Nonlinear translation-invariant transformations for speaker-independent speech recognition

Florian Müller and Alfred Mertins

Institute for Signal Processing
University of Lübeck, Germany
mueller@isip.uni-luebeck.de

Abstract. The spectral effects of vocal tract length (VTL) changes are one reason of why the recognition rate of today's speaker-independent automatic speech recognition (ASR) systems is considerably lower than the one of speaker-dependent systems. By using certain types of filterbanks these effects can be described by a translation in subband-index space. In this paper, nonlinear translation-invariant transforms that originally have been proposed in the field of pattern recognition are investigated for their applicability in speaker-independent ASR tasks. It will be shown that the combination of different types of such transforms leads to features that are more robust against VTL changes than the standard Mel-frequency cepstral coefficients and that almost yield the performance of vocal tract length normalization without any adaption to individual speakers.

1 Introduction

The vocal tract length (VTL) is one of the intrinsic variabilities which causes the error rate of today's speaker-independent automatic speech recognition (ASR) systems to be two to three times higher than for speaker-dependent ASR systems [1]. Besides its shape it is the length of the vocal tract that determines the location of the resonance frequencies (formants). The formants determine the overall envelope of the short-time spectra of a voiced utterance. Given speakers A and B their short-time spectra are approximately related as $X_A(\omega) = X_B(\alpha{\cdot}\omega)$ in case of the same utterance. The factor α is the VTL ratio of the two speakers and is known as the warping factor. Typically, α ranges from about 0.8 to 1.2 in a speaker-independent ASR task [2].

Several techniques for handling this warping effect have been proposed. One group of techniques tries to adapt the acoustic models of the recognition system to the individual speakers, e.g, [3]. Other methods try to normalize the spectral effects of different VTLs at the feature extraction stage [4, 2]. A third group of methods tries to generate features that are independent of the warping factor [5–8].

Standard approaches for the time-frequency (TF) analysis of speech signals locate the frequency centers of the filters equally spaced on the quasi-logarithmic ERB scale that approximately represents the frequency resolution of the human auditory system. In this domain linear frequency warping becomes a translation. This effect can be utilized for the computation of features on basis of the TF-analysis that are invariant under translation [6–8].

Nonlinear transformations that lead to translation-invariant features have been investigated and successfully applied for decades in the field of pattern recognition. Following the concepts of [9], the idea of invariant features is to find a mapping T that is able to extract the features that are the same for possibly different observations of the same equivalence class with respect to a group action. Such a transformation T maps all observations of an equivalence class into one point of the feature space. Given a transformation T, the set of all observations that are mapped into one point is denoted as the set of invariants of an observation. The set of all possible observations within one equivalence class is called orbit. 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. On the other hand, incomplete transformations can lead to the same features from observations of different equivalence classes and thus cannot distinguish them [9].

The idea of the method proposed in this paper is to gain features that are robust against VTL changes by using nonlinear transformations that are invariant to translations. Well known transforms of this type are, for example, the cyclic autocorrelation of a sequence and the modulus of the discrete Fourier transform (DFT). A general class of translation-invariant transforms was introduced by [10] and further investigated in [11, 12] in the field of pattern recognition. It is called the class $\mathbb{C}T$.

In this paper, different transforms of the class $\mathbb{C}T$ will be investigated with the aim of obtaining a feature set that has a high degree of completeness under the group action induced by VTL changes. Experimental results will be given for phoneme recognition tasks in which there is a mismatch in the average vocal tract length between the training and test sets. While the experiments showed that, under such conditions, the individual transforms of the class $\mathbb{C}T$ as well as previously investigated individual transforms [6, 7] achieve at most a recognition performance that is comparable to the one of Mel-Frequency Cepstral Coefficients (MFCCs), it turned out that combinations of different transforms increase the degree of completeness, so that combined invariant transforms significantly outperform the MFCCs with respect to the problem of VTL changes.

The paper is organized as follows. The next section introduces the class of transforms $\mathbb{C}T$ and explains our method for using these transforms for the calculation of features for speech recognition tasks. Section 3 describes the experimental setup. The results are presented in Section 4 with a subsequent conclusion in the last section.

