Contextual invariant-integration features for improved speaker-independent speech recognition

Florian Müller*, Alfred Mertins

Institute for Signal Processing, University of Lübeck, Ratzeburger Allee 160, 23538 Lübeck, Germany

Abstract

This work presents a feature-extraction method that is based on the theory of invariant integration. The invariant-integration features are derived from an extended time period, and their computation has a very low complexity. Recognition experiments show a superior performance of the presented feature type compared to cepstral coefficients using a mel filterbank (MFCCs) or a gammatone filterbank (GTCCs) in matching as well as in mismatching training-testing conditions. Even without any speaker adaptation, the presented features yield accuracies that are larger than for MFCCs combined with vocal tract length normalization (VTLN) in matching training-test conditions. Also, it is shown that the invariant-integration features (IIFs) can be successfully combined with additional speaker-adaptation methods to further increase the accuracy. In addition to standard MFCCs also contextual MFCCs are introduced. Their performance lies between the one of MFCCs and IIFs.

Key words: Speech recognition, speaker-independency, invariant-integration

1. Introduction

A wide variety of applications of *automatic speech recognition* (ASR) systems can be found nowadays. Though major advances are reported regularly, generally, the performance of ASR systems is still far below the human one, which degrades the user acceptance. Different reasons for the performance lag can be made up; these can be, for example, sensitivities to environmental noise or to the characteristics of the speech-transmission

*Corresponding author.

Email addresses: mueller@isip.uni-luebeck.de (Florian Müller*), mertins@isip.uni-luebeck.de (Alfred Mertins)

channel. A general problem in speaker-independent ASR is the high variability that is inherent in human speech. Benzeghiba et al. [1] give a detailed review of different kinds of variabilities in ASR. Two broad groups of variabilities can be defined: extrinsic (non-speaker related) and intrinsic (speaker-related) variabilities. Environmental noise and transmission artifacts are two examples of extrinsic variabilities. Besides varying accents, speaking-styles and -rates, age, and emotional state, it is the shape of the vocal tract that intrinsically contributes to the variable occurrence of speech signals representing the same textual content. The problems originating from different *vocal tract lengths* (VTLs) become especially apparent in mismatching training-testing conditions. For example, if children use an ASR system whose acoustic models have only been trained with adult data, the recognition performance degrades significantly compared to the performance of adult users. Therefore, in speaker-independent ASR systems, one often uses speaker-adaptation techniques to reduce the influence of speaker-related variabilities. There is ongoing work for addressing each of the variabilities mentioned above. The work presented in this paper focuses on the VTL as a source of variability between individual speakers.

A common model of human speech production is the source-filter model [15]. In this model, the source corresponds to the air stream originated from the lungs and the filter corresponds to the vocal tract, which is located between the glottis and the lips, and is composed of different cavities. The VTL describes the distance between the glottis and the lips. On average, the VTL is about 14 cm for adult women and 17 cm for men [2]. The locations of the vocal tracts' resonance frequencies (the "formants") shape the overall short-time spectrum and define the phonetic content. The spectral effects of different VTLs have been widely studied, see [1] and references therein. An important observation is that, while the absolute formant positions of individual phones are speaker specific, their relative positions for different speakers are somewhat constant. A relation that describes this observation is given by considering a uniform tube model with length l. Here, the resonances occur at frequencies $F_i = c \cdot (2i-1)/(4l)$, $i = 1, 2, 3, \ldots$, where c is the speed of sound [7]. Using this model, the spectra S of the same utterance from two speakers A and B with different VTLs are related by a scaling factor α_S, which is also known as the

frequency-warping factor:

$$S_A(\omega) = S_B(\alpha_S \cdot \omega). \tag{1}$$

Though Eq. (1) is only a rough approximation for the real relationship between spectra from speakers with different VTLs, methods that try to achieve speaker independency for an ASR system commonly take this relationship as their fundamental assumption.

A TF analysis of the speech signal is usually the first operation in an ASR feature-extraction stage after possible preprocessing steps such as pre-emphasis or noise cancellation. This analysis tries to simulate the human auditory system up to a certain degree, and different methods have been proposed. As it is done for the computation of the well-known *mel frequency cepstral coefficients* (MFCCs), a basic approach is the use of the *fast Fourier transformation* (FFT) applied on windowed short-time signals whose output is weighted by a set of triangular bandpass filters in the spectral domain [15]. Another common filterbank approach uses gammatone filters [35]. These filters were shown to fit the impulse response of the human auditory filters well. Both types of TF analysis methods have in common that they locate the center frequencies of the filters evenly spaced on nonlinear auditory motivated scales; in case of the MFCCs the mel scale is used [6], and in case of a gammatone filterbank the *equivalent rectangular bandwidth* (ERB) scale is used [18, 29, 35]. Different works make use of the observation that both the mel and the ERB scale approximately map the spectral scaling as described in Eq. (1) to a translation along the subband-index space of the TF analysis. More details will be given below.

Present methods that try to achieve speaker independency can be roughly grouped into three categories. These groups act on different stages of the ASR process and often may be combined within the same ASR system. One group tries to normalize the features after the extraction [22, 46, 49] by estimating the implicit warping factors of the utterances. These techniques are commonly referred to as *VTL normalization* (VTLN) methods. A second group of methods adapts the acoustic models to the features of each utterance [10, 23]. The use of *maximum-likelihood linear regression* (MLLR) methods is part of most state-of-the-art recognition systems nowadays. It was shown in [36] that certain types of VTLN methods are equivalent to constrained MLLR. The third group of methods works on the feature-extraction stage and tries to compute features that are speaker independent.

In theory, this group of methods does not need an additional speaker-adaption step and, hence, promises lower computational costs than the first two mentioned groups of methods. While this is true in principle, practical experiments carried out in this work show that even these methods may benefit from further speaker-adaptation techniques.

The concept of computing features that are independent of the VTL has been taken up by several works in the past, and different methods were proposed. In the following, a brief summary of these ideas is presented.

To begin with, Cohen [5] introduced the scale transformation which was further investigated for its applicability in the field of ASR by Umesh et al. [47]. Its use in ASR is motivated by the relationship given in Eq. (1). One property of the scale transformation is that the magnitudes of the transformations of two scaled versions of one and the same signal are the same. Thus, the magnitudes can be seen to be scaling invariant. The scale cepstrum, which has the same invariance property, was also introduced in the work of Umesh et al. [47]. The scale transformation is a special case of the Mellin transformation. The work of Patterson [34] describes a so-called *auditory image model* (AIM) that was extended with the Mellin transformation in the work of Irino and Patterson [17]. Further studies about the Mellin transformation have been conducted, for example, by Sena and Rocchesso [44].

Various works rely on the assumption that the effects of inter-speaker variability caused by VTL differences is mapped to translations along the subband-index space of an appropriate filterbank analysis [26–28, 30–32, 37]. Mertins and Rademacher [26, 27] used a wavelet transformation for the *time-frequency* (TF) analysis and proposed so-called *vocal tract length invariant* (VTLI) features based on auto- and cross-correlations of wavelet coefficients. Subsequent work [37] showed that a gammatone filterbank instead of a previously used wavelet filterbank leads to a higher robustness against VTL changes.

