An Interval-Based Algorithm for Feature Extraction from Speech Signals*

Andreas Rauh
University of Rostock, Chair of Mechatronics,
Justus-von-Liebig-Weg 6, D-18059 Rostock, Germany
Andreas.Rauh@uni-rostock.de

Susann Tiede
Speech Therapist
Evangelisches Schulzentrum Demmin: Katharina von Bora, Waldstraße 20, D-17109 Demmin, Germany
s.tiede.speechtherapy@gmail.com

Cornelia Klenke
Speech Therapist
Lloydstraße 3, D-17192 Waren/ Müritz, Germany
klenke.koerper-sprache@email.de

Abstract

Language disorders can be classified into the three linguistic levels of pronunciation, lexicon, and grammar. To identify the most important linguistic processes leading to disorders in the before-mentioned fields, both standardized test procedures and the analysis of freely spoken language are used in the everyday work of speech therapists. However, especially the analysis of freely spoken language may become a tedious and time consuming task for a speech therapist because it involves a repeated listening to the recorded speech of each individual patient. Therefore, a current research project, focusing on the German language, aims at the development of a computer-based assistance system for the automatic detection of pronunciation and grammar disorders to enable a speech therapist to spend her/ his valuable time more efficiently on therapeutic interventions and therapy planning. In this paper, interval-based algorithms are presented which form a basic building block for the automatic segmentation of speech signals into individual phonemes. Moreover, a first classification scheme for individual sounds is presented that will be employed in future

*Submitted: December 4, 2016; Revised: June 12, 2017; Accepted: August 22, 2017.

work for the automatic pronunciation analysis and for the automatic detection and classification of pronunciation disorders.

Keywords: Signal processing, Speech analysis, Phoneme and pattern recognition, Extended Kalman Filters, Branch-and-bound algorithms, Interval likelihood functions
AMS subject classifications: 65G30,93E11,60G35,68T10

1 Introduction

To develop a computer-based assistance system for speech therapy, it is essential to distinguish between the linguistic levels of lexicon, grammar, and pronunciation [2] and to deal with (i) the automatic transcription and preprocessing of speech involving erroneous pronunciation, (ii) the classification of pronunciation disorders, and (iii) a grammatical analysis[1].

For the tasks (i) and (ii), this contribution exploits an online frequency analysis of speech signals based on stochastic filtering approaches and the identification of points of time at which transitions between subsequent phonemes can be expected (cf. [7, 8]). The online frequency analysis makes use of a dynamic system model in discrete-time state-space representation. This model is the prerequisite for the design of an appropriate Extended Kalman Filter approach that allows for determining both the characteristic frequency components of each phoneme and their bandwidths.

This Kalman Filter-based frequency estimation, in general, needs to be able to account for the fact that phonemes can be classified into voiced and unvoiced sounds [5, 7, 14]. Voiced sounds (e.g. normal vowels) are characterized by several relatively sharp formant frequencies produced by vibrations of the vocal folds. In the case of voiced sounds, the vocal folds represent a laminar fluidic resistance against the outflow of air expelled from the lungs.

In contrast, unvoiced sounds (e.g. whispered vowels and fricatives such as *ch*, *ss*, *sch*, *f* in the words *Bach* (J.S. Bach, 1685–1750) [baχ], *Wasser* (eng. water) [ˈvasɐ], *Schiff* (eng. ship) [ʃɪf])[2] are caused by a turbulent, partially irregular, air flow with negligible vibrations of the vocal folds. To some extent, unvoiced sounds are produced by fizzing sounds originating between teeth and lips as well as between tongue and hard or soft palate. Due to the before-mentioned considerations motivated by principles from fluid mechanics, sharp formant frequencies are characteristic for voiced phonemes, whereas wide frequency bands are typical for unvoiced ones.

For both voiced and unvoiced phonemes, a stochastic filtering approach [7, 9] can be employed to estimate the expected values of the formant frequencies and their associated covariances, where the broad-band nature of unvoiced speech is reflected by (co-)variance estimates that are significantly larger than for the voiced case. Transitions between subsequent phonemes are then indicated by rapid changes in the above-mentioned estimation results [8].

In this contribution, an approach for the discrete-time modeling of speech signals is firstly reviewed in Sec. 2 together with a short description of the Extended Kalman Filter procedure that was published in [7]. On this basis, a novel interval algorithm

[1] Note that there exists a wide range of applications of computer-based assistance systems for speech therapy. Firstly, children with developmental language disorders — with a prevalence up to 50% in primary education [12] — require logopedic interventions. Secondly, speech therapy may be indicated for adults and elderly people with neurological diseases.

[2] The phonetic transcription of these German sample words is given by the symbols of the International Phonetic Alphabet (IPA).

is presented for the segmentation of speech signals into individual phonemes in Sec. 3 as well as for the matching of individual sounds against the phoneme features stored in a suitable reference database (Sec. 4). Besides a pattern recognition for correctly pronounced sounds, this procedure helps to identify and classify pronunciation disorders[3]. Here, expert knowledge of speech therapists will be inevitable if the features of a mispronounced phoneme are not yet included in the database. Moreover, the identification procedure is designed in such a way that it can be extended in future work to detect further disorders such as stuttering. This kind of disorder is characterized by (rapid) repetitions of individual and/ or multiple phonemes or syllables. From this point of view, characteristic features of stuttering can be detected from an occurrence count of the extracted phoneme features over a short time window. Numerical results for a benchmark speech signal highlight the algorithmic features in Sec. 5 before this contribution is concluded in Sec. 6 with an outlook on future work.

2 Mathematical Modeling and Frequency Analysis of Speech Signals

Phonemes are the basic building block of syllables, from which individual words are formed. They are characterized by specific features, namely, a speaker-dependent basis frequency and higher formant frequencies that are specific for each phoneme. Typically, higher formant frequencies are not integer multiples of the basis frequency. Moreover, the bandwidth of the included frequency ranges serves as a feature to distinguish between voiced and unvoiced phonemes.

