Generalized Cyclic Transformations in Speaker-Independent Speech Recognition

Florian Müller [1], Eugene Belilovsky, and Alfred Mertins

Institute for Signal Processing, University of Lübeck
Ratzeburger Allee 160, 23538 Lübeck, Germany
[1] mueller@isip.uni-luebeck.de

***Abstract*—A feature extraction method is presented that is robust against vocal tract length changes. It uses the generalized cyclic transformations primarily used within the field of pattern recognition. In matching training and testing conditions the resulting accuracies are comparable to the ones of MFCCs. However, in mismatching training and testing conditions with respect to the mean vocal tract length the presented features significantly outperform the MFCCs.**

I. Introduction

The vocal tract length (VTL) is a source of variability that is especially relevant for speaker-independent automatic speech recognition (ASR) systems. Adult VTLs may differ by up to 25 percent [1]. A common approximation is that in the linear spectral domain, the short-time spectra of two speakers A and B, when uttering the same phone, are approximately related by $S_A(\omega) = S_B(\alpha \cdot \omega)$. Without further processing within the ASR system, this spectral scaling causes a degradation of recognition rate.

Various methods were proposed that try to counteract the scaling effect. On the one hand, there are methods that try to adapt the acoustic models to the features of each utterance [2], [3], also known as (constrained) MLLR techniques. On the other hand, there are methods that try to normalize the features to reduce the mismatch between training and testing conditions. The VTL normalization (VTLN) methods [1], [4] belong to this group. In [3] it was shown that VTLN can be seen as a constrained MLLR. The mentioned methods have in common that they need an additional adaptation step within the recognition process of the ASR system. Another class of methods tries to extract features that are invariant to the spectral effects of VTL changes [5], [6], [7]. Though not as mature as the adaptation and normalization techniques, speaker-invariant feature extraction methods with low computational costs could simplify ASR systems by omitting speaker-normalization or -adaptation stages.

The field of pattern recognition has acquired many methods that compute invariant features with respect to different groups of transformations. Various works showed that by using a filter bank with frequency centers located evenly spaced on an auditory motivated scale, like the mel or equivalent rectangular bandwidth (ERB) scale, the spectral scaling is approximately mapped to translation. The magnitude of the Fourier transformation is commonly known for its translation-invariance. Other well known nonlinear translation-invariant transformations belong to a group known as *$\mathbb{C}$T-transformations* [8], [9]. Based on this group, modifications were also presented [10]. It has been shown in previous works [7], [11] that the application of translation-invariant transformations as feature extraction method for ASR systems can yield features that are more robust in mismatching training-testing conditions (w.r.t. the mean VTL) than the standard mel frequency cepstral coefficients (MFCCs). In this work another class of transformation is investigated for its applicability in the field of speaker-independent speech recognition. The members of this class are generally known as *generalized cyclic transformations* (GCT) [12]. Instances of this class were successfully used in the field of pattern recognition [13].

The next section introduces the class of GCTs and shows how to use it for feature extraction in ASR systems. A series of phoneme recognition experiments has been conducted for this work. Section III describes the recognition system and the experiments that compare the presented feature extraction methods. It is shown that the translation-invariant feature types work best when they are combined with each other.

II. Generalized Cyclic Transformations & Translation-Invariant Features

A. Definition of the GCT

In this section the class of generalized cyclic transformations is described following the presentation in [12]. For this, let $\boldsymbol{x} \in \mathbb{R}^N$ be an input vector and let $\hat{\boldsymbol{x}} \in \mathbb{R}^N$ be the transformation of $\boldsymbol{x}$. These vectors are related by the transformation matrix $\boldsymbol{A}_N \in \mathbb{R}^{N \times N}$ as

$$\hat{\boldsymbol{x}} = \boldsymbol{A}_N \cdot \boldsymbol{x}. \tag{1}$$

The notion of *negacyclic matrices* [14], also known as skew circulants, is instrumental in the following sections. A negacyclic matrix $\boldsymbol{C} \in \mathbb{R}^{N \times N}$ is defined by a coefficient vector $\boldsymbol{c} := (c_0, c_1, \ldots, c_{N-1})$ as

$$\boldsymbol{C}[\boldsymbol{c}] := \begin{bmatrix} c_0 & c_1 & \cdots & \cdots & c_{N-1} \\ -c_{N-1} & c_0 & c_1 & \cdots & c_{N-2} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ -c_2 & -c_3 & -c_4 & \ddots & c_1 \\ -c_1 & -c_2 & -c_3 & \cdots & c_0 \end{bmatrix}. \tag{2}$$