Previous works of the authors of this manuscript investigated translation-invariant transformations that were originally developed within the field of image analysis [30–32]. In the present work, we propose a feature-extraction method that is based on the principle of invariant integration. We refer to these features as *invariant-integration features* (IIFs) in the following. While preliminary investigations [31] about this method showed the

applicability as a proof-of-concept, here the method is consequently enhanced and studied in detail. Formally, the features are designed to be invariant to translations along the subband-index space of the TF representation. The method for their computation is based on the general approach of integrating feature functions over a group action, which is mathematically well founded. It is shown that IIFs can be found that are more robust to different VTLs than standard MFCCs and that, with an optimum parameter choice, even perform superior to MFCCs in combination with speaker adaptation. Besides a comparison to MFCCs in combination with VTLN and MLLR, this article gives a detailed study about the influence of the parameters on the recognition performance. Furthermore, its robustness in different training-testing scenarios as well as their performance on different corpora is investigated.

The following section explains general notions of the theory of invariants and describes three canonical approaches for obtaining invariants. Section 3 describes the approach of invariant integration and shows how it can be applied to the field of speaker-independent ASR. In that section the IIFs are formally defined. Starting with a basic definition of an IIF founded on a theoretical basis, it is further developed to incorporate contextual information for an ASR task. Since the proposed method has a high degree of freedom for its choice of parameters, an appropriate feature-selection method is described. The fourth part of this article describes the experiments together with their results. In this section, in addition to contextual IIFs, also the contextual selection of MFCCs is investigated. Conclusions and an outlook to future work are given in the last section.

2. Invariant features and their construction

Nonlinear transformations that lead to invariant features have been investigated and successfully applied for decades in the field of pattern recognition. A brief introduction to the general notions and concepts for the construction of invariants is given in the following. It is based on the book chapter by Burkhardt and Siggelkow [4] and the thesis by Schulz-Mirbach [42].

The idea of invariant features is to find a mapping T that is able to extract features which are the same for different observations $\boldsymbol{x}$ of the same equivalence class with respect

to a group action G. Such a transformation T maps all observations of an equivalence class into one point of the feature space:

$$\boldsymbol{x}_1 \overset{G}{\sim} \boldsymbol{x}_2 \Rightarrow T(\boldsymbol{x}_1) = T(\boldsymbol{x}_2). \tag{2}$$

In our case, it means that all frequency-warped versions of the same utterance should result in the same sequence of feature vectors. Given a transformation T and an observation $\boldsymbol{x}$, the set of all observations that are mapped into one point in the feature space is denoted as the *set of invariants* $\mathbf{I}_T(\boldsymbol{x})$ of an observation:

$$\mathbf{I}_T(\boldsymbol{x}) = \{\, \boldsymbol{x}_i \,|\, T(\boldsymbol{x}_i) = T(\boldsymbol{x}) \,\}. \tag{3}$$

The set of all possible observations within one equivalence class is called *orbit* $\mathbf{O}(\boldsymbol{x})$: Given a prototype $\boldsymbol{x}$, all other equivalent observations can be generated by applying the group action G,

$$\mathbf{O}(\boldsymbol{x}) := \{\, \boldsymbol{x}_i \,|\, \boldsymbol{x}_i \overset{G}{\sim} \boldsymbol{x} \,\}. \tag{4}$$

A transformation T is said to be *complete* if both the set of invariants of an observation and the orbit of the same observation are equal. Complete transformations have no ambiguities regarding the class discrimination. Incomplete transformations, on the other hand, have the property

$$\mathbf{O}(\boldsymbol{x}) \subseteq \mathbf{I}_T(\boldsymbol{x})$$

and may lead to the same features from observations of different equivalence classes and thus cannot distinguish them [4]. The described terms are visualized in Figure 1.

[Figure 1 to be inserted here]

In practice, a "high degree of completeness" is desired, which means that $\mathbf{I}_T(\boldsymbol{x})$ should not be much larger than $\mathbf{O}(\boldsymbol{x})$. Principles to systematically construct invariants with this property can be divided into three categories:

1. *Normalization.* Here, the observations are transformed with respect to extreme points on the orbit. Usually, a certain subset of the observations is chosen for the parameter estimation of the transformation.

2. *Differential approach.* This group of techniques obtains invariant features by solving partial differential equations (PDEs). Its main idea is that the features should be insensitive to infinitesimally small variations of the parameter(s) of the group action. Let these parameters be denoted by $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_N)$. For a group element $g \in G$ and a term $g(\boldsymbol{\beta})$ denoting a transformation $g \in G$ with parameters $\boldsymbol{\beta}$ and an input signal $\boldsymbol{x}$, it is demanded that

$$\frac{\partial T(g(\boldsymbol{\beta})\boldsymbol{x})}{\partial \beta_i} \equiv 0, \quad i = 1, 2, \ldots, N. \tag{5}$$

The solutions of the partial differential equations are the invariants.

3. *Integral approach.* The idea of the third group of methods is to compute averages of arbitrary functions on the entire orbit. Hurwitz invented the principle of integrating over the transformation group for constructing invariant features in 1897 [16]. The resulting integral is independent of the parameter(s) of the group action G:

$$T_f(\boldsymbol{x}) = \frac{1}{|G|} \int_G f(g\boldsymbol{x}) dg, \tag{6}$$

where $|G| := \int_G dg$ is the "power"of the group, and f is a (possibly complex-valued) kernel function.

One drawback of the normalization approach is the problem of a proper choice of a subset of the parameters. Another disadvantage is that a reduction of dimensionality is not given with this method. For the differential approach, solving the PDEs is the main difficulty in practice. In contrast, the integral approach does not need any preprocessing and, as will be shown in the following, is easily applied in the context of this work. Successful applications of this approach in the field of image processing were described, for example, by Schulz-Mirbach [43] and Siggelkow [45].

The three described principles represent general approaches for the construction of invariants for arbitrary transformation groups with a high degree of completeness. Certainly, there exist other methods that are invariant to specific transformations. However, the set of invariants $\mathbf{I}_T(\boldsymbol{x})$ of these methods is generally much larger than the orbit $\mathbf{O}(\boldsymbol{x})$. This means that the generated features are invariant to more operations than desired. Examples of translation-invariant transformations are the magnitude of the Fourier transformation as well as the auto- and cross correlation functions. Both the Fourier-transform

magnitude and the autocorrelation of a signal remain unchanged when applied to translated versions of the same signal. For the cross-correlation of two signals, both signals may be translated by the same arbitrary amount without affecting the result. Known invariant transformations from the field of image analysis are the generalized cyclic transformations [25] and the transformations of the class $\mathbb{C}T$ [3], which includes the rapid transformation [38]. Although, theoretically, these transformations have a larger set of invariants compared to the methods of the described three groups, all of them proved to be valuable in different kinds of applications [9, 25].

3. Feature computation and selection

This section introduces in its first part the contextual IIFs and their application to ASR. As it turns out, the proposed features have a high degree of freedom for the choice of parameters. A feature-selection method is described in the second part of this section.

3.1. Invariant-integration features

The proposed features rely on a TF analysis that approximately maps the spectral scaling due to different VTLs to translations along the subband-index space. With respect to the notions introduced in Section 2, this translation effect can be attributed to the action of the group G of translations. With this assumption, the TF representations S of two speakers A and B of the same utterance are related by a translation parameter α_T,

$$S_A(\zeta(\omega)) = S_B(\alpha_T + \zeta(\omega)), \tag{7}$$

where $\zeta(\omega) = \log(\omega)$ or some other function like the mel or the ERB scale. Formally, let $v_k(n)$ denote the TF representation of a speech signal, where n is the time index, $1 \leq n \leq N$, and k is the subband index with $1 \leq k \leq K$. A *frame* for time index n is then given by $\boldsymbol{v}_n = (v_1(n), v_2(n), \ldots, v_K(n))^\top$. When considering a translation according to α_T in the subband-index space, some boundary conditions need to be introduced. Periodic boundary conditions, where all subband indices are understood modulo K, have been used in [31, 32], because they were required by the applied invariance transformations. In this paper, we use repeated boundary conditions $v_k(n) = v_1(n)$ for $k < 1$ and $v_k(n) = v_K(n)$ for $k > K$, because they form a closer match to frequency warping in the analog domain.