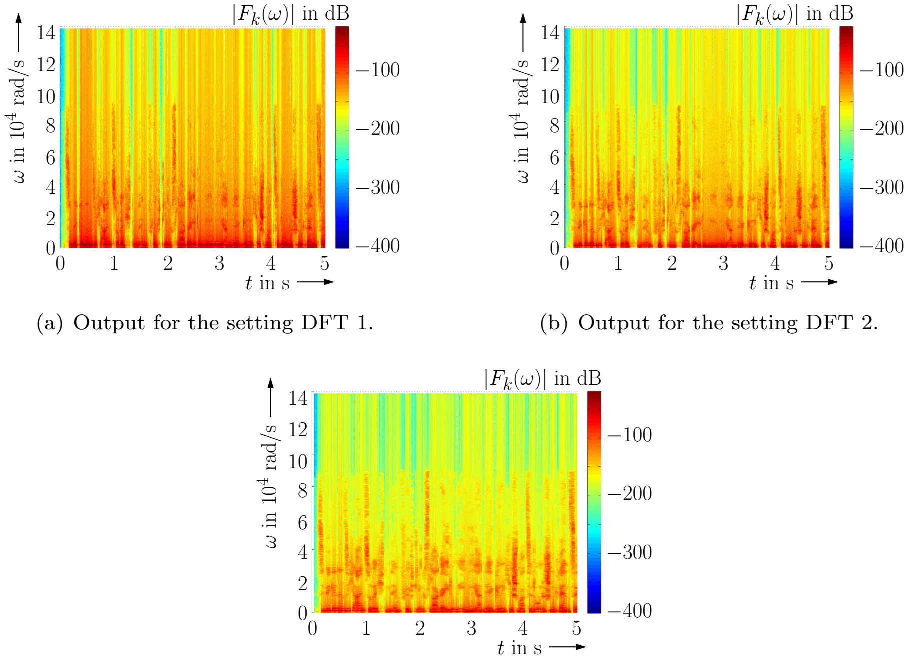
As mentioned in the introduction, voiced phonemes are, for example, *normal vowels*, that are characterized by several *sharp formant frequencies.* In contrast, unvoiced phonemes, such as *whispered vowels* and *fricatives* are characterized by *wide blurred formant frequency ranges.*

2.1 Fourier Analysis of Speech Signals

In the frame of automatic speech recognition systems [1, 4, 10], an offline frequency analysis is commonly performed, which consists of the following stages:

1. Cut the sound sequence into short temporal slices of typically $10 - 50\,\mathrm{ms}$ length
2. Perform a short-time Fourier analysis for each of these time slices (partly with overlapping time windows after the application of windowing functions aiming at the suppression of the leakage phenomenon)
3. Determine a measure of similarity with phoneme-dependent frequency spectra (usually by the application of cross-correlation functions in the frequency domain)

[3] Although a lot of research work exists in the frame of speech recognition (cf. [13] and the references therein), no assistance system is yet available that automatically detects and classifies pronunciation disorders. The reason is that most speech recognition systems perform a matching of speech signals against correctly pronounced reference patterns, leading to a replacement of mispronounced sounds by seemingly correct substitutes. However, this replacement obviously contradicts the aim of developing the desired assistance system for speech therapy.



(a) Output for the setting DFT 1.

(b) Output for the setting DFT 2.

(c) Output for the setting DFT 3.

Figure 1: DFT analysis of a benchmark speech signal [7].

For all numerical results concerning frequency analysis and pattern recognition of speech signals in this paper, a 5-second excerpt from a German TV news broadcast is used. It consists of an audio stream (pure speech of the anchorman without background sounds) sampled with the frequency $f_\mathrm{s} = 44.1\,\mathrm{kHz}$.

Results of an offline Discrete Fourier Transformation (DFT) of this signal are summarized in Fig. 1. There, frequency ranges which are characteristic for a specific phoneme become visible in a dark red color code. These frequencies correspond to sufficiently large values in the amplitude response $|F_k(\omega)|$ that is computed for windows of N equidistant sampling points of the speech signal. Moreover, large temporal variations of the frequency contents represent the transition points between different subsequent sounds in the speech signal. Hence, they coincide with those points of time at which sharp vertical lines appear in the spectrum in Fig. 1.

According to the parameters in Tab. 1, an increase in the number of sampling points N for each DFT evaluation leads to smaller values Δf of the spectral resolution. This improves the detectability of the formant frequencies (i.e., frequencies with large amplitude values) as well as their temporal variations in the speech signal (i.e., the before-mentioned vertical lines).

As mentioned above, points of time with significant changes in the spectrum represent candidates for the transition between two subsequent phonemes. These points of time are determined algorithmically in the following after the introduction of a

Table 1: Different parameterizations for the DFT analysis of a benchmark signal.

	DFT 1	DFT 2	DFT 3
Length of time window in samples (N)	512	1024	2048
Length of time window in ms	11.6	23.2	46.4
Sample shift between two DFT evaluations	64	64	64
Spectral resolution Δf in Hz	86.5	43.2	21.6

filter-based online frequency analysis which can serve as a computationally efficient substitute for the offline DFT.

2.2 Filter-Based Frequency Analysis

For the implementation of an online, real-time capable filter approach, the measured speech signal is represented by a superposition of different harmonic components according to

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

with the basis frequency $\omega_1 > 0$, further harmonic signal components $\omega_2, \ldots, \omega_n$, $\omega_{i+1} > \omega_i$, $i \in \mathbb{N}$, the signal amplitudes α_i, and the phase shifts ϕ_i.

The i-th component of the signal model (1) coincides with the solution of the continuous-time system model [6, 7]

$$\begin{aligned} \dot{x}_{3i-2}(t) &= -x_{3i}(t) \cdot \alpha_i \cdot \sin\left(x_{3i}(t) \cdot t + \phi_i\right) =: x_{3i-1}(t) \\ \dot{x}_{3i-1}(t) &= -x_{3i}^2(t) \cdot \alpha_i \cdot \cos\left(x_{3i}(t) \cdot t + \phi_i\right) \\ \dot{x}_{3i}(t) &= 0 \end{aligned} \tag{2}$$

that can be summarized in the state-space representation

$$\dot{\mathbf{x}}_i(t) = \mathbb{A}_i\left(x_{3i}(t)\right) \cdot \mathbf{x}_i(t) \ , \quad \mathbf{x}_i(t) = \begin{bmatrix} x_{3i-2}(t) \\ x_{3i-1}(t) \\ x_{3i}(t) \end{bmatrix} \in \mathbb{R}^3 \tag{3}$$

with the frequency-dependent system matrices

$$\mathbb{A}_i\left(x_{3i}(t)\right) = \left[\begin{array}{cc:c} 0 & 1 & 0 \\ -x_{3i}^2(t) & 0 & 0 \\ \hdashline 0 & 0 & 0 \end{array}\right] \in \mathbb{R}^{3\times 3} \quad . \tag{4}$$

For this set of state equations, the i-th component of the signal model (1) is given by

$$y_i(t) = \check{\mathbf{c}}_i^T \cdot \mathbf{x}_i(t) \quad \text{with} \quad \check{\mathbf{c}}_i^T = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \quad . \tag{5}$$