978-1-4244-5479-2/09/$26.00 © 2009 IEEE

By introducing matrices $\boldsymbol{T}_{N/2}, \boldsymbol{T}_{N/4}, \ldots, 1$, the transformation matrix A_N of the GCT can be defined recursively as

$$A_N := \begin{bmatrix} T_{N/2} & -T_{N/2} \\ A_{N/2} & A_{N/2} \end{bmatrix}, \tag{3}$$

where $A_1 := 1$. The derivation of the following result can be found in [12]. Based on the observations in [15], the idea is to choose the matrices $\boldsymbol{T}_{N/2}, \boldsymbol{T}_{N/4}, \ldots, 1$ such that the *absolute value spectrum* (AVS) of $\hat{\boldsymbol{x}}$, which is introduced in the next subsection, stays unchanged under cyclic translation of $\boldsymbol{x}$. It can be shown that this is fulfilled by defining

$$\boldsymbol{T}^{\boldsymbol{b}_M} := -\boldsymbol{C}[\boldsymbol{b}_M], \tag{4}$$

where $\boldsymbol{b}_M = (b_0, b_1, \ldots, b_{M-1})$ is a coefficient vector. The matrix $\boldsymbol{T}$ is called a *generalized cyclic matrix* (GCT). A set of $\mathrm{ld}(N)$ coefficient vectors $\boldsymbol{b}^1_{N/2}, \boldsymbol{b}^2_{N/4}, \ldots, \boldsymbol{b}^{\mathrm{ld}(N)}_1$ defines a transformation matrix $\boldsymbol{A}_N$. Now, by unfolding the recursion in Equation (3), the transformation matrix $\boldsymbol{A}_N$ can be written as

$$\boldsymbol{A}_N := \begin{bmatrix} \boldsymbol{T}^{\boldsymbol{b}^1_{N/2}} & & & \boldsymbol{0} \\ & \boldsymbol{T}^{\boldsymbol{b}^2_{N/4}} & & \\ & & \ddots & \\ \boldsymbol{0} & & & 1 \end{bmatrix} \cdot \left(\prod_{i=1}^{\mathrm{ld}(N)-1} \mathrm{diag}\left(\boldsymbol{I}_{N-2^i}, \boldsymbol{H} \otimes \boldsymbol{I}_{2^{i-1}}\right) \right) \cdot \left(\boldsymbol{H} \otimes \boldsymbol{I}_{N/2}\right), \tag{5}$$

where $\mathrm{diag}(\cdot, \cdot)$ defines a diagonal block matrix with two submatrices. $\boldsymbol{I}_N$ is the identity matrix of size N, and $\boldsymbol{H}$ is a Hadamard matrix of order 2:

$$\boldsymbol{H} := \begin{bmatrix} +1 & -1 \\ +1 & +1 \end{bmatrix}. \tag{6}$$

The last two bracket terms in Equation (5) can be interpreted as a rationalized form of the *modified Walsh Hadamard transformation* (MWHT) [16]. The matrix containing the GCMs specifies the properties of the resulting transformation and is defined by the characteristic coefficient vector

$$\tilde{\boldsymbol{b}} := (\boldsymbol{b}^1_{N/2}, \boldsymbol{b}^2_{N/4}, \ldots, \boldsymbol{b}^{\mathrm{ld}(N)}_1). \tag{7}$$

The rows of the transformation matrix $\boldsymbol{A}_N$ have a period-wise order. It is noteworthy that the rationalized MWHT and the *square wave transformation* (SWT) [17] can be realized with the GCT by choosing appropriate characteristic coefficients. Besides the coefficients for the MWHT and the SWT, three additional generating rules for the coefficients were listed in [13]. Based on these, we used the coefficient types as listed in Table I.