By assuming that a finite set of translations along the subband-index space describes the occurring spectral effects due to different VTLs sufficiently[1], a finite group of translations $\widehat{G}$ with $|\widehat{G}|$ elements can be defined. For reasons of simplicity, integer translations are used in the rest of this article. Nevertheless, non-integer translations could be used if an appropriate interpolation scheme were incorporated in the following definitions. According to the "integration approach" as described in Section 2, Eq. (6) becomes

$$\widehat{T}_f(\boldsymbol{x}) = \frac{1}{|\widehat{G}|} \sum_{g \in \widehat{G}} f(g\boldsymbol{x}). \tag{8}$$

The question of how to define the function f arises. According to Noether's theorem [33, 41, 43], a complete transformation

$$\widehat{T}(\boldsymbol{x}) = \left(\widehat{T}_{f_1}(\boldsymbol{x}), \widehat{T}_{f_2}(\boldsymbol{x}), \ldots, \widehat{T}_{f_F}(\boldsymbol{x})\right)^\top \tag{9}$$

can be constructed by only considering monomials for the kernel functions f. Given the vectors $\boldsymbol{k} = (k_1, k_2, \ldots, k_M)$ and $\boldsymbol{l} = (l_1, l_2, \ldots, l_M)$, containing element indices and integer exponents with $\boldsymbol{k} \in \mathbb{N}^M$ and $\boldsymbol{l} \in \mathbb{N}_0^M$, respectively, a non-contextual monomial $m(n; w, \boldsymbol{k}, \boldsymbol{l})$ with M components is defined in the following as

$$m(n; w, \boldsymbol{k}, \boldsymbol{l}) := \left[\prod_{i=1}^{M} v_{k_i+w}^{l_i}(n)\right]^{1/\gamma(m)}, \tag{10}$$

where $w \in \mathbb{N}_0$ is a spectral offset parameter that is used for ease of notation in the following definitions. The value $\gamma(m)$ denotes the *order of a monomial* m:

$$\gamma(m) := \sum_{i=1}^{M} l_i. \tag{11}$$

The all-encompassing exponent $1/\gamma(m)$ in Eq. (10) acts as a normalizing term with respect to the order of the monomial. Noether showed that for input signals of dimensionality K and finite groups with $|\widehat{G}|$ elements, the group averages of monomials with order less or equal $|\widehat{G}|$ form a generating system of the input space. Such a basis has at most $\binom{|\widehat{G}|+K}{K}$ elements. It has to be pointed out that this is an upper bound, and it was shown in

[1]This is similar to the assumptions made by a typical grid-search based VTLN method, where a warping factor is chosen out of a finite set of possible warping factors.

many applications that the number of practically needed basis functions is considerably smaller [e.g., 41, 43, 45]. We can assume that the maximal translation that occurs as effect of VTL changes in the subband-index space is limited to a certain range W [46]. A non-contextual *invariant-integration feature* (IIF) $A_m(n)$ for a frame at time n, as a group average on the basis of a monomial m, is defined as

$$A_m(n) := \frac{1}{2W+1} \sum_{w=-W}^{W} m(n; w, \boldsymbol{k}, \boldsymbol{l}), \tag{12}$$

with $W \in \mathbb{N}_0$ determining the window size. To give an explanatory example, we consider a monomial of order 3 with exponents $l_i = 0$ for all $i \in \{1, 2, \ldots, K\} \setminus \{k_1, k_2, k_3\}$ and $l_i = 1$ for $i \in \{k_1, k_2, k_3\}$. The corresponding IIF $A_m(n)$ with window parameter $W = 1$ is then given by

$$A_m(n) = \frac{1}{3} \left[v_{k_1-1}(n) v_{k_2-1}(n) v_{k_3-1}(n) + v_{k_1}(n) v_{k_2}(n) v_{k_3}(n) + v_{k_1+1}(n) v_{k_2+1}(n) v_{k_3+1}(n) \right]. \tag{13}$$

Figure 2(a) shows a schematic plot of the computation of an IIF as defined in Eq. (12).

[Figure 2 to be inserted here]

A monomial for frame n following the definition in Eq. (10) is only evaluated on the components of frame n and temporal context is not considered. Since the spectral translation between different speakers is assumed to be time independent (c.f. Eq. (7)) the definition of a monomial in Eq. (10) can be extended such that it also considers neighboring frames for its computation. The resulting feature type with contextual monomials is called *contextual IIF* here. Given a vector $\boldsymbol{m} \in \mathbb{N}_0^M$ containing temporal offsets, a *contextual monomial* $\hat{m}$ with M components is defined as

$$\hat{m}(n; w, \boldsymbol{k}, \boldsymbol{l}, \boldsymbol{m}) := \left[\prod_{i=1}^{M} v_{k_i+w}^{l_i}(n + m_i) \right]^{1/\gamma(\hat{m})}. \tag{14}$$

Consequently, a contextual IIF $A_{\hat{m}}(n)$ is then given by replacing m by $\hat{m}$ in Eq. (12):

$$A_{\hat{m}}(n) := \frac{1}{2W+1} \sum_{w=-W}^{W} \hat{m}(n; w, \boldsymbol{k}, \boldsymbol{l}, \boldsymbol{m}). \tag{15}$$

With this definition, the non-contextual IIFs are a special case of the contextual ones. A schematic plot for the computation of the contextual IIFs according to Eq. (15) is shown in Figure 2(b). Following Noether's theorem as described above, an adequately chosen IIF set

$$\mathbf{A} := \{A_{\hat{m}_1}, A_{\hat{m}_2}, \ldots, A_{\hat{m}_F}\} \tag{16}$$

yields features that, on the one hand, are invariant to the translational spectral effects due to different VTLs, and, on the other hand, allow for discriminating between the individual classes.

The contextual IIFs can be applied on any TF representation that fulfills the assumption that spectral scaling is mapped to translation. This is approximately the case when a mel or an ERB scale is used for locating the frequency centers within the TF analysis [28, 48]. Depending on the application, a mean normalization can be applied to the features of each utterance to reduce the effect of channel sensitivity.

3.2. Feature selection for invariant-integration features

The set of parameters of the IIFs, consisting of the window size, element indices, exponents, and temporal offsets causes a huge number of possible combinations. Generally, with F features, T possible temporal offsets, B possible window sizes, K subbands, and a maximum order of D, the total count C of possible IIF sets is given by

$$C = \binom{B \cdot \sum_{d=1}^{D} (K \cdot T)^d}{F}. \tag{17}$$

Depending on the choice of parameter constraints, a count of more than 10^{28} different feature sets is possible in practice. For finding a good subset of IIFs, an appropriate feature selection has to be done. As stated above, Noether's theorem gives an upper bound for the order of the monomials that is needed to construct a basis for the observation space. Hence, one constraint for the feature selection that can be defined is an assumption about the maximum number of group elements, i.e., the maximum number of different translations that can be observed in our setting. Although the maximum reasonable monomial order is constrained by this theorem, the experiments of this work show that already an order of up to three leads to good results.