The complete signal model can then be formulated according to

$$\dot{\mathbf{x}}(t) = \mathbf{A}\left(\mathbf{x}(t)\right) \cdot \mathbf{x}(t) \quad \text{with} \quad \mathbf{x}(t) \in \mathbb{R}^{3n} \quad , \tag{6}$$

where the frequency-dependent system matrix corresponds to a block diagonal concatenation of the subsystem matrices defined in (4) according to

$$\mathbf{A}\left(\mathbf{x}(t)\right) = \begin{bmatrix} \mathbb{A}_1\left(x_3(t)\right) & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbb{A}_2\left(x_6(t)\right) & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbb{A}_n\left(x_{3n}(t)\right) \end{bmatrix} \in \mathbb{R}^{3n\times 3n} \quad . \tag{7}$$

A superposition of the individual signal components resulting from (3) and (5) leads to the output representation

$$y_{\mathrm{m},n}(t) = \mathbf{c}^T \cdot \mathbf{x}(t) \quad \text{with} \quad \mathbf{c}^T = \begin{bmatrix} \check{\mathbf{c}}_1^T & \check{\mathbf{c}}_2^T & \dots & \check{\mathbf{c}}_n^T \end{bmatrix} \ . \tag{8}$$

Under the assumption that the sampling frequency f_s is sufficiently large, temporal frequency variations can be neglected between two subsequent discretization points t_k and t_{k+1}. Therefore, a discrete-time system model is derived from (6), (7), and (8) in terms of the difference equations

$$\mathbf{x}_{k+1} = \exp\left(T_\mathrm{s} \cdot \mathbf{A}\left(\mathbf{x}_k\right)\right) \cdot \mathbf{x}_k + \mathbf{w}_k =: \mathbf{A}_k^\mathrm{d} \cdot \mathbf{x}_k + \mathbf{w}_k \tag{9}$$

with the measured output

$$y_k = y_{\mathrm{m},n,k} := y_{\mathrm{m},n}(t_k) = \mathbf{c}^T \cdot \mathbf{x}_k + v_k \ . \tag{10}$$

In the equations (9) and (10), the additive terms $\mathbf{w}_k$ and v_k are normally distributed noise processes

$$f_{w,k}\left(\mathbf{w}_k\right) = \mathcal{N}(\boldsymbol{\mu}_{w,k}, \mathbb{C}_{w,k}) \quad \text{and} \quad f_{v,k}\left(v_k\right) = \mathcal{N}\left(\mu_{v,k}, \mathbb{C}_{v,k}\right) \tag{11}$$

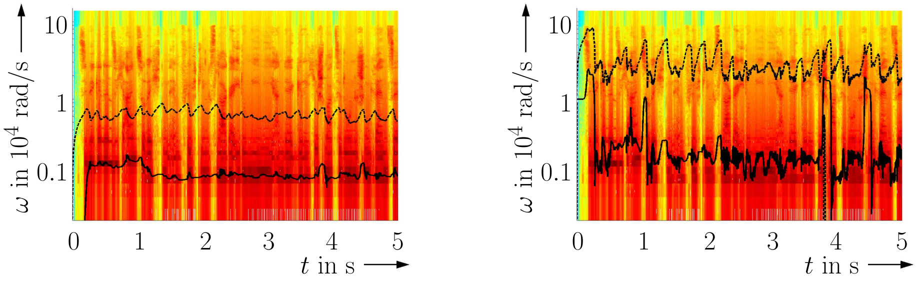
with zero mean values $\boldsymbol{\mu}_{w,k} = \mathbf{0}$ and $\mu_{v,k} = 0$ as well as the (co-)variances $\mathbb{C}_{w,k}$ and $\mathbb{C}_{v,k}$. These noise processes are taken into consideration in the design of an Extended Kalman Filter[4] (EKF) that is used for the online estimation of the formant frequencies ω_i and their associated standard deviations $\left(\sqrt{\mathbb{C}^e_{x,k,(3i,3i)}}\right)$. A typical estimation result for an EKF with $n = 2$ is shown in Fig. 2.

3 Phoneme-Based Segmentation of Speech Signals

3.1 Threshold Classification

Candidates for transition points between two subsequent phonemes [3] in a speech signal are characterized by significant variations in the formant frequencies (i.e., the entries in the vectors of estimated expected values) as well as in the bandwidth of the included frequency components (related to diagonal entries in the estimated covariance matrices). For that purpose, the following threshold estimator makes use of the Extended Kalman Filter outputs after a normalization by the corresponding average

[4] For linear dynamic systems with additive Gaussian process and measurement noise, the Kalman Filter provides the solution of Bayesian state estimation in terms of parameterizing the exact, normally distributed probability density functions of all state variables by their expected values and covariances. For nonlinear processes, the Bayesian estimation does usually not have analytic solutions. To obtain Gaussian approximations of the probability densities in such cases, the process and measurement models are locally approximated by linearized state-space representations with additive noise. The Extended Kalman Filter is thus the application of the Kalman Filter routine to this linearized model [11]. For the application of further stochastic filtering approaches aiming at the estimation of frequencies in speech signals, the reader is referred to [9].



(a) Estimate for ω_1.

(b) Estimate for ω_2.

Figure 2: Estimation results for $n = 2$: expected value $\mu^e_{x,k,3i}$ (solid lines) and upper frequency bound $\mu^e_{x,k,3i} + 3 \cdot \sqrt{\mathbb{C}^e_{x,k,(3i,3i)}}$ (dashed lines) [7].

values over a windows of L samples (typically the complete speech signal under consideration). Hence, the *normalized expectation* of the i-th formant frequency and the corresponding *normalized (co-)variance* are given by

$$\tilde{\mu}^e_{x,k,3i} = \frac{\mu^e_{x,k,3i}}{\frac{1}{L}\sum\limits_{k=1}^{L} \mu^e_{x,k,3i}} \quad \text{and} \quad \tilde{\mathbb{C}}^e_{x,k,(3i,3i)} = \frac{\mathbb{C}^e_{x,k,(3i,3i)}}{\frac{1}{L}\sum\limits_{k=1}^{L} \mathbb{C}^e_{x,k,(3i,3i)}} \ , \quad i = 1, \ldots, n \ . \tag{12}$$

Based on the absolute variation rate

$$\Delta\tilde{\mu}^e_{x,k,3i} = \left|\tilde{\mu}^e_{x,k+1,3i} - \tilde{\mu}^e_{x,k,3i}\right| \tag{13}$$

of the estimated frequencies as well as the absolute variation rate

$$\Delta\tilde{\mathbb{C}}^e_{x,k,(3i,3i)} = \left|\tilde{\mathbb{C}}^e_{x,k+1,(3i,3i)} - \tilde{\mathbb{C}}^e_{x,k,(3i,3i)}\right| \tag{14}$$

of the corresponding (co-)variances, candidates for phoneme boundaries are characterized by

$$\Delta\tilde{\mu}^e_{x,k,3i} > \overline{\Delta\tilde{\mu}} \quad \text{and} \quad \Delta\tilde{\mathbb{C}}^e_{x,k,(3i,3i)} > \overline{\Delta\tilde{\mathbb{C}}} \ , \tag{15}$$

where either of the absolute variations (13) and (14) exceeds empirically chosen threshold values $\overline{\Delta\tilde{\mu}}$ or $\overline{\Delta\tilde{\mathbb{C}}}$.