B. Translation-Invariant Features

In order to obtain translation-invariant features, a nonlinear function has to be applied on $\hat{\boldsymbol{x}}$. Based on the transformed signal $\hat{\boldsymbol{x}}$, two types of translation-invariant features are described in this subsection. While the first method is based on the aforementioned AVS, the second method uses the cyclic autocorrelation function for defining a translation-invariant extended group spectrum.

Similar to the computation of the power spectrum of the Walsh-Hadamard transformation [16], an AVS for a GCT transformed signal can be defined. It is shown in [13] that the period-wise addition of the absolute values of the transformed signal $\hat{\boldsymbol{x}}$ yields a translation-invariant spectrum. Formally, let

$$F_i := N\left(1 - 0.5^i\right), \quad i \in \mathbb{N}_0, \tag{8}$$

be a supplementary function that is used for the ease of notation in the following. Then, $\mathrm{AVS}(\hat{\boldsymbol{x}}) : \mathbb{R}^N \rightarrow \mathbb{R}^{\mathrm{ld}(N)+1}$ with $\mathrm{AVS}(\hat{\boldsymbol{x}}) = (s_0, s_1, \ldots, s_{\mathrm{ld}(N)})$ is defined as

$$s_i := \begin{cases} \sum_{k=F_i}^{F_{i+1}-1} |\hat{x}_k|, & i = 0, \ldots, \mathrm{ld}(N) - 1, \\ |\hat{x}_N|, & i = \mathrm{ld}(N). \end{cases} \tag{9}$$

A second translation-invariant feature type for the GCT can be defined on base of a cyclic autocorrelation function. Based on [18], the cyclic autocorrelation for the GCT is defined as

$$\boldsymbol{R}^{\boldsymbol{U}}_{\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}} := \begin{bmatrix} \boldsymbol{U}_0 & & & \boldsymbol{0} \\ & \boldsymbol{U}_1 & & \\ & & \ddots & \\ \boldsymbol{0} & & & 1 \end{bmatrix} \cdot \hat{\boldsymbol{x}}, \tag{10}$$

where

$$\boldsymbol{U}_i := \boldsymbol{C}\left[\left(\hat{x}_{F_i}, \hat{x}_{F_i+1}, \ldots, \hat{x}_{F_{i+1}-1}\right)\right] \tag{11}$$

for $i = 0, 1, \ldots, \mathrm{ld}(N) - 1$. According to [19], the autocorrelation of a GCT transformed signal is highly redundant. However, by introducing the signum function to the equation given in (10), a translation-invariant spectrum, denoted as *extended group spectrum* (EGS), can be computed. The $\mathrm{EGS} : \mathbb{R}^N \rightarrow \mathbb{R}^N$ of a transformed signal $\hat{\boldsymbol{x}}$ is defined as

$$\mathrm{EGS}(\hat{\boldsymbol{x}}) := \boldsymbol{R}^{\boldsymbol{V}}_{\hat{\boldsymbol{x}}\hat{\boldsymbol{x}}}, \tag{12}$$

where

$$\boldsymbol{V}_i := \boldsymbol{C}\left[\mathrm{sgn}\left(\hat{x}_{F_i}, \hat{x}_{F_i+1}, \ldots, \hat{x}_{F_{i+1}-1}\right)\right]. \tag{13}$$

The number of separable patterns of the EGS is larger than the one of the AVS method. The drawback of the EGS function is the higher dimension of its image compared to the AVS. It is noteworthy that the AVS is contained within the EGS [19].

C. Application to ASR

For the application of the described feature types within an ASR system, a time-frequency (TF) representation of an input speech signal has to be computed. Commonly, a filter bank is used for the TF analysis. Auditory motivated scales like the mel or the ERB scale are typically used to locate the frequency centers of the filters evenly spaced on these scales. It was shown that these kinds of filter banks map the spectral changes induced by VTL changes to translations along the subband index space of the TF representation [5], [20]. Let $\boldsymbol{y}(n, k)$ denote the TF representation of an input speech signal, where n is the frame index, $1 \leq n \leq N$, and k is the