Because of the high degree of freedom and the large amount of data, a requirement for the feature-selection method is a low computational complexity. We used a method for feature selection based on the so-called *feature finding neural network* (FFNN) [12]. This method was applied successfully in the field of speech recognition [e.g., 12, 19, 20] especially in case of small sets of training data. The FFNN approach works iteratively with a single-layer perceptron at its basis. A fast training of this linear classifier is guaranteed by its closed-form solution. It should be emphasized that the linear classifier does not need to form complex decision boundaries, but to generalize well [12]. The feature-selection method can be summarized in the following four steps:

1. Start with a set of $F + 1$ features whose parameters are randomly chosen.
2. Use the linear classifier for computing the relevance of each feature.
3. Remove the feature with the least relevance.
4. If a stopping criterion (e.g., the total number of iterations) is not fulfilled, add a new, randomly generated feature to the current feature set and go back to the second step, otherwise stop.

During the feature selection, the parameter set that leads to the highest mean relevance is memorized and returned at the end of the selection process. The linear classifier performs a frame-wise phone classification. The relevance of a feature i is computed on basis of the computation of the *root mean square* (RMS) error of the linear classifier. The relevance of feature i is defined as the difference between the RMS error when using all features and the RMS error without feature i. It follows that this approach yields a ranking of the features according to their relevance. For determining the relevance, again, there are different setups possible. For example, it can be determined by using a matching (with respect to the mean VTL) training-test scenario. Another possibility would be to compute the relevance for each feature as the mean relevance for different mismatching training-test scenarios in which, for example, male utterances are used for training and female ones are used for testing, and vice versa. The experimental part considers both ways of relevance computation.

4. Experiments

A series of experiments has been conducted with the contextual IIFs. In the following, we compare the results with those for MFCCs and GTCCs. First, we consider IIFs with monomials of order one. Then we investigate whether the feature selection method used to find IIFs can also be applied to select MFCCs from a time window. Experiments with IIFs of higher order are considered afterwards. In a second part of the experiments, we include speaker adaptation based on VTLN and MLLR. The generality of the selected feature sets is investigated in the last part of the experiments.

4.1. Data and setup

The TIMIT corpus [11] with a sampling rate of 16 kHz was used for finding the feature sets and for the first part of the recognition experiments, in which we look at phone recognition. Overall, it consists of 6300 utterances from 630 speakers (438 male and 192 female). For training, the NIST training set [14] was used. It consists of 462 speakers, excludes the dialect (SA) sentences, and contains 3696 utterances. The complete test set [14] contains 1344 utterances from 168 speakers (SA sentences excluded) with no overlap between training and test sets. The last part of the experiment used the TIDIGITS corpus [24] downsampled to a sampling rate of 16 kHz. Overall, this corpus consists of about 25.000 read digit sequences produced by adult and children speakers. The corpus provides predefined training and test sets, which were used in this work.

Similar to the feature-selection process described in Section 3.2, different training-test scenarios were defined for both corpora in order to simulate matching and mismatching training-test conditions. The matching scenario in TIMIT refers to the standard training and test sets. Two mismatching scenarios were defined for TIMIT. The first used only the female utterances from the training set for training and the male utterances from the test set for testing. Similarly, the second setup used only the male utterances from the training and the female utterances from the test set. Accordingly, the mismatching scenarios are denoted as "F-M" or "M-F" in the following. In the M-F setting, the amount of test data is reduced to one third of the complete test set, so that the obtained accuracies are less statistically significant than for the complete test set used in the matching scenario.

However, the number of utterances is still more than twice of the core test set [14], so that statistical significance of M-F results is not a serious issue that could lead to a misinterpretation of the properties of different feature sets.

For the TIDIGITS corpus, two scenarios were defined. The first used only the corresponding utterances by adults for training and test. The second scenario used the utterances by adults for training and those by children for testing. These two scenarios are denoted in the following as "A-A" and "A-C", respectively.

While this work explicitly looks at the VTL as source of variability, other speaker dependencies also affect the results. For example, the corpora contain different dialects and speaking rates, so that good results on these corpora also indicate robustness to such variabilities. Of course, in general, one would want robustness to even more types of variation, which cannot be studied with these corpora. Experiments with regard to other variations are therefore planned for the future.

The TF analysis methods differed for the considered feature types: In case of the MFCCs, the standard HTK setup was used that consists of a 26-channel mel filterbank [6, 50]. A gammatone filterbank implemented with an FFT approach [8] was used in case of the GTCCs and IIFs. The minimum center frequency of the filters was set to 40 Hz, and the maximum center frequency was set to 8 kHz. With these constraints, the center frequencies were evenly spaced on the ERB scale. In case of GTCCs, 26 channels were used. In order to have sufficiently many spectral values for the computation of the IIFs, the number of channels was chosen as 110 in this case. Certainly, the number of channels could be smaller if the parameter for the window size would be non-integer and if an appropriate interpolation scheme would be used. However, minimizing the number of channels is not the scope of this article. The frame length was set to 20 ms and a frame shift of 10 ms was chosen. A power-law compression of the spectral values with an exponent of 0.1 was applied in order to resemble the nonlinear compression found in the human auditory system.

The recognizer was based on the *Hidden-Markov model toolkit* (HTK) [50] in all experiments. Phone recognition experiments were conducted on the TIMIT corpus. State-clustered, cross-word triphone models with diagonal covariance modeling were used. All

phone models had three emitting states with a left-to-right topology. Additionally, a long silence and a short silence model were included. A bigram language model was used that was computed on base of the TIMIT training data. According to [21] the phone-recognition results were folded to yield 39 final classes. The number of Gaussian mixtures was optimized for both feature types. While for the cepstral coefficient based feature types a mixture of 16 Gaussian distributions per state was chosen, it turned out to be beneficial to use mixtures of eight Gaussian distributions when using IIFs as features. Word recognition experiments were conducted with TIDIGITS. Here, whole-word left-to-right HMMs without skips over the states were trained. The number of states was chosen according to the average length of the individual words and varied between 9 and 15 states. A mixture of up to eight Gaussian distributions was used for all states.

For baseline accuracies, MFCCs and GTCCs with 12 coefficients plus log-energy together with first and second order derivatives were computed. Cepstral mean subtraction and variance normalization were performed. As a standard VTLN method, the approach described by Welling et al. [49] was used. It estimates maximum-likelihood warping factors based on whole utterances with a grid-search approach. The considered warping factors α_S were chosen from the set $\{0.88, 0.9, \ldots, 1.12\}$ in case of MFCCs. For the GTCCs, the warping parameters α_T were chosen empirically from $\{-1.5, -1.25, \ldots, 1.5\}$. During training, individual warping factors were first estimated for all speakers. Then these warping factors were used to train speaker-independent models. The decoding of the test data used a two-pass strategy: First, a hypothesis was computed based on the unwarped features. Then, the hypothesis was aligned to a set of warped features, and the warping factor with the highest confidence was used for the final decoding. Speaker-adaptive training and speaker adaptation with MLLR for the tests were also considered during the experiments. When MLLR was used, a regression class tree with eight terminal nodes was employed.

4.2. Feature selection

For the feature selection, the parameters were constrained as follows: The number of features F and the maximum order of the used monomials were varied within the individual experiments. The temporal offsets $\boldsymbol{m}$ were allowed to be within an interval

of ± 3 frames, which corresponds to a maximum temporal context of 80 ms. The maximum window-size parameter W was limited to 80 subbands. This number has been chosen so high to be on the safe side, and in fact, the features that were selected had values for W of up to 65. According to Eq. (17) the total count of possible feature sets for $F = 30$ and a maximum order of 1 is of magnitude 10^{110}. Because of technical limitations, only every tenth utterance was used for the feature selection. Therefore, about 370 male and female utterances with about 110.000 frames were considered. Of course, a larger amount of data for the feature selection could lead to better generalization capabilities, but the found feature sets already show good robustness, even if an entirely different corpus is used for testing. The last part of the experiments described here investigates the generalization capabilities of the TIMIT-based feature sets by using these features sets for word-recognition experiments on the TIDIGITS corpus.