According to [8], both criteria in (15) can be merged into a joint detection routine for possible phoneme boundaries. Phoneme boundaries are set at those points in the speech signal where the inequalities above are satisfied with a minimum temporal distance $\Delta T_{\min}$ between two subsequent boundaries. For typical human speech signals, this temporal distance can be chosen in the order of magnitude of $\Delta T_{\min} = 20\,\text{ms}$.

3.2 Branch-and-Bound Procedure

As an alternative to the previous threshold-based detection scheme, a novel interval-based branch-and-bound procedure is introduced in this subsection. In analogy to the previous classification scheme, a stochastic frequency estimation is performed in a first stage by using the Extended Kalman Filter, cf. Sec. 2.

Then, the following algorithm is applied.

Step 1 Determine the convex interval hull (CIH) for each of the estimated mean values and standard deviations of the analyzed speech signal in terms of mutually independent intervals for each frequency. These intervals are defined such that their bounds correspond to the smallest, respectively largest, values of the quantity of interest over the considered time span.

Step 2 Bisect the time interval if

- the length is larger than $\Delta T'_{\min} < \Delta T_{\min}$ and
- either of the maximum variation rates

$$\max_{k\in\mathcal{K}} \left\{\tilde{\mu}^e_{x,k,3i}\right\} - \min_{k\in\mathcal{K}} \left\{\tilde{\mu}^e_{x,k,3i}\right\} > \overline{\Delta\tilde{\mu}} \tag{16}$$

or

$$\max_{k\in\mathcal{K}} \left\{\tilde{\mathbb{C}}^e_{x,k,(3i,3i)}\right\} - \min_{k\in\mathcal{K}} \left\{\tilde{\mathbb{C}}^e_{x,k,(3i,3i)}\right\} > \overline{\Delta\tilde{\mathbb{C}}} \tag{17}$$

exceeds given threshold values, where $\mathcal{K}$ denotes the range of time indices for the considered temporal slice.

- A subsequent merging of neighboring time intervals is possible if the inequalities (16) and (17) are not fulfilled for the union of these two neighboring temporal slices.

4 Phoneme Identification

4.1 Relevant Features

In order to identify the phoneme that is located between its before-mentioned boundaries, it is necessary to compare its relevant features with the ones that are stored in an offline generated reference database.

To avoid speaker-dependent influences, all formant frequencies and standard deviations are normalized by a multiplicative scaling factor so that the mean of the estimated frequency ω_1 over the complete speech signal corresponds to the mean of ω_1 in the reference database. This procedure is admissible for sufficiently long audio streams.

The most relevant phoneme features are represented by a CIH (cf. *Step 1* in Sec. 3.2) over the expected values of each formant frequency for the complete phoneme duration as well as by the CIH over the respective standard deviations of each formant frequency and their associated amplitudes.

As soon as a single phoneme is characterized by multiple temporal subintervals due to the segmentation routines in Sec. 3, a weighted arithmetic average of the interval values is used, where the corresponding subinterval durations serve as weighting factors for the following pre-classifier as well as for *Algorithm 1* in Sec. 4.4. Alternatively, a CIH can be formed over all subslices as well.

4.2 Database of Reference Values

To compare the estimated CIHs against some reference values, it is necessary to create a database of features that represent the phonemes of a specific language of interest. For example, for the German language from which the benchmark scenario in Sec. 2 was chosen, a list of approximately 75 characteristic, different phonemes exists (including, for example, short and long variants of a vowel as separate database entries). The corresponding features are stored in terms of their CIH in a reference database, where naturally spoken language without pronunciation disorders serves as the reference. To avoid a bias due to random errors in this reference database, the selected features need to be observed multiple times during the creation of the reference. The typical interval features are then obtained by averaging the corresponding CIHs or by forming a further CIH after normalization to the average of the basis frequency.

4.3 Pre-Classification of Phonemes by Interval-Based Similarity Measures

4.3.1 Unvoiced Phonemes

As mentioned in Sec. 2, unvoiced phonemes are characterized by *large* estimated standard deviations (especially for $i \geq 2$). Based on the normalized standard deviation of the second formant $[\tilde{\sigma}_2^e]$ as well as based on the duration τ of a phoneme, a pre-classification scheme can be implemented to distinguish between candidates that most likely are unvoiced sounds and definitely not voiced ones. This pre-classification uses the following criteria with the threshold value σ^* chosen by comparison of the DFT frequency spectra of unvoiced and voiced phonemes:

Criterion 1:

$$\frac{\sup\{[\tilde{\sigma}_2^e]\} - \sigma^*}{\sigma^* - \inf\{[\tilde{\sigma}_2^e]\}} > 0.5 \tag{18}$$

Criterion 2:

$$(\inf\{[\tilde{\sigma}_2^e]\} > \sigma^*)\,\&\,(\tau < 65\,\mathrm{ms}) \tag{19}$$

Either Criterion 1 or Criterion 2 has to be satisfied in order to denote the current phoneme as a candidate for an unvoiced sound.

4.3.2 Voiced Phonemes

In contrast to unvoiced phonemes, voiced phonemes are characterized by *small* estimated standard deviations (especially for $i \geq 2$).