TABLE I
CHARACTERISTIC COEFFICIENT VECTOR FOR DIFFERENT TYPES OF TRANSFORMATIONS

Transformation	Coefficient vector $\tilde{b}$
SWT	$(\underbrace{1, 1, \ldots, 1}_{N-3 \text{ coeffs.}}, -1, -1, 1)$
MWHT	$(\underbrace{0, 0, \ldots, 0, -1}_{N/2 \text{ coeffs.}}, \underbrace{0, 0, \ldots, 0, -1}_{N/2^2 \text{ coeffs.}}, \ldots, -1)$
C$_1$	$\left(2^{N/2-1}, 2^{N/2-2}, \ldots, 2^0, 2^{N/4-1}, 2^{N/4-2}, \ldots, 2^0, \ldots, 0, -1, -1, -1\right)$
C$_2$	$\left(\tilde{b}_1, \tilde{b}_2, \ldots, \tilde{b}_N\right)$ with $\tilde{b}_k = -1/N \cdot \cos(\pi \cdot (k + 1/2)/N))$ for $k = 1, \ldots, N-1$, $\tilde{b}_N = 1$
C$_3$	$(r_0, r_1, \ldots, r_{N-1})$ with $r_k \in \mathcal{N}(0, 1)$

subband index, $1 \leq k \leq K$. A vector $\boldsymbol{f} = (f_1, f_2, \ldots, f_K) = \boldsymbol{y}(n, k)$, $k = 1, 2, \ldots, K$, containing all spectral values for a time index n is called a *frame*.

The way the GCT is applied for the feature extraction within ASR is inspired from [13]. For each frame the GCT is applied on all subframes according to a chosen subframe length and a subframe shift. An example for a frame of length 8, a chosen subframe length of 4 and a subframe shift of 2 is shown in Fig. 1. The combination of a chosen subframe length and subframe shift is called a *subframing scheme* in the following.

An exemplary computation of AVS and EGS features of a signal and a translated version of it is shown in Fig. 2.

III. PHONEME RECOGNITION EXPERIMENTS

A. Experimental Setup

We conducted phoneme recognition experiments for the evaluation of the described feature extraction methods. We used the TIMIT corpus with a sampling rate of 16 kHz. The "SA" sentences have not been used to avoid an unfair bias for certain phonemes [21]. In order to simulate mismatching training and testing conditions with respect to the mean VTL, the training and testing data was split into male and female subsets and three scenarios were defined:

1) Training on both male and female data and testing on male and female data (FM-FM),
2) training on male data and testing on female data (M-F) and
3) training on female data and testing on male data (F-M).

For the TF analysis we used a gammatone filter bank [22]. The number of filters was set to 64 and corresponds to the choice in [23]. We used this number of filters as compromise between frequency resolution and size of feature vector. The minimum center frequency was 40 Hz, the maximum center frequency was set to 8000 Hz. The final frame length was set to 20 ms and the frame shift was 10 ms. The bandwidth was chosen as 1 ERB for each filter. A power-law compression with an exponent of 0.1 was applied in order to resemble the nonlinear compression found in the human auditory system.

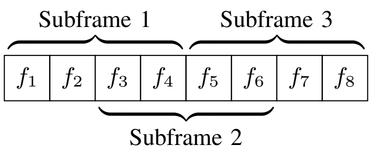
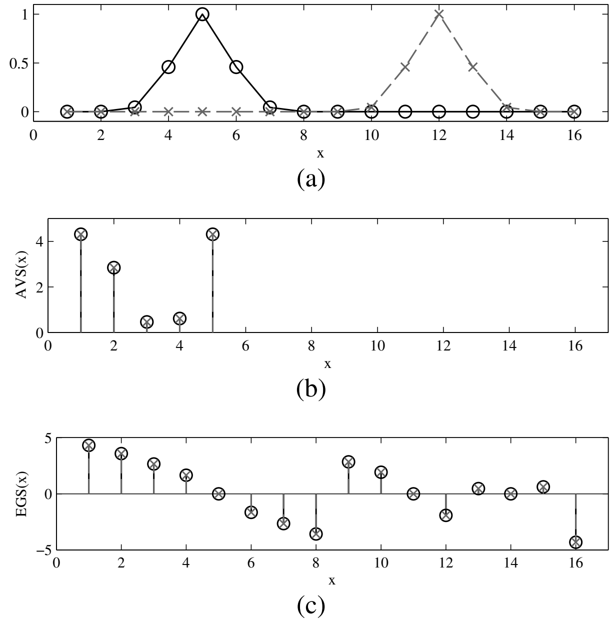


Fig. 1. Exemplary application of the GCT to subframes of length 4 and a subframe shift of 2.