The result of a feature-selection process according to the described method depends on its initialization and the randomly chosen features during its runtime. Thus, for each experiment, several repetitions of the feature-selection process were made. A number of ten repetitions has been experimentally determined, and an increase beyond ten did not significantly improve the results. No further heuristics for the initialization of the parameters have been used. The overall process can be described as follows:

1. Compute the TF representation of the training data as described in Section 4.1.
2. Perform feature selection as described in Section 3.2 ten times.
3. Decide for the feature set with the highest mean relevance.

With this choice of parameters, up to 15.000 different feature sets can be examined during the feature selection. The evolution of the smallest RMS error and the corresponding phone error rates (PER) are shown for an exemplary feature selection in Figure 3.

[Figure 3 to be inserted here]

For this example, it can be observed that the mean RMS error correlates well with the PER for the first 400 iterations. During the following 1000 iterations, the PER increases and decreases while the best RMS error is decreasing. This behavior in the vicinity of the optimum could be expected, as the results obtained with a linear classifier can predict

the performance of an HMM-based recognizer only to a limited extent. Generally, it was observed that after 1500 iterations of the described feature selection, the IIF set with the smallest mean RMS error yielded a significant improvement in accuracy compared to the randomly initialized IIF set. The IIFs that were obtained after the feature selection usually have small as well as large integration windows and involve the whole range of allowed parameter ranges. Considering the relevances of the individual features as described in Section 3.2, we observed that IIFs with large as well as with small integration windows were considered as highly relevant by the feature selection. Two exemplary contextual IIFs $A_{m_1}(n)$ and $A_{m_2}(n)$ that were selected during the experiments are shown in the following. They use monomials of order one, which corresponds to computing the mean spectral value for a certain frequency range,

$$A_{m_1}(n) = \frac{1}{21} \sum_{w=-10}^{10} v_{22+w}(n+1), \qquad A_{m_2}(n) = \frac{1}{47} \sum_{w=-23}^{23} v_{60+w}(n-1). \tag{18}$$

Here, the integration range for $A_{m_1}(n)$ is from about 150 Hz to about 480 Hz and for $A_{m_2}(n)$ from about 600 Hz to about 3200 Hz.

4.3. Invariant integration features of order one

Three IIF feature sets of order one and with sizes 10, 20, and 30 were selected in a matching scenario as described above. For the decoding, the log-energy was appended together with first and second order derivatives. Then, a *linear discriminant analysis* (LDA) was used to project the feature vectors down to 55 dimensions. Here, the phone segments were considered as individual classes [13]. The dimensionality-reduction step was omitted for the feature set that consists of 10 IIFs. Finally, a *maximum likelihood linear transformation* (MLLT) [39] was applied to allow for diagonal covariance modeling. In the following, the resulting accuracies will be compared with those for MFCCs and GTCCs with and without VTLN in a phone recognition task on TIMIT. Further results on cepstral coefficient based feature types with speaker adaptation will be reported in Section 4.6. Experiments with contextually enhanced MFCCs are described in the next section. To study the robustness to VTL mismatches, the usual matching case as well as the two mismatching scenarios M-F and F-M were considered for the comparison. Table 1 summarizes the results. The first two rows list the results for the MFCCs. By comparing

Table 1: Accuracies [%] of phone recognition experiments on TIMIT for MFCCs, GTCCs, and for IIFs of order one (with $F = 10, 20, 30$). The feature vector dimensions are shown in brackets.

Scenario / Features	match	mismatch	
	FM-FM	F-M	M-F
MFCC (39)	72.2	53.7	54.7
MFCC+VTLN (39)	73.3	67.7	70.0
GTCC (39)	72.5	55.4	54.0
GTCC+VTLN (39)	73.9	66.8	68.7
10 IIF (33)	73.4	61.6	62.9
20 IIF (55)	74.8	60.3	61.4
30 IIF (55)	75.3	60.2	61.1

the results for the MFCC matching scenario with those for the corresponding mismatching ones, it can be seen that the performance of standard MFCCs differs by about 17 percentage points. The enhancement when using MFCCs with VTLN is larger for the mismatching scenarios than for the corresponding matching scenario. The next two rows show the same behavior for GTCCs.

Analyzing the results for the IIFs, it can be seen that all three feature sets outperform the MFCCs and GTCCs without VTLN in all scenarios. Furthermore, for the matching case an increase of accuracy is observable with an increasing number of IIFs. Interestingly, the accuracies of the IIFs for the mismatching scenarios is highest for the smallest feature set consisting of 10 IIFs and is lowest with the largest feature set. An explanation may be that the larger IIF sets are better adapted to the corpus they were selected on and, therefore, do not generalize as well as the smaller IIF set. Interestingly, all IIF feature sets yield accuracies that are comparable (10 IIFs) or even higher (20 IIFs, 30 IIFs) than the ones of the cepstral coefficient based feature types with VTLN in the matching scenario, which represents the normal mode of operation of ASR systems.

4.4. Contextually selected MFCCs

Besides approximations of first- and second order time derivatives of the feature components, another common approach for considering temporal context information is to

concatenate feature vectors followed by an LDA. We also carried out LDA-based combinations of MFCC and GTCC feature vectors for an 80 ms context. Because of their similar results, only the ones for the MFCCs are shown in the upper part of Table 2. The comparison with Table 1 shows that the accuracies for the LDA extension are higher when no speaker-adaptation is used. However, when MLLR or VTLN+MLLR are applied, the features with the time-derivative extensions yield in most cases slightly higher accuracies with our setup. The properties of the LDA-based approach have, for example, been discussed by Schlüter et al. in [40], where it was pointed out that the performance of the LDA method often drops when features are highly correlated, too many frames are concatenated, or the training set is too small. To further investigate contextually enhanced MFCCs, we will describe another approach for including contextual information now which leads to significant improvements under matching as well as under mismatching conditions.

Contextual IIFs of order one are sums of weighted spectral values where the weighting coefficients have a rectangular shape. In case of the IIFs the rectangular shape originates from the idea of integrating over the group of all possible translations. This procedure is similar to the computation of MFCCs in which the spectral values at the output of a mel filterbank are weighted with triangular shaped coefficients and integrated. For a further comparison between IIFs and MFCCs we selected 30 MFCCs from an 80 ms context using the same feature selection algorithm as for the IIFs. Similar to the IIFs, log-energy and first- and second order time derivatives were appended to the feature vectors and finally reduced with an LDA to 55 dimensions. The lower part of Table 2 shows the results of these experiments. The recognition rates for selected GTCCs are not listed, because they were slightly inferior to selected MFCCs.

The comparison of these results with the accuracies of the MFCCs as listed in Table 1 shows that the contextually selected MFCCs yield much higher accuracies than the standard MFCCs for most of the setups. It is noteworthy that the accuracies of the contextual MFCCs increase especially under mismatching conditions. Our conclusions of these findings are that a good feature selection of individual components within a contextual temporal window may be beneficial compared to a (linear) combination of components as

Table 2: Accuracies [%] of phone recognition experiments on TIMIT for LDA-reduced concatenated MFCCs ($MFCC_{LDA}$), and for contextually selected MFCCs ($MFCC_{select}$). The feature vector dimensions are shown in brackets.