Besides the interval for the normalized standard deviation of the second formant $[\tilde{\sigma}_2^e]$, intervals for the estimated formant frequencies $[\tilde{\omega}_1^e]$ and $[\tilde{\omega}_2^e]$ as well as the corresponding estimated amplitudes $[\tilde{\alpha}_1^e]$, $[\tilde{\alpha}_2^e]$ and the phoneme duration τ are included in the pre-classifier for voiced phonemes. The resulting criterion is given by:

Criterion:

$$\begin{aligned}(\sup\{[\tilde{\sigma}_2^e]\} < \sigma^*)\&\,(\inf\{[\tilde{\omega}_2^e]\} > \sup\{[\tilde{\omega}_1^e]\})\ldots\\ \&\,(\tau \geq 50\,\mathrm{ms})\,\&\left(\frac{\sup\{[\tilde{\alpha}_1^e]\}}{\sup\{[\tilde{\alpha}_2^e]\}} > 0.8\right)\end{aligned} \tag{20}$$

Note that the results that are summarized in Sec. 5.2 for the pre-classification of phonemes are restricted to using the option of averaged interval boxes for the relevant phoneme features if a phoneme is represented by several subintervals.

4.4 Matching of Phonemes

4.4.1 Interval Box Representation of Phoneme Features: Algorithm 1

The matching of phonemes with the entries that are stored in a reference database can be performed by the evaluation of the following similarity measure. It makes use of the short-hand notation

$$\operatorname{vol}\left(\operatorname{diam}\{[\mathbf{z}]\}\right) = \prod_{i=1}^{n} \left(\overline{z}_i - \underline{z}_i\right) \tag{21}$$

for the pseudo-volume of an n-dimensional interval box $[\mathbf{z}]$.

The feature box $[\mathbf{z}]_{\mathrm{ref},j}$ for the j-th phoneme in the reference database ($j = 1, \ldots, j_{\max}$) contains the normalized intervals of formant frequencies, estimated standard deviations, and amplitudes over the complete phoneme duration.

With the help of the feature box $[\mathbf{z}]$ of the current phoneme, the similarity measure

$$\rho_j = \frac{\operatorname{vol}\left(\operatorname{diam}\{[\mathbf{z}]\}\right) + \operatorname{vol}\Big(\operatorname{diam}\{[\mathbf{z}]_{\mathrm{ref},j}\}\Big)}{\operatorname{vol}\Big(\operatorname{diam}\{[\mathbf{z}] \cup [\mathbf{z}]_{\mathrm{ref},j}\}\Big)} \tag{22}$$

is evaluated for each $j = 1, \ldots, j_{\max}$. The ratio ρ_j in (22) is a measure for the degree of similarity between the current feature box and the j-th reference in terms of maximum overlapping with a minimum excess volume between the volume of the convex axis-aligned hull of two interval boxes denoted by the operator $\cup$ and the sum of the individual volumes.

Hence, the best match for the corresponding phoneme is the maximizer of the similarity measure according to

$$j^* = \underset{j=1,\ldots,j_{\max}}{\arg\max} \left\{\rho_j\right\} \quad . \tag{23}$$

4.4.2 Interval Likelihood Representation: Algorithm 2

As an alternative algorithm for phoneme matching, interval likelihood functions can be used as a generalization of a point-valued normal distribution for the phoneme feature representation in the time step k. A point-valued probability density function with corresponding mean values and covariances determined with the help of the Extended Kalman Filter algorithm is given by

$$f_{x,k}\left(\mathbf{x}_k\right) = \frac{1}{\sqrt{(2\pi)^{3n} \left|\mathbb{C}_{x,k}\right|}} \exp\left\{-\frac{1}{2}\left(\mathbf{x}_k - \boldsymbol{\mu}_{x,k}\right)^T \mathbb{C}_{x,k}^{-1}\left(\mathbf{x}_k - \boldsymbol{\mu}_{x,k}\right)\right\} \quad . \tag{24}$$

A reasonable generalization to an interval-based likelihood representation can be defined for the complete phoneme according to

$$f_{r,\mathcal{K},j} \in \left[f_{r,\mathcal{K}}\right]_j := w_j \cdot \frac{1}{\sqrt{(2\pi)^{3n} \left|\left[\mathbb{S}_{x,\mathcal{K}}\right]_j\right|}} \exp\left\{-\frac{1}{2}\left[\mathbf{r}_{x,\mathcal{K}}\right]_j^T \left[\mathbb{S}_{x,\mathcal{K}}\right]_j^{-1} \left[\mathbf{r}_{x,\mathcal{K}}\right]_j\right\} \quad , \tag{25}$$

where

$$\left[\mathbf{r}_{x,\mathcal{K}}\right]_j := \left[\boldsymbol{\mu}_{x,\mathcal{K}}\right] - \left[\boldsymbol{\mu}_{x,\mathcal{K}}\right]_{\mathrm{ref},j} \tag{26}$$

is an interval residuum value and

$$[\mathbb{S}_{x,\mathcal{K}}]_j := [\mathbb{C}_{x,\mathcal{K}}] + [\mathbb{C}_{x,\mathcal{K}}]_{\mathrm{ref},j} \tag{27}$$

a suitable interval definition of the error (co-)variance. Both the interval residuum $[\mathbf{r}_{x,\mathcal{K}}]_j$ in (26) and the interval covariance $[\mathbb{S}_{x,\mathcal{K}}]_j$ in (27) are defined as CIHs over the complete index set $k \in \mathcal{K}$ of the corresponding phoneme duration. In a fundamental setting, each scaling factor $w_j > 0$ is set to $w_j \equiv 1$.

With the information obtained from the interval likelihood functions (25), the most likely phoneme is given by

$$j^* = \underset{j=1,\ldots,j_{\max}}{\arg\max} \left\{ \sup \left\{ [f_{r,\mathcal{K}}]_j \right\} \right\} \quad , \tag{28}$$

which detects the maximum supremum of all possible interval likelihood functions $[f_{r,\mathcal{K}}]_j$ for the complete list of phoneme database entries $j = 1, \ldots, j_{\max}$.

In addition to using the complete estimated feature vector (provided by the Extended Kalman Filter routine in Sec. 2), also projections of the interval residua and the corresponding interval (co-)variances onto a subspace formed by individual components of $\boldsymbol{\mu}_{x,\mathcal{K}}$ are possible. These components have to be chosen such that they carry the most relevant information for detecting the phonemes of interest. Typically, this information is given by the estimated frequencies and the duration of the phoneme. Information about the duration is necessary, if distinctions between short and long sounds such as short and long vowels are made.

A necessary prerequisite for the applicability of this second classification algorithm is the invertability of the interval matrices $[\mathbb{S}_{x,\mathcal{K}}]_j$ defined in (27) for all $j = 1, \ldots, j_{\max}$. Although this is ensured for time horizons $\mathcal{K}$ that consist only of a single time step k, interval-related overestimation due to the use of CIHs over the complete duration of a phoneme may lead to a loss of regularity of the corresponding interval matrix. If this loss of regularity occurs, the first classification procedure can be used as a fallback solution.