(a)

(b)

(c)

Fig. 2. Example of a signal and its AVS and EGS using MWHT coefficients. (a) Signal $\boldsymbol{x}$ (circles) and a translated version of it (crosses). (b) Average group spectrum (AVS) and (c) extended group spectrum (EGS) of the signal and its shifted version. One can see that the shift does not change the AVS or EGS.

The recognizer was based on the hidden-Markov model toolkit (HTK) [24]. Monophone models with three states per phoneme, 8 Gaussian mixtures per state and diagonal covariance matrices were used together with bigram statistics. According to [21], 48 phonetic models were trained and the recognition results were folded to yield 39 final phoneme classes that had to be distinguished. All feature vectors were supplemented with the first and second order time derivatives. When the size of the feature vector was greater than 47 elements, a linear discriminant analysis was performed such that the feature vectors were reduced to a length of 47.

MFCCs were used to obtain baseline recognition accuracies. They were computed by using the standard HTK setup which yields 12 coefficients plus the logarithmized energy for each

TABLE II
ACCURACIES OF C_1 TRANSFORMATION AS REPRESENTATIVE FOR DIFFERENT SUBFRAMING SCHEMES

Subframe length 64, subframe shift 0

	FM-FM	**M-F**	**F-M**
AVS	52.01	49.18	46.48
EGS	57.14	52.55	50.03

Subframe length 16, subframe shift 16

	FM-FM	**M-F**	**F-M**
AVS	56.63	50.69	49.61
EGS	62.82	55.66	54.87

Subframe length 16, subframe shift 8

	FM-FM	**M-F**	**F-M**
AVS	59.04	51.48	51.23
EGS	64.21	56.09	56.17

frame. The accuracies for the three scenarios when using the MFCCs are as follows: FM-FM: 66.57%, M-F: 55.00% and F-M: 52.42%. It can be seen that the accuracy of the MFCCs declines significantly when the training and testing conditions do not match.

B. Comparison of Individual Transformations & Subframing Schemes

The first part of the experiments evaluated the two feature types AVS and EGS for the coefficients shown in Table I. Three different subframing schemes were considered in these experiments:

1) trivial subframing scheme with subframe length of 64, i.e., the GCT applied on the full frame
2) subframe length of 16 and subframe shift of 16
3) subframe length of 16 and subframe shift of 8

A general observation is that all five coefficient-generating functions from Table I had very similar performances. Thus, the results of the transformation C_1 are shown in Table II as a representative.

Overall, it can be stated that the EGS feature type performs better than the AVS feature type for a given subframing scheme. Using a trivial subframing scheme, both feature types perform worse then the MFCCs. However, using a nontrivial subframing scheme increases the accuracies of both the AVS and the EGS features. The accuracy of the EGS for the mismatching training-testing scenarios is slightly higher than the one of the MFCCs which was about 55% and 52% for the M-F and F-M scenarios, respectively. In comparison with the non-overlapping subframing scheme, the subframing scheme with overlapping subframes shows slightly higher accuracies for both feature types.

C. Combining Translation-Invariant Feature Types

Previous works showed that combinations of translation-invariant feature types are a promising approach for increasing the robustness of features in mismatching training-testing conditions [6], [25]. The second part of the experiments thus investigated combinations of the presented feature types. In addition, feature types based on the autocorrelation and cross correlation [25] of frames were considered in this experiment as well. These feature types are denoted as ACF and CCF, respectively. Again, all five coefficient types as shown in Table I were considered for the GCT. Based on the results described in the previous section, the best subframing scheme was used. Therefore, a subframe length of 16 and a subframe shift of 8 was chosen. For all considered coefficient types, the EGS was computed as GCT-based feature type. All possible combinations of the GCT- and correlation-based feature types with size 2, 3 and 4 have been evaluated. The last part of this experiment supplemented the considered combinations with MFCCs. This was necessary in [25] to boost the performance of the translation-invariant feature types.