2 Method

2.1 Translation-invariant transformations of class $\mathbb{C}T$

A general class of translation-invariant transforms was originally introduced by [10] and later given the name $\mathbb{C}T$ [11]. Their computation is based on a generalization of the linear, fast Walsh-Hadamard transform (WHT). Given a vector $\boldsymbol{x} := (x_0, x_1, \ldots, x_{N-1})$ with $N = 2^M$ as input and following the notation of [9], members of the class $\mathbb{C}T$ are defined by the following recursive transformation T with commutative operators $f_1(.,.)$, $f_2(.,.)$:

$$T(\boldsymbol{x}) := \big(T(f_1(\boldsymbol{x}_{1|2}, \boldsymbol{x}_{2|2})),\, T(f_2(\boldsymbol{x}_{1|2}, \boldsymbol{x}_{2|2}))\big)\,, \tag{1}$$

where $\boldsymbol{x}_{1|2}$ and $\boldsymbol{x}_{2|2}$ denote the first and second halves of the vector $\boldsymbol{x}$, respectively. The recursion starts with $T(x_i) = x_i$. Fig. 1 shows a corresponding signal-flow diagram for $N = 4$.

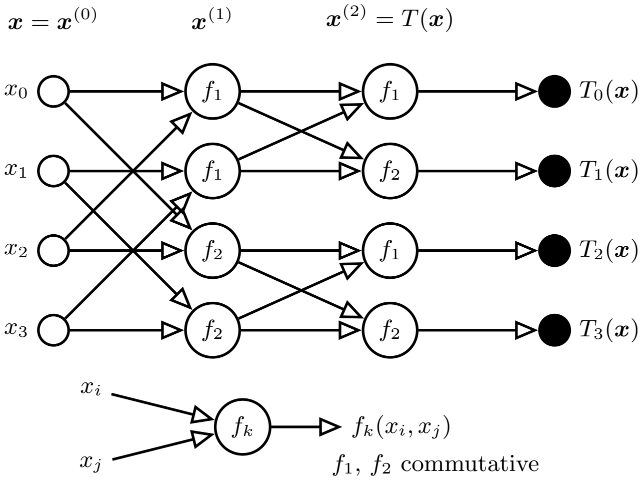

Fig. 1. Signal-flow diagram for transformations of the class $\mathbb{C}T$ for $N = 4$.

The pairs of commutative operators that are examined in this work have found applications in pattern recognition tasks before [13, 10, 11]. One representative of the class $\mathbb{C}T$ is the “Rapid Transform” (RT) which has found a notably wide application [14, 15, 9]. In comparison to the RT, it was shown in [10] that taking the $\min(.,.)$ and $\max(.,.)$ functions as f_1 and f_2, respectively, can lead to a higher separability. This transformation with its pair of functions is denoted as MT. It was shown in [11] that the power spectrum of the modified WHT can be computed with a transformation of $\mathbb{C}T$ by choosing $f_1 := a+b$ and $f_2 := (a-b)^2$. This transformation is denoted as QT. The mentioned transformations together with their according pairs of commutative functions are shown in Table 1.

Table 1. Common pairs of commutative operators

	RT	MT	QT
f_1	$a+b$	$\min(a,b)$	$a+b$
f_2	$\|a-b\|$	$\max(a,b)$	$(a-b)^2$

In [12] a preprocessing operator b for the RT was presented that destroys the unwanted property of invariance under reflection of the input data. This operator works elementwise and is defined as

$$x_i' = b(x_i, x_{i+1}, x_{i+2}) := x_i + |x_{i+1} - x_{i+2}|. \qquad (2)$$

This particular preprocessing followed by the RT is called "Modified Rapid Transform" (MRT).

2.2 Translation-invariant feature candidates for ASR

The translation-invariant features are computed on the basis of primary features given by the result of a TF-analysis of an input signal $\boldsymbol{x}$. The TF analysis will be denoted by $y_{\boldsymbol{x}}(t,k)$ in the following. Here, t is the frame index, $1 \leq t \leq T$, and k is the filter index with $1 \leq k \leq K$. The transformations RT, MRT, MT and QT are applied framewise to the primary features. In addition to the transformations described above, individual translation-invariant features from previous work [6, 7] will also be considered in this study. These are based on the logarithmized correlation sequences of spectral values,

$$\log r_{yy}(t,d,m) \quad \text{with} \quad r_{yy}(t,d,m) = \sum_k y_{\boldsymbol{x}}(t,k) y_{\boldsymbol{x}}(t-d,k+m), \qquad (3)$$

and on the correlation sequences of logarithmized spectral values,

$$c_{yy}(t,d,m) = \sum_k \log(y_{\boldsymbol{x}}(t,k)) \cdot \log(y_{\boldsymbol{x}}(t-d,k+m)). \qquad (4)$$