As it is shown in Sec. 5.3, the use of factors $w_j \neq 1$ can noticeably enhance the detection rates for phonemes which can hardly be distinguished by residuum vectors (26) that only consist of the estimated formant frequencies.

5 Results

5.1 Phoneme-Based Segmentation of the Speech Signal

As a fundamental prerequisite for the matching of phonemes, the possible boundaries of the individual phonemes need to be detected. For the phoneme-based segmentation of the benchmark speech signal, the two possible options of using either a pure threshold segmentation (Sec. 3.1) or the interval-based segmentation (Sec. 3.2) are compared.

The results of both algorithms are then compared to phoneme boundaries obtained manually from repeated listening to excerpts of the speech signal and to the visual frequency variations in the DFT output (cf. the vertical band structure in Fig. 1).

For the automatic segmentation according to Sec. 3, the formant frequencies are firstly determined by means of the Extended Kalman Filter procedure. It provides information about the frequencies up to the order n as well as the corresponding (co-)variances and the signal amplitudes.

As shown in Fig. 3, the threshold classification procedure allows for an accurate detection of phoneme boundaries. Here, the boundaries determined manually by the human listener (marked by red vertical lines) are in good coincidence with the automatic segmentation and the DFT results. This statement is equally true for minimum temporal slices chosen from the set $\{15\,\mathrm{ms}, 20\,\mathrm{ms}, 25\,\mathrm{ms}\}$.

At first glance, the branch-and-bound version in Fig. 4 leads to very similar results. However, due to the splitting and merging steps of time intervals described in Sec. 3.2, the algorithm determines the visible boundary points in the DFT spectrum more accurately than the pure threshold implementation. Moreover, it has the advantage that also points of time with characteristic variations of formant frequencies and bandwidths within a single phoneme are determined more accurately.

This is also confirmed by the fact that the lengths of the individual subintervals are less homogeneously distributed than for the algorithm according to Sec. 3.1. The distribution of the estimated lengths of the time slices between points with significant feature variations is exemplarily shown for the branch-and-bound procedure in Fig. 5.

Note that — without the automatic segmentation procedure — inner phoneme boundaries could only be extracted from the offline DFT. For a human listener, these boundaries are not detectable at all, since phonemes can only be perceived as a whole.

Concerning the matching of phonemes, these intermediate points of significant frequency variations are especially useful to detect plosives that only become audible for a human listener by their combination with a preceding or subsequent voiced phoneme (typically a vowel). This is usually the case for bilabial stops such as the voiced bilabial plosive *b* ([b], [b̥]) or the unvoiced bilabial plosive *p* ([p], [pʰ]) with a preceding or following vowel. These phoneme combinations are commonly characterized by significant variations of at least one of the formant frequencies over the duration of the complete joint sound. Examples, for which these variations can be perceived, are [ba], [da], [ga], [di], [da], [du].

5.2 Results of the Phoneme Pre-Classification

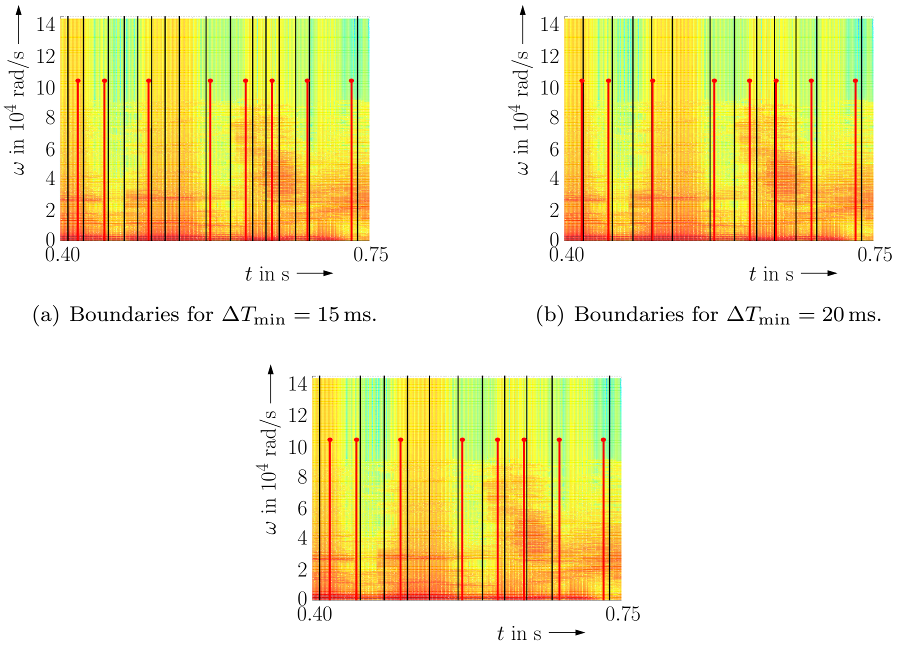
In this subsection, the results of the phoneme pre-classification are presented. The audio signal consists of a 2-second excerpt from the news broadcast already used in Sec. 2. From this excerpt, the following outcome was obtained, where the repetition of sounds represents their multiple occurrences:

- Unvoiced: `‘s’ ‘n’ ‘n’ ‘s’ ‘n’ ‘ch’ ‘t’`
- Voiced: `‘i’ ‘i’ ‘e’ ‘n’`
- Undecided: `‘b’ ‘e’ ‘g’ ‘e’ ‘i’ ‘ch’ ‘ei’ ‘z’ ‘u’ ‘r’ ‘i’`

In IPA transcription, the respective sounds are:

- Unvoiced: [z] [n] [n] [z] [n] [ç] [t]
- Voiced: [iː] [ɪ] [ə] [n̩]
- Undecided: [b] [əˈ] [g] [ə] [ɪ] [ç] [aɪ̯] [ts] [uː] [ʀ] [ɪ]

Here, all sounds pre-classified as unvoiced were actually unvoiced phonemes. In the set of voiced ones, only the phoneme `’n’` ([n̩]) is misclassified. This sound is hard to detect even for a human listener since it follows a short voiced `’i’` ([ɪ]). Due to some jitter in the detection of the phoneme boundaries, this voiced sound `’i’` was mistakenly included in the list of undecided candidates.



(a) Boundaries for $\Delta T_{\min} = 15\,\text{ms}$.

(b) Boundaries for $\Delta T_{\min} = 20\,\text{ms}$.

(c) Boundaries for $\Delta T_{\min} = 25\,\text{ms}$.