Table III shows the combinations that led to the highest accuracies within the described experiments. Generally, it can be observed that the feature type that is based on the cross correlation of two frames is always part of the best feature-type combinations. This indicates the importance of contextual information for the feature extraction in ASR systems. At the top of Table III, results for the best combinations of two feature types are shown. Compared to the accuracies that were achieved with individual GCT-based features, the accuracies in all scenarios further increased to around 65.5% in the matching and slightly less than 60% in the mismatching training-testing conditions. The introduction of a third feature type increases the accuracies in all three scenarios slightly. Again, it can be observed that the highest accuracies with combinations of three feature types were achieved by including the correlation based feature types ACF and CCF. Supplementary, the results for the combination of ACF and CCF features are shown in the middle part of Table III. The fact that the combination of only ACF and CCF leads to lower accuracies in the scenarios shows that the GCT feature types do contain additional discriminative information. The combination of four feature types yields only slight further improvements.

As shown in the last part of Table III the accuracies of the scenarios do not change significantly when the translation-invariant features are supplemented by MFCCs. While the supplementing step was necessary when using only the correlation based feature types ACF and CCF in [25] it was shown in previous work that the combination of correlation based features and another translation-invariant feature type did not

TABLE III
ACCURACIES OF FEATURE-TYPE COMBINATIONS

	Feature types	**FM-FM**	**M-F**	**F-M**
	C_3 + CCF	65.51	59.46	59.58
	MWHT + CCF	65.54	59.39	59.98
	C_3 + ACF + CCF	65.74	60.72	60.59
	MWHT + ACF + CCF	65.78	60.86	60.52
	ACF + CCF	63.17	58.05	56.84
	C_1 + C_2 + ACF + CCF	66.04	60.61	60.71
	C_3 + SWT + ACF + CCF	65.98	60.43	60.92
MFCC	C_3 + CCF	66.5	59.84	59.20
	C_3 + ACF + CCF	66.60	60.80	61.22
	C_1 + C_2 + ACF + CCF	66.56	61.14	61.62

benefit from additional MFCCs [6]. The same observation is made here. The GCT based feature types in combination with ACF and CCF seem to include the discriminative information of the MFCCs.

IV. Conclusions & Future Work

In this work we described the principles of the GCT together with two translation-invariant feature extraction methods as they were introduced in the field of pattern recognition. The transformation can be parametrized with different choices of coefficients. We conducted phoneme recognition experiments for the evaluation of the methods. Different choices of coefficients led to results with no significant differences in our experiments. With respect to the two feature extraction methods, the EGS outperforms the AVS. However, the size of the feature vector of the EGS is bigger than the one of the AVS. The use of a subframing scheme with overlapping subframes lead to higher accuracies than a subframing scheme without overlap or no subframing at all. While the EGS with overlapping subframing scheme leads to accuracies comparable to the one of the MFCCs, the combination of different GCT- and correlation-based feature types increases the accuracies by more than 5% in the scenarios in which the training and testing conditions do not match.

Future work will deal with the refinement of the feature types in order to further increase their robustness against mismatching training-testing conditions. The inclusion of contextual information into the feature extraction methods will be part of this refinement. In theory, the GCT can be parametrized with complex valued coefficients. This could also help to further improve the GCT based features. Optimal parameters for a subframing scheme may also be investigated. Furthermore, previous publications by the authors presented other robust feature types. An evaluation and comparison of these feature types on different corpora and under different noise conditions will yield more distinctive results.

Acknowledgments

This work has been supported by the German Research Foundation under Grant No. ME1170/2-1. Acknowledgements are also directed to the German Academic Exchange Service, that made the stay at the University of Lübeck for Eugene Belilovsky possible.

References

[1] L. Lee and R. C. Rose, "A frequency warping approach to speaker normalization," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 1, pp. 49–60, Jan. 1998.

[2] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[3] M. Pitz and H. Ney, "Vocal tract normalization equals linear transformation in cepstral space," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 5 Part 2, pp. 930–944, 2005.