Scenario / Features	match	mismatch	
	FM-FM	F-M	M-F
$MFCC_{LDA}$ (55)	73.8	57.0	57.3
$MFCC_{LDA}$+MLLR (55)	74.2	60.4	61.2
$MFCC_{LDA}$+VTLN (55)	74.8	68.6	69.6
$MFCC_{LDA}$+VTLN+MLLR (55)	75.3	69.3	70.8
$MFCC_{select}$ (55)	74.7	60.3	61.5
$MFCC_{select}$+MLLR (55)	75.2	63.2	64.4
$MFCC_{select}$+VTLN (55)	76.1	70.9	71.2
$MFCC_{select}$+VTLN+MLLR (55)	76.4	71.2	72.1

it is done, e.g., by an LDA. The pure selection of MFCCs from a context window has to the best of our knowledge not been proposed before.

4.5. Invariant integration features of higher order and the feature-selection scenario

In the next experiment, the allowed order of the monomials was constrained to two and three, respectively. Furthermore, each feature selection was performed on the matching scenario, as well as on the mismatching scenarios M-F and F-M. The chosen size of the feature sets was set to 30. For comparison purposes, an IIF set of order one was also selected on the mismatching scenarios. Table 3 shows the results of the experiments.

It can be seen that the accuracies in the matching scenario degrade when IIFs of higher order are included, whereas the accuracies in the corresponding mismatching scenarios do increase. This effect is most noticeable when the IIFs are selected on the basis of mismatching scenarios. For example, the IIF set whose results are shown in the last row of Table 3 yields similar accuracies as the MFCCs when combined with VTLN. Thus, the accuracy for the normal matching cases can be traded off to increase the robustness to larger mismatches between training and testing.

Table 3: Accuracies [%] of phone recognition experiments on TIMIT for 30 IIFs of higher order. "FS scenario" denotes the scenario which was used for the relevance computation during the feature selection (matching or mismatching), "order" denotes the maximum monomial order of the IIFs.

FS scenario	Order	match	mismatch	
		FM-FM	F-M	M-F
matching	1	75.3	60.2	61.1
	≤ 2	74.4	62.4	61.4
	≤ 3	74.2	62.2	61.7
mismatching	1	74.2	62.3	62.2
	≤ 2	73.3	63.3	64.9
	≤ 3	73.5	64.3	64.1

4.6. Invariant integration features combined with VTLN and/or MLLR

Adaptation and normalization methods are commonly part of state-of-the art ASR systems nowadays. In the following, we investigate whether the superior properties of the IIFs are still observable when VTLN and/or MLLR are also used within the ASR system. A block-diagonal structure with three blocks was set as constraint for all MLLR-transform estimations, because it turned out to be beneficial for all considered feature types. While in the experiments in Sections 4.3 and 4.5 an LDA was applied on the IIFs together with their derivatives, the experiments in this part had a different sequence of feature-processing steps: For the IIF sets of size 30, an LDA with a final dimensionality of 20 was applied. No dimensionality reduction was performed with the IIF sets of size 10 and 20. Then, the log-energy and first and second order derivatives were appended, and a 3-block-constrained MLLT was computed to decorrelate the features. Speaker-adaptive training was performed with *constrained MLLR* (CMLLR), while a combination of CMLLR and MLLR was applied in the decoding stage. When VTLN was used with IIFs, which are using a 110-channel gammatone filterbank, the considered warping parameters α_T were $-8, -7, \ldots, 8$. In case of MFCCs and GTCCs, the warping parameters were chosen as described in Section 4.1. Table 4 shows the results of the experiments.

As expected, the cepstral coefficient-based systems that use both MLLR and VTLN

Table 4: Accuracies [%] of phone recognition experiments on TIMIT for MFCCs, GTCCs, and for IIFs of order one ($F = 10, 20, 30$) when adaptation methods are used. The feature vector dimensions are shown in brackets.

Features \ Scenario	match	mismatch	
	FM-FM	F-M	M-F
MFCC+MLLR (39)	75.2	65.3	66.9
MFCC+VTLN (39)	73.3	67.7	70.0
MFCC+VTLN+MLLR (39)	75.4	69.6	71.8
GTCC+MLLR (39)	74.9	66.2	66.3
GTCC+VTLN (39)	73.9	66.8	68.7
GTCC+VTLN+MLLR (39)	76.3	70.4	72.4
10 IIF+MLLR (33)	74.9	68.0	68.6
20 IIF+MLLR (63)	75.4	67.8	69.4
30 IIF+MLLR (63)	76.2	68.1	69.4
10 IIF+VTLN (33)	75.4	69.7	70.5
20 IIF+VTLN (63)	76.2	70.1	70.1
30 IIF+VTLN (63)	77.2	71.4	70.9
10 IIF+VTLN+MLLR (33)	76.0	71.2	72.5
20 IIF+VTLN+MLLR (63)	77.1	72.4	72.3
30 IIF+VTLN+MLLR (63)	77.4	73.4	72.4

yield higher accuracies in all scenarios than the cepstral coefficient-based systems that use only one of the adaptation methods or none of them. Generally, it can be observed that IIF-based ASR systems do benefit from the use of MLLR and/or VTLN. Compared to the accuracies from Table 1, it can be seen that the additional use of MLLR and VTLN increases the accuracy of the MFCC- and GTCC-based systems in the matching scenario by about 3 and 4 percentage points, respectively. Combining MLLR and VTLN with IIFs yields increases in accuracy for the matching scenario between 2.1 and 2.6 percentage points. Overall, the IIF set of size 30 in combination with MLLR and VTLN yields the highest accuracies in the experiments. Also, in comparison to the accuracies of the

Table 5: Accuracies [%] of word recognition experiments on TIDIGITS for MFCCs, GTCCs, and for IIFs of order one ($F = 10, 20, 30$). The feature vector dimensions are shown in brackets.

Features \ Scenario	match	mismatch
	A-A	A-C
MFCC (39)	99.52	96.02
MFCC+VTLN (39)	99.59	97.25
GTCC (39)	99.65	97.67
GTCC+VTLN (39)	99.66	98.94
10 IIF (33)	99.59	97.40
20 IIF (55)	99.64	97.95
30 IIF (55)	99.68	97.89
10 IIF+VTLN (33)	99.65	99.22
20 IIF+VTLN (55)	99.73	99.38
30 IIF+VTLN (55)	99.76	99.30

contextually selected MFCCs as listed in Table 2 the IIF set with 30 features yields the highest accuracies in most of the setups.

4.7. Experiments on TIDIGITS

Since the features were selected on base of an individual corpus (in our case TIMIT), the question arises, how good the IIF sets perform on another corpus. Therefore, the last part of the experiments considered the "generalization capabilities" of the features. The same IIF sets that were selected on the TIMIT corpus were used for the feature extraction for the word-recognition task on TIDIGITS. The results of theses experiments are shown in Table 5.

The first four lines of Table 5 show the results obtained by the cepstral coefficient-based ASR systems without adaptation and with VTLN, respectively. From line five on, the accuracies of the IIF-based systems without any adaptation and with VTLN are shown. For both scenarios, it can be seen that the IIF-based systems without VTLN yield accuracies that are equally high (10 IIF, A-A) or higher than for the MFCC-based systems. It is particularly remarkable that the IIF-based system without adaptation outperforms

the MFCC-based one with VTLN even in the A-C setting, where the recognizer is trained on adult speech and tested with children speech. Due to the very low number of errors in the A-A case, the statistical significance of the results may be in question for this scenario. However, in the A-C setting, where the accuracy is generally lower, it could be improved by up to 0.7 percentage points when using IIFs instead of MFCCs with VTLN. In absolute terms, this means that the number of correctly recognized digits could be increased by the IIFs by up to 90 digits compared to MFCCs, which can be seen as a significant increase of accuracy. The comparison of IIFs with GTCCs without VTLN shows comparable accuracies. However, when VTLN is used, the accuracy of the IIFs in the A-C scenario still is about 0.4 percentage points higher than the GTCC+VTLN case, which corresponds to 30 more correctly recognized digits.