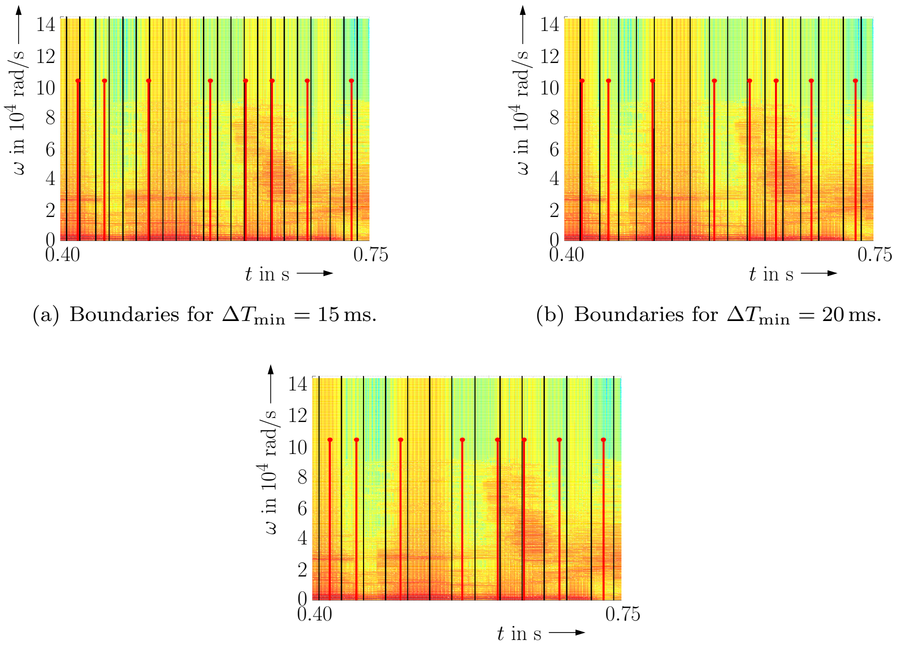
Figure 3: Comparison of manually detected phoneme boundaries (red vertical lines) with the automatic threshold classification results (black lines).

Due to the separation of phonemes with the attributes *voiced*, *unvoiced*, and *undecided*, the pre-classification stage significantly reduces the effort of the following phoneme matching. For the classes *voiced* and *unvoiced*, only the corresponding subsets from the reference database need to be investigated, while a full list search only becomes necessary for *undecided* phonemes.

In future work, the pre-classification criteria will be extended by some probabilistic confidence measures that aim at the reduction of the misclassification likelihood of sounds such as for the unvoiced `'n'` above.

5.3 Results of the Phoneme Matching Procedure

Fig. 6 gives a comparison of the two phoneme matching procedures. For the sake of compactness, only a matching of selected vowels is investigated. It can be seen that the interval likelihood representation in *Algorithm 2* clearly reduces the misclassification rate in comparison with the box-valued *Algorithm 1* if it is applied in a two-stage manner. Firstly, $w_j \equiv 1$ is used to rule out all unlikely phonemes (Fig. 6(b)). Candidates within a certain distance from the most likely phoneme are kept for the evaluation



(a) Boundaries for $\Delta T_{\min} = 15\,\text{ms}$.

(b) Boundaries for $\Delta T_{\min} = 20\,\text{ms}$.

(c) Boundaries for $\Delta T_{\min} = 25\,\text{ms}$.

Figure 4: Comparison of manually detected phoneme boundaries (red vertical lines) with the automatic branch-and-bound segmentation procedure (black lines).

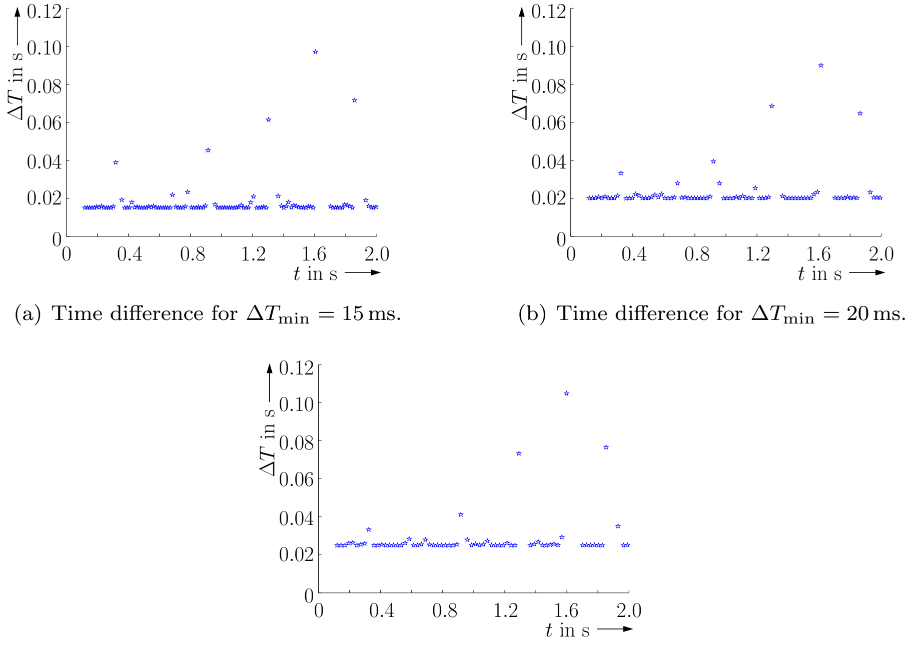
and are investigated further by choosing the weighting factors now as

$$w_j = \frac{1}{\sum\limits_{i=1}^{n} \left(\operatorname{diam}\left\{ \left[\mu^e_{x,\mathcal{K},3i-2}\right] \right\} - \operatorname{diam}\left\{ \left[\mu_{x,\mathcal{K},3i-2}\right]_{\mathrm{ref},j} \right\} \right)} \ . \tag{29}$$

In order to reduce the number of remaining misclassifications further, it is promising to improve the quality of the reference database by extracting the phoneme features from a larger number of frequency estimates. Currently, the reference database is made up only of averages of up to five recordings of each of the investigated sounds so that outliers can still have a negative impact on the matching quality.

6 Conclusions and Outlook on Future Work

In this paper, the first building blocks of a signal processing unit were presented which is currently under development for the automatic classification of pronunciation disorders in speech therapy. When the detection of mispronounced phonemes is of interest in a therapeutic context, it is not possible to use state-of-the-art speech recognition systems based on pattern recognition procedures such as those presented in [13]. The reason for this is obvious: As soon as replacements of mispronounced sounds by seemingly correct ones occur during the comparison of the frequency features of a speech



(a) Time difference for $\Delta T_{\min} = 15\,\mathrm{ms}$.

