%	**56.77%**	56.49%	**90.11%**	**93.40%**
	$n_y=2$	$\mathcal{M}=100$	65.32%	79.93%	**50.29%**	**60.47%**	**51.32%**	**60.33%**	43.16%	**50.77%**	72.77%	**81.87%**
		$\mathcal{M}=250$	68.41%	**83.29%**	**53.44%**	**59.85%**	**50.59%**	**60.31%**	47.86%	51.22%	82.05%	87.02%
		$\mathcal{M}=1000$	74.52%	86.08%	**54.53%**	**59.28%**	51.74%	**58.26%**	49.56%	**59.84%**	79.74%	89.05%

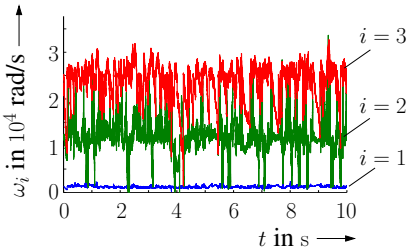


Fig. 2. Frequency estimates by the UKF with $n = 3$ extended by the sliding mode derivative estimation.

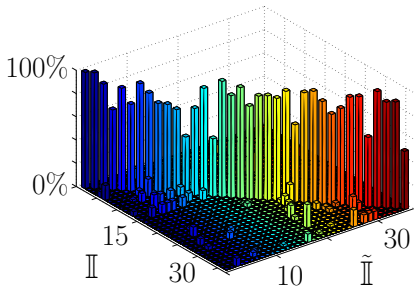


Fig. 3. Confusion matrix for the classification of the phonemes $\mathbb{I}$.

VI. CONCLUSIONS AND OUTLOOK ON FUTURE WORK

In this paper, a novel combination of filter-based frequency estimators for speech signals with a neural network classification scheme was presented. To make this procedure applicable to the detection of pronunciation disorders in the frame of speech therapy, the classification is performed on the level of individual phonemes. This is the major difference to state-of-the-art speech recognition systems in which estimated frequencies are usually compared against a dictionary of correctly pronounced syllables or words, leading to an inevitable loss of information about the disorders of interest.

Future work will deal with a further enhancement of the estimation scheme by systematically determining optimal values for the numbers n of frequencies considered in the UKF and for the numbers of neurons $\mathcal{M}$. Moreover, time series-based neural networks will be compared with the currently implemented variant which aggregates information about each phoneme in terms of point values representing either average or maximum values. The advantage of considering time series directly will be the detection of frequency variations within individual sounds. This is especially helpful to classify unvoiced bilabial stops such as [b] or [p] which only become audible in combinations with vowels.

REFERENCES

[1] J. S. Damico, N. Müller, and M. J. Ball, *The Handbook of Language and Speech Disorders*, ser. Blackwell Handbooks in Linguistics. Chichester, West Sussex, UK: Wiley, 2010.

[2] S. Tiede and J.-U. Braun, "Ist Chancengerechtigkeit für Kinder mit Sprachentwicklungsstörungen schon Realität? — Eine empirische Querschnittstudie zur Quantifizierung des Bedarfs sprachtherapeutischer Interventionen im Primarbereich (Has the Equity of Opportunities Already Become Reality for Children with Speech Acquisition Disorders? — An Empirical Cross Sectional Study for the Quantification of the Needs for Speech Therapeutic Interventions in Primary Education)," *Forschung Sprache*, vol. 5, no. 1, pp. 21–39, 2017.

[3] R. Stengel, *Optimal Control and Estimation*. Dover Publications, Inc., 1994.

[4] S. J. Julier, J. K. Uhlmann, and H. F. Durrant-Whyte, "A New Approach for the Nonlinear Transformation of Means and Covariances in Filters and Estimators," *IEEE Transactions on Automatic Control*, vol. 45, no. 3, pp. 477–482, 2000.

[5] E. C. Zsiga, *The Sounds of Language: An Introduction to Phonetics and Phonology*, ser. Linguistics in the World. Chichester, West Sussex, UK: Wiley, 2012.

[6] P. Ladefoged and I. Maddieson, *The Sounds of the World's Languages*, ser. Phonological Theory. Chichester, West Sussex, UK: Wiley, 1996.

[7] D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*, ser. Signals and Communication Technology. London: Springer-Verlag, 2015.

[8] A. Rauh, S. Tiede, and C. Klenke, "Comparison of Different Filter Approaches for the Online Frequency Analysis of Speech Signals," in *Proc. of 22nd IEEE Intl. Conference on Methods and Models in Automation and Robotics MMAR 2017*, Miedzyzdroje, Poland, 2017.

[9] A. Levant, "Higher-Order Sliding Modes, Differentiation and Output-Feedback Control," *International Journal of Control*, vol. 76, no. 9–10, pp. 924–941, 2003.

[10] A. Rauh, S. Tiede, and C. Klenke, "Observer and Filter Approaches for the Frequency Analysis of Speech Signals," in *Proc. of 21st IEEE Intl. Conference on Methods and Models in Automation and Robotics MMAR 2016*, Miedzyzdroje, Poland, 2016.

[11] ——, "Stochastic Filter Approaches for a Phoneme-Based Segmentation of Speech Signals," in *Proc. of 21st IEEE Intl. Conference on Methods and Models in Automation and Robotics MMAR 2016*, Miedzyzdroje, Poland, 2016.