In addition to using the TF analysis $y_{\boldsymbol{x}}(t,k)$ directly, we also consider multiscale representations of it. The method of multi-scale analysis has been successfully applied to various fields of speech processing [16–19]. Therefore, multiple scales of spectral resolution of each frame were computed. The length of each frame of scale n was half of the length of scale $n-1$. Each scale was used as input to the described transformations and the results of the transformations on each scale were concatenated. Following this procedure, the resulting number of features for an input of size $N = 2^M$ is $2^{M+1} - 1$. Features of this type are denoted with the subscript "Scales". For the experiments, also a subset of 50 features of the "Scales"-versions of the features was determined by applying a feature selection method according to [20]. These feature sets are denoted with the subscript "Scales-50".

3 Experimental setup

On the basis of the described feature types, different feature sets have been defined and evaluated in a number of phoneme recognition experiments. The experiments have been conducted using the TIMIT corpus with a sampling rate of 16 kHz. To avoid an unfair bias for certain phonemes, we chose not to use the "SA" sentences in training and testing similar to [21]. The remaining training and test sets were both split into female and male utterances. This was used to create three different training and testing scenarios: Training/testing on both, male and female data (FM-FM), training on male and testing on female data (M-F) and training on female and testing on male data (F-M). 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.

The recognizer was based on the Hidden-Markov Model Toolkit (HTK) [22]. Monophone models with three states per phoneme, 8 Gaussian mixtures per state and diagonal covariance matrices were used together with a bigram language model.

MFCCs were used to obtain baseline recognition accuracies. The MFCCs were calculated by using the standard HTK setup which yields 12 coefficients and a single energy feature for each frame. For comparison with a vocal tract length normalization (VTLN) technique, the method of [2] was used.

We chose to use a complex-valued Gammatone filterbank [23] with 90 filters equally-spaced on the ERB scale as basis for computing the translation-invariant features. This setup was chosen to allow for a comparison with previous works (cf. [6, 7]). The magnitudes of the subband signals were lowpass filtered in order to decrease the time resolution to 20 ms. These filtered magnitudes were then subsampled to obtain a final frame rate of one frame every 10 ms. Because the transforms of the class $\mathbb{C}T$ require the length of the input data to be a power of two, the output of the filterbank was frame-wise interpolated to 128 data points, and then a power-law compression [24] with an exponent of 0.1 was applied in order to resemble the nonlinear compression found in the human auditory system.

The following feature types known from [6, 7] were investigated in addition to class-$\mathbb{C}T$ features: The first 20 coefficients of the discrete cosine transform (DCT) of the correlation term (3) with $d = 0$ ("ACF") have been used as well as the first 20 coefficients of the DCT of the term (4) with $d = 4$ ("CCF"). The features belonging to the class $\mathbb{C}T$ as described in the previous section were considered together with their "Scales" and "Scales-50" versions. The "Scales-50" versions were used for feature set combinations of size four and five.

All feature sets were amended by the logarithmized energy of the original frames together with delta and delta-delta coefficients [22]. The resulting features were reduced to 47 features with linear discriminant analysis (LDA). The reduction matrices of the LDAs were based on the 48 phonetic classes contained in both, the male and female utterances.

4 Results and Discussion

At first, each of the previously described feature types were tested individually in the three scenarios. The resulting percentage accuracies [22] of these experiments are shown in Table 2. It can be seen that the MFCCs have the highest accuracy for the FM-FM scenario compared to the other considered feature types. The features resulting from the RT and MRT obtain similar accuracies as the MFCCs in the gender-separated scenarios, but perform worse in the general FM-FM scenario. The inclusion of different scales in the feature sets leads to accuracies that are comparable to those of the MFCCs in the FM-FM scenario and already outperform the MFCCs in the gender-separated scenarios M-F and F-M. Using only the 50 best features from the "Scales"-feature sets leads to accuracies that are similar to the feature sets that include all scales. However, in the gender separated scenarios the "Scales-50" versions perform worse than the "Scales" version.