(b) Time difference for $\Delta T_{\min} = 20\,\mathrm{ms}$.

(c) Time difference for $\Delta T_{\min} = 25\,\mathrm{ms}$.

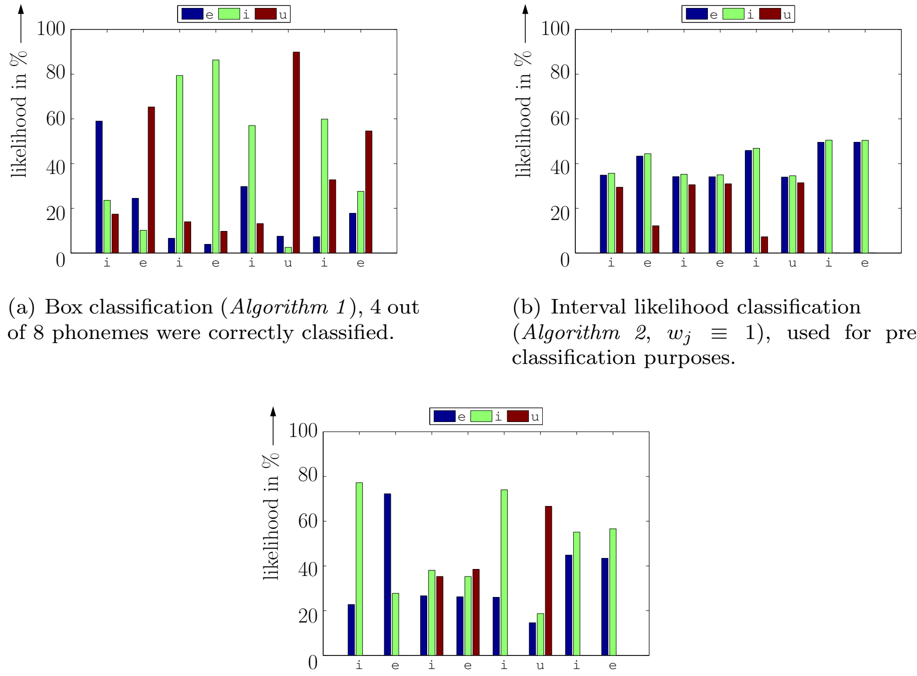
Figure 5: Temporal distance between two subsequent segmentation points for the branch-and-bound procedure.

signal with words that are included in some reference dictionary, every mispronounced sound is replaced by something that is seemingly correct. However, exactly the information of pronunciation disorders[5] is of interest for the therapist to be assisted in her/ his everyday work. Therefore, also learning-type approaches from artificial intelligence aiming at an automatic correction of mispronunciation are not reasonable for the development of the assistance system *SUSE* (A *S*oftware assistance system for *U*ncovering speech disorders by *S*tochastic *E*stimation techniques).

For this reason, an online filter-based frequency estimation procedure was presented which allows for determining characteristic features on the level of individual phonemes such as formant frequencies and their estimated (co-)variances. Here, the (co-)variance information is especially useful to distinguish between voiced and unvoiced sounds in a pre-classification phase. Moreover, various results have shown that novel interval-based segmentation procedures for the detection of phoneme boundaries and interval-likelihood representations of phoneme features can be used to reliably determine the boundaries of individual sounds and to find those sounds from a list of reference signals which represent the actually observed sound with maximum probability.

Future work firstly has to deal with extending the phoneme database to a complete

[5] Note that such replacements may be admissible in cases of a pure grammar analysis which was not considered in this paper.



(a) Box classification (*Algorithm 1*), 4 out of 8 phonemes were correctly classified.

(b) Interval likelihood classification (*Algorithm 2*, $w_j \equiv 1$), used for pre-classification purposes.

(c) Interval likelihood classification (*Algorithm 2*, $w_j \neq 1$ acc. to (29)), 6 out of 8 phonemes were correctly classified.

Figure 6: Overview of the outcomes of the different phoneme matching procedures, where the correct phonemes are listed below the horizontal axis.

list of all sounds that are relevant for a specific language (e.g. approximately 75 fundamental sounds for the German language). Moreover, detection rates of individual phonemes need to be thoroughly analyzed if a speaker-independent reference database is employed and if this is equally applied to analyze the speech of various speakers including children, adults, and elderly persons, each of them with and without pronunciation disorders. Besides the implementation of further estimation procedures (such as Extended Kalman Filters with a bandpass filtering of the input signal aiming at an improved observability as shown in [7, 9]) and the use of multi-hypothesis filter techniques [11], options will be investigated to include self-learning features into the classification scheme. There, especially mispronounced sounds not yet included in the database, and weakly distinguishable sounds will be in the focus. In both cases, the therapist will have to provide expert knowledge in order to trigger the learning by rejecting possibly incorrect estimates made by the assistance system. One possibility for this will be the adaptation of weighting factors during the learning phase such as those included in Sec. 5.3.

References

[1] M. Cutajar, E. Gatt, I. Grech, O. Casha, and J. Micallef. Comparative Study of Automatic Speech Recognition Techniques. *IET Signal Processing*, 7(1):25–46, 2013.

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

[3] G. Gosztolya and L. Tóth. Detection of phoneme boundaries using spiking neurons. In L. Rutkowski, R. Tadeusiewicz, L. A. Zadeh, and J. M. Zurada, editors, *Proc. of 9th Intl. Conference on Artificial Intelligence and Soft Computing – ICAISC 2008: Zakopane, Poland, June 22-26, 2008*, pages 782–793. Springer, Berlin, Heidelberg, 2008.

[4] F. Jelinek. *Statistical Methods for Speech Recognition.* Language, Speech, & Communication: A Bradford Book. MIT Press, Cambridge, MA, 1997.

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

[6] A. Rauh, J. Ehret, and H. Aschemann. Observer-Based Real-Time Frequency Analysis for Combustion Engine-Based Power Trains with Applications to Identification and Control. In *Proc. of 19th IEEE Intl. Conference on Methods and Models in Automation and Robotics MMAR 2014*, Miedzyzdroje, Poland, 2014.

[7] 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.

[8] A. Rauh, S. Tiede, and C. Klenke. 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.

[9] 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.

[10] W. Reichl and G. Ruske. Syllable Segmentation of Continuous Speech with Artificial Neural Networks. In *Third European Conference on Speech Communication and Technology*, 1993.

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

[12] 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*, 5(1):21–39, 2017.

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

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