[4] L. Welling, H. Ney, and S. Kanthak, "Speaker adaptive modeling by vocal tract normalization," *IEEE Trans. Speech and Audio Processing*, vol. 10, no. 6, pp. 415–426, Sept. 2002.

[5] J. J. Monaghan, C. Feldbauer, T. C. Walters, and R. D. Patterson, "Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 123, no. 5, pp. 3066–3066, Jul. 2008.

[6] F. Müller and A. Mertins, "Robust features for speaker-independent speech recognition based on a certain class of translation-invariant transformations," in *LNCS*, in press.

[7] F. Müller and A. Mertins, "Invariant-integration method for robust feature extraction in speaker-independent speech recognition," in *Proc. Int. Conf. Spoken Language Processing (Interspeech 2009-ICSLP)*, Brighton, Sept. 2009.

[8] M. Wagh and S. Kanetkar, "A class of translation invariant transforms," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 25, no. 2, pp. 203–205, Apr. 1977.

[9] H. Burkhardt and X. Müller, "On invariant sets of a certain class of fast translation-invariant transforms," *IEEE Trans. Acoustic, Speech, and Signal Processing*, vol. 28, no. 5, pp. 517–523, Oct. 1980.

[10] M. Fang and G. Häusler, "Modified rapid transform," *Applied Optics*, vol. 28, no. 6, pp. 1257–1262, Mar. 1989.

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

[12] V. Lohweg and D. Müller, "Nonlinear generalized circular transforms for signal processing and pattern recognition," in *IEEE-Eurasip Workshop on Nonlinear Signal and Image Processing*, Baltimore, Jun. 2001.

[13] V. Lohweg, C. Diederichs, and D. Müller, "Algorithms for hardware-based pattern recognition," *EURASIP Journal on Applied Signal Processing*, vol. 2004, pp. 1912–1920, Jan. 2004.

[14] P. J. Davis, *Circulant matrices*, Chelsea Publishing, New York, USA, 1979.

[15] N. Ahmed, K. R. Rao, and A. L. Abdussattar, "BIFORE or Hadamard transform," *IEEE Trans. Audio Electroacoustics*, vol. 19, pp. 225–234, Sept. 1971.

[16] N. U. Ahmed and K. R. Rao, *Orthogonal Transforms for Digital Signal Processing*, Springer, New York, USA, 1975.

[17] J. Pender and D. Covey, "New square wave transform for digital signal processing," *IEEE Trans. Signal Processing*, vol. 40, no. 8, pp. 2095–2097, Aug. 1992.

[18] N. Ahmed, K. R. Rao, and A. L. Abdussattar, "On cyclic autocorrelation and the Walsh-Hadamard transform," *IEEE Trans. Electromagnetic Compatibility*, vol. EMC-15, no. 3, pp. 141–146, Aug. 1973.

[19] V. Lohweg and D. Müller, "A complete set of translation invariants based on the cyclic correlation property of the generalized circular transforms," in *Proc. 6th Digital Image Computing Techniques and Applications (DICTA'02)*, Melbourne, Australia, Jan. 2002, pp. 134–138, Australian Pattern Recognition Society.

[20] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-warping in speech," *Proc. International Conference on Spoken Language (ICSLP 96)*, vol. 1, no. 1, pp. 414–417, Oct. 1996.

[21] K. F. Lee and H. W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.

[22] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand, "Complex sounds and auditory images," in *Auditory Physiology and Perception*, Y. Cazals, L. Demany, and K. Horner, Eds., Pergamon, Oxford, 1992, pp. 429–446.

[23] R. Sinha and S. Umesh, "Non-uniform scaling based speaker normalization," in *Proc. IEEE Int. Conf. Acoust. Speech and Signal Processing (ICASSP'02)*, Orlando, USA, May 2002, vol. 1, pp. I–589 – I–592.

[24] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK version 3.4)*, Cambridge University Engineering Department, Cambridge, Dec. 2006.

[25] A. Mertins and J. Rademacher, "Frequency-warping invariant features for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing*, Toulouse, France, May 2006, vol. V, pp. 1025–1028.