Table 2. Percentage accuracies of individual feature types [%]

Feature type	FM-FM	M-F	F-M
1. MFCC	66.57	55.00	52.42
3. RT	58.39	55.30	51.99
4. MRT	57.90	53.88	50.75
5. QT	53.00	48.03	46.12
6. MT	59.96	56.53	54.45
7. RT$_{\text{Scales}}$	64.29	57.36	56.67
8. MRT$_{\text{Scales}}$	64.27	58.90	58.42
9. QT$_{\text{Scales}}$	62.64	56.75	55.34
10. MT$_{\text{Scales}}$	64.05	58.79	58.02
11. RT$_{\text{Scales-50}}$	64.47	55.49	54.28
12. MRT$_{\text{Scales-50}}$	64.08	55.66	54.03
13. QT$_{\text{Scales-50}}$	62.25	53.07	52.15
14. MT$_{\text{Scales-50}}$	64.19	53.77	52.38
15. ACF	58.85	46.97	48.76
16. CCF	62.46	54.54	53.41

As a further performance benchmark, the VTLN method has been tested on the three scenarios. Since this method adapts to the vocal tract length of each individual speaker, it gave the best performance in all cases. The results were as follows: FM-FM: 68.61%, M-F: 64.02%, F-M: 63.39%.

To investigate in how far the performance of the translation-invariant features can be increased through the combination of different feature types, all possible combinations of the "Scales"-versions of the features and the ACF and CCF features have been considered. These include feature sets of two, three, four, and five types of features. For each of these feature sets of different size, the results for the best combinations are shown in Table 3.

Table 3. Highest percentage accuracies for feature sets with different sizes and energy amendment [%]

Feature type combination + energy	FM-FM	M-F	F-M
$\text{MT}_{\text{Scales}}$ + CCF	65.74	61.13	60.52
MRT + CCF	65.36	60.60	60.51
$\text{MRT}_{\text{Scales}}$ + CCF + ACF	65.90	61.75	61.94
$\text{MRT}_{\text{Scales}}$ + $\text{MT}_{\text{Scales}}$ + CCF	65.71	62.01	61.94
$\text{MRT}_{\text{Scales-50}}$ + CCF + ACF + $\text{RT}_{\text{Scales-50}}$	66.01	61.27	60.59
$\text{MRT}_{\text{Scales-50}}$ + CCF + ACF + $\text{RT}_{\text{Scales-50}}$ + $\text{QT}_{\text{Scales-50}}$	65.94	61.77	61.17

As the results show, already the combination of two well-selected feature sets leads to an accuracy that is comparable to the MFCCs in the general FM-FM scenario. In contrast to the MFCCs the gender separated scenarios lead to an accuracy that is 5.6% to 7% higher in the M-F scenario and 8.1% to 9.5% higher in the F-M scenario. In particular, the results indicate that the information contained in the CCF features is quite complementary to the one contained in the class-$\mathbb{C}T$ features. Also the MRT and MT features seem to contain complementary information. The fact that the accuracies do not increase by considering combinations of four or five feature sets could either be explained by the fact that the “Scales-50” features in comparison to the “Scales” features have a much lower accuracy for the gender separated scenarios or by the assumption that the RT, MRT and QT do contain quite similar information.

In a third experiment, we amended the previously considered, fully translation-invariant features with MFCCs, as this had been necessary to boost the performance with the method in [6, 7]. The results of the experiment are shown in Table 4. It is notable that the MFCCs do not increase the accuracies significantly in the FM-FM scenarios in all combinations. This means that the MFCCs do not carry additional discriminative information compared to the feature set combinations that consist of translation-invariant features.

Table 4. Highest percentage accuracy for feature type combinations with different sizes and MFCC amendment [%]

Feature type combination + MFCC	FM-FM	M-F	F-M
$\text{MT}_{\text{Scales}}$ + CCF	65.58	61.62	61.46
$\text{MRT}_{\text{Scales}}$ + CCF + ACF	66.45	62.08	62.40
$\text{MRT}_{\text{Scales-50}}$ + CCF + ACF + $\text{RT}_{\text{Scales-50}}$	66.34	61.90	61.23
$\text{MRT}_{\text{Scales-50}}$ + CCF + ACF + $\text{RT}_{\text{Scales-50}}$ + $\text{QT}_{\text{Scales-50}}$	66.46	61.85	61.96

Using the improved feature set (based on Gammatone analysis) presented in the previous work [7] within the described experimental setup of this work leads to the following accuracies: FM-FM: 65.70%, M-F: 60.75% and F-M: 59.90%. These results indicate a better performance in the gender separated scenarios than the MFCCs. However, the new translation-invariant feature sets presented in this paper perform even better.