5. Discussion and conclusions

We presented a feature-extraction method that is based on the theory of invariant integration. A major assumption is that the spectral effects caused by different VTLs is mapped to translations along the subband-index space of the TF representation. *Invariant-integration features* were defined such that they incorporate information about temporal context. Since the proposed IIFs are based on monomials, their computation has a very low complexity. However, the definition of these features has a high degree of parametric freedom, so that an appropriate feature-selection method is needed. Within this work, a method based on a linear classifier was used that iteratively enhances a feature set of a fixed size.

Experiments showed that contextually selected MFCCs yield much higher accuracies in matching as well as in mismatching scenarios than standard MFCCs and LDA-reduced concatenated MFCCs. Invariant-integration features can be seen as a further refinement that tries to find important spectral cues within a certain temporal context and enhances the robustness of the resulting features to VTL changes.

When no speaker-adaptation was used, the experiments showed a superior performance of the IIFs compared to cepstral coefficients (MFCCs and GTCCs) in matching, and especially in mismatching training-testing conditions. In the matching scenario, IIFs

of order one perform better than IIFs based on higher order monomials. However, using higher order monomials yields better performances on mismatching training-testing scenarios. The experiments showed that the combination of IIFs with MLLR and/or VTLN further increases their accuracy. Within the experiments, the IIF-based systems lead to the highest accuracies and outperformed the cepstral coefficient-based systems in matching training-test conditions for the TIMIT phone-recognition task by more than one percentage point. The last part of the experiments showed that the TIMIT-based IIF sets can equally well be used for word recognition on the TIDIGITS corpus. Here, the IIF-based systems also yield accuracies that are higher than those of the MFCC-based system without and with VTLN.

Future work will be focused on finding a general set of IIFs that works equally well on different corpora and under a larger class of variabilities. A comprehensive study about the robustness of the IIFs to different types of distortions will also be conducted. Finally, the application of more sophisticated TF-analysis methods can be combined with the presented feature type and might lead to further improvements.

Acknowledgements

This work has been supported by the German Research Foundation under Grant No. ME1170/2-1.

References

[1] Benzeghiba, M., Mori, R. D., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., Fissore, L., Laface, P., Mertins, A., Ris, C., Rose, R., Tyagi, V., Wellekens, C., Oct.-Nov. 2007. Automatic speech recognition and speech variability: a review. Speech Communication 49 (10-11), 763–786.

[2] Boë, L.-J., Granat, J., Badin, P., Autesserre, D., Pochic, D., Zga, N., Henrich, N., Ménard, L., Dec. 2006. Skull and vocal tract growth from newborn to adult. In: Proc. 7th Int. Seminar on Speech Production (ISSP7). Ubatuba, Brazil, pp. 75–82.

[3] Burkhardt, H., Müller, X., Oct. 1980. On invariant sets of a certain class of fast translation-invariant transforms. IEEE Trans. Acoustic, Speech, and Signal Processing 28 (5), 517–523.

[4] Burkhardt, H., Siggelkow, S., 2001. Invariant features in pattern recognition – fundamentals and applications. In: Nonlinear Model-Based Image/Video Processing and Analysis, C. Kotropoulos and I. Pitas (Eds.). John Wiley & Sons, pp. 269–307.

[5] Cohen, L., Dec. 1993. The scale representation. IEEE Trans. Signal Processing 41 (12), 3275–3292.

[6] Davis, S., Mermelstein, P., Aug. 1980. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans. Acoustics, Speech and Signal Processing 28 (4), 357–366.

[7] Deller, J. R., Proakis, J. G., Hansen, J. H. L., 1993. Discrete-time processing of speech signals. Macmillan, New York.

[8] Ellis, D. P. W., Jun. 2009. Gammatone-like spectrograms. web resource: http://www.ee.columbia.edu/~dpwe/resources/matlab/gammatonegram.

[9] Fang, M., Häusler, G., Mar. 1989. Modified rapid transform. Applied Optics 28 (6), 1257–1262.

[10] Gales, M. J. F., Apr. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language 12 (2), 75–98.

[11] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., 1993. DARPA TIMIT acoustic phonetic speech corpus. Linguistic Data Consortium, Philadelphia.

[12] Gramss, T., Oct. 1991. Word recognition with the feature finding neural network (FFNN). In: Proc. IEEE Workshop Neural Networks for Signal Processing. Princeton, NJ, USA, pp. 289–298.

[13] Haeb-Umbach, R., Ney, H., Mar. 1992. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing. Vol. 1. San Francisco, CA, USA, pp. 13–16.

[14] Halberstadt, A. K., Nov. 1998. Heterogeneous acoustic measurements and multiple classifiers for speech recognition. Ph.D. thesis, Massachusetts Institute of Technology.

[15] Huang, X., Acero, A., Hon, H., 2001. Spoken Language Processing: A Guide to Theory, Algorithm, and System Development. Prentice Hall PTR Upper Saddle River, New York, USA.

[16] Hurwitz, A., 1897. Ueber die Erzeugung der Invarianten durch Integration. Nachrichten von der Königl. Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse, 71–90.

[17] Irino, T., Patterson, R., Mar. 2002. Segregating information about the size and the shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform. Speech Communication 36 (3), 181–203.

[18] Ishizuka, K., Nakatani, T., Minami, Y., Miyazaki, N., 2006. Speech feature extraction method using subband-based periodicity and nonperiodicity decomposition. The Journal of the Acoustical Society of America 120 (1), 443–452.

[19] Kleinschmidt, M., Sep. 2002. Robust speech recognition based on spectro-temporal processing. Ph.D. thesis, Universität Oldenburg.

[20] Kleinschmidt, M., Gelbart, D., Sept. 2002. Improving word accuracy with Gabor feature extraction. In: Proc. Int. Conf. Spoken Language Processing. pp. 25–28.

[21] Lee, K. F., Hon, H. W., Nov. 1989. Speaker-independent phone recognition using hidden Markov models. IEEE Trans. Acoustics, Speech and Signal Processing 37 (11), 1641–1648.

[22] Lee, L., Rose, R. C., Jan. 1998. A frequency warping approach to speaker normalization. IEEE Trans. Speech and Audio Processing 6 (1), 49–60.

[23] Leggetter, C., Woodland, P., Apr. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer Speech and Language 9 (2), 171–185.

[24] Leonard, R. G., Doddington, G., 1993. TIDIGITS. Linguistic Data Consortium, Philadelphia.

[25] Lohweg, V., Diederichs, C., Müller, D., Jan. 2004. Algorithms for hardware-based pattern recognition. EURASIP J. Applied Signal Processing 2004, 1912–1920.

[26] Mertins, A., Rademacher, J., Nov. 27 -Dec. 1 2005. Vocal tract length invariant features for automatic speech recognition. In: Proc. 2005 IEEE Automatic Speech Recognition and Understanding Workshop. San Juan, Puerto Rico, pp. 308–312.

[27] Mertins, A., Rademacher, J., May 2006. Frequency-warping invariant features for automatic speech recognition. In: Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing. Vol. V. Toulouse, France, pp. 1025–1028.

[28] Monaghan, J. J., Feldbauer, C., Walters, T. C., Patterson, R. D., Jul. 2008. Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition. J. Acoustical Society of America 123 (5), 3066–3066.

[29] Moore, B. C. J., Glasberg, B. R., Mar. 1996. A revision of Zwicker's loudness model. Acta Acustica united with Acustica 82 (11), 335–245.

[30] Müller, F., Belilovsky, E., Mertins, A., Dec. 2009. Generalized cyclic transformations in speaker-independent speech recognition. In: Proc. IEEE Automatic Speech Recognition and Understanding Workshop. Merano, Italy, pp. 211–215.

[31] Müller, F., Mertins, A., Sept. 2009. Invariant-integration method for robust feature extraction in speaker-independent speech recognition. In: Proc. Int. Conf. Spoken Language Processing (Interspeech 2009-ICSLP). Brighton, UK, pp. 2975–2978.

[32] Müller, F., Mertins, A., Feb. 2010. Nonlinear translation-invariant transformations for speaker-independent speech recognition. In: Sole-Casals, J., Zaiats, V. (Eds.), Advances in Nonlinear Speech Processing. Vol. 5933 of LNAI. Springer, Heidelberg, Germany, pp. 111–119.

[33] Noether, E., Mar. 1915. Der Endlichkeitssatz der Invarianten endlicher Gruppen. Mathematische Annalen 77 (1), 89–92.

[34] Patterson, R. D., 2000. Auditory images: How complex sounds are represented in the auditory system. J. Acoustical Society of Japan (E) 21 (4), 183–190.

[35] Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D., Zhang, C., Allerhand, M., 1992. Complex sounds and auditory images. In: Cazals, Y., Demany, L., Horner, K. (Eds.), Auditory Physiology and Perception. Advanced Bioscience. Vol. 83. Pergamon, Oxford, pp. 429–446.

[36] Pitz, M., Ney, H., Sept. 2005. Vocal tract normalization equals linear transformation in cepstral space. IEEE Trans. Speech and Audio Processing 13 (5), 930–944.

[37] Rademacher, J., Wächter, M., Mertins, A., Sept. 2006. Improved warping-invariant features for automatic speech recognition. In: Proc. Int. Conf. Spoken Language Processing (Interspeech 2006 - ICSLP). Pittsburgh, PA, USA, pp. 1499–1502.

[38] Reitboeck, H., Brody, T. P., Aug. 1969. A transformation with invariance under cyclic permutation for applications in pattern recognition. Information and Control 15 (2), 130–154.

[39] Saon, G., Padmanabhan, M., Gopinath, R., Chen, S., Jun. 2000. Maximum likelihood discriminant feature spaces. In: Proc. Int. Conf. Audio Speech and Signal Processing. pp. 1129–1132.

[40] Schlüter, R., Zolnay, A., Ney, H., Sept. 2006. Feature combination using linear discriminant analysis and its pitfalls. In: Proc. Int. Conf. Spoken Language Processing (ICSLP/Interspeech). Pittsburgh, USA, pp. 345–348.

[41] Schulz-Mirbach, H., Aug. 1992. On the existence of complete invariant feature spaces in pattern recognition. In: Proc. Int. Conf. Pattern Recognition. Vol. 2. Hague, Netherlands, pp. 178–182.

[42] Schulz-Mirbach, H., 1995. Anwendung von Invarianzprinzipien zur Merkmalgewinnung in der Mustererkennung. Tr-402-95-018, Universitaet Hamburg, Hamburg, Germany.

[43] Schulz-Mirbach, H., 1995. Invariant features for gray scale images. In: Mustererkennung 1995, 17. DAGM-Symposium. Springer, London, UK, pp. 1–14.

[44] Sena, A. D., Rocchesso, D., Nov. 2005. A study on using the mellin transform for vowel recognition. In: Proc. Sound and Music Conf. Salerno, Italy.

[45] Siggelkow, S., 2002. Feature histograms for content-based image retrieval. Ph.D. thesis, Fakultät für Angewandte Wissenschaften, Albert-Ludwigs-Universität Freiburg, Breisgau, Germany.

[46] Sinha, R., Umesh, S., May 2002. Non-uniform scaling based speaker normalization. In: Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP'02). Vol. 1. Orlando, USA, pp. I–589 – I–592.

[47] Umesh, S., Cohen, L., Marinovic, N., Nelson, D. J., Jan. 1999. Scale transform in speech analysis. IEEE Trans. Speech and Audio Processing 7 (1), 40–45.

[48] Umesh, S., Kumar, S. V. B., Vinay, M. K., Sharma, R., Sinha, R., May 2002. A simple approach to non-uniform vowel normalization. In: Proc. int . Conf. Acoustics, Speech, and Signal Processing (ICASSP'02). Vol. 1. Orlando, USA, pp. I–517–I–520.

[49] Welling, L., Ney, H., Kanthak, S., Sept. 2002. Speaker adaptive modeling by vocal tract normalization. IEEE Trans. Speech and Audio Processing 10 (6), 415–426.

[50] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. A., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2009. The HTK Book (for HTK Version 3.4.1). Cambridge University Engineering Department, Cambridge, UK.

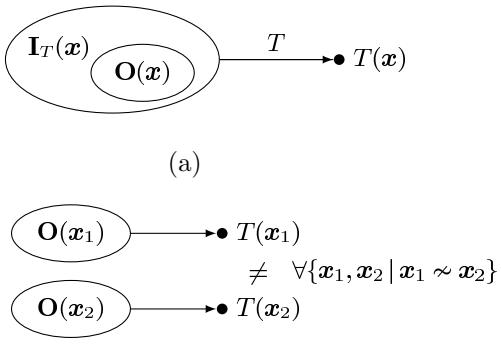

Figure 1: (a) Relation of invariant set $\mathbf{I}_T$ and orbit $\mathbf{O}$ for a given transformation T and observation $\boldsymbol{x}$, (b) notion of *completeness* of a transformation T. After Burkhardt and Siggelkow [4].

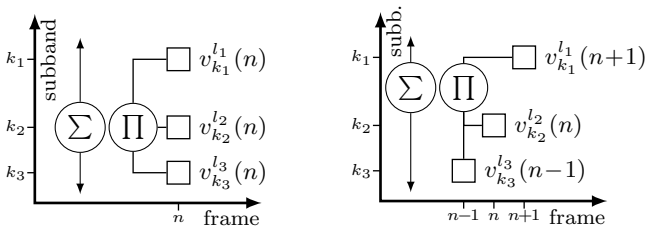

Figure 2: (a) Schematic plot of a non-contextual IIF with exponents $l_1, l_2, l_3 \neq 0$, (b) schematic plot of a contextual IIF with exponents $\boldsymbol{l} = (l_1, l_2, l_3)$, $l_1, l_2, l_3 \neq 0$ and corresponding temporal offsets $\boldsymbol{m} = (m_1, m_2, m_3) = (+1, 0, -1)$.

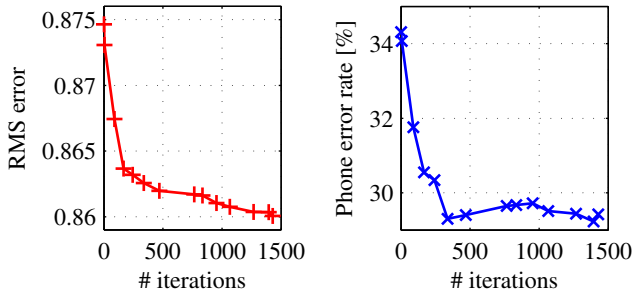

Figure 3: Exemplary feature selection: (a) evolution of the smallest mean RMS error, and (b) corresponding phone error rates.