5 Conclusions

Vocal tract length changes lead to translations in the subband-index space of time-frequency analyses when they are performed on a (quasi-) logarithmic frequency scale. Well-known translation-invariant transformations that were originally proposed in the field of pattern recognition have been applied in this paper in order to obtain features that are more robust to the effects of VTL changes. We showed that combining certain types of translation-invariant feature leads to accuracies that are similar to those of MFCCs in case of training and testing on male and female data and outperform MFCCs in case of gender-separated training and testing. This may lead to significantly more robustness in scenarios in which VTLs differ significantly, as, for example, in children speech. Therefore, children speech and further feature optimization will be subject of future investigations of nonlinear feature-extraction methods. Compared to the VTLN method, our features do not require any speaker adaptation and are therefore much faster to compute and to use than VTLN.

Acknowledgments

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.: Automatic speech recognition and speech variability: a review. Speech Communication **49**(10-11) (Oct.-Nov. 2007) 763–786
2. Welling, L., Ney, H., Kanthak, S.: Speaker adaptive modeling by vocal tract normalization. IEEE Transactions on Speech and Audio Processing **10**(6) (Sept. 2002) 415–426
3. Leggetter, C., Woodland, P.: Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. Computer, Speech and Language **9**(2) (Apr. 1995) 171–185
4. Lee, L., Rose, R.C.: A frequency warping approach to speaker normalization. IEEE Transactions on Speech and Audio Processing **6**(1) (Jan. 1998) 49–60
5. Umesh, S., Cohen, L., Marinovic, N., Nelson, D.J.: Scale transform in speech analysis. IEEE Transactions on Speech and Audio Processing **7**(1) (Jan. 1999) 40–45

6. Mertins, A., Rademacher, J.: Frequency-warping invariant features for automatic speech recognition. In: Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing. Volume V., Toulouse, France (May 2006) 1025–1028
7. Rademacher, J., Waechter, M., Mertins, A.: Improved warping-invariant features for automatic speech recognition. Proc. International Conference on Spoken Language Processing (Interspeech2006 - ICSLP) (Sept. 2006) 1499–1502
8. Monaghan, J.J., Feldbauer, C., Walters, T.C., Patterson, R.D.: Low-dimensional, auditory feature vectors that improve vocal-tract-length normalization in automatic speech recognition. The Journal of the Acoustical Society of America **123**(5) (Jul. 2008) 3066–3066
9. Burkhardt, H., Siggelkow, S.: Invariant features in pattern recognition – fundamentals and applications. In: Nonlinear Model-Based Image/Video Processing and Analysis. John Wiley & Sons (2001) 269–307
10. Wagh, M., Kanetkar, S.: A class of translation invariant transforms. IEEE Transactions on Acoustics, Speech, and Signal Processing **25**(2) (Apr. 1977) 203–205
11. Burkhardt, H., Müller, X.: On invariant sets of a certain class of fast translation-invariant transforms. IEEE Transactions on Acoustic, Speech, and Signal Processing **28**(5) (Oct. 1980) 517–523
12. Fang, M., Häusler, G.: Modified rapid transform. Applied Optics **28**(6) (Mar. 1989) 1257–1262
13. Reitboeck, H., Brody, T.P.: A transformation with invariance under cyclic permutation for applications in pattern recognition. Inf. & Control. **15** (1969) 130–154
14. Wang, P.P., Shiau, R.C.: Machine recognition of printed chinese characters via transformation algorithms. Pattern Recognition **5**(4) (1973) 303–321
15. Gamec, J., Turan, J.: Use of Invertible Rapid Transform in Motion Analysis. Radioengineering **5**(4) (Dec. 1996) 21–27
16. Pinkowski, B.: Multiscale fourier descriptors for classifying semivowels in spectrograms. Pattern Recognition **26**(10) (1993) 1593–1602
17. Stemmer, G., Hacker, C., Noth, E., Niemann, H.: Multiple time resolutions for derivatives of Mel-frequency cepstral coefficients. IEEE Workshop on Automatic Speech Recognition and Understanding. (Dec. 2001) 37–40
18. Mesgarani, N., Shamma, S., Slaney, M.: Speech discrimination based on multiscale spectro-temporal modulations. IEEE International Conference on Acoustics, Speech, and Signal Processing **1** (May 2004) I–601–I–604
19. Zhang, Y., Zhou, J.: Audio segmentation based on multi-scale audio classification. IEEE International Conference on Acoustics, Speech, and Signal Processing **4** (May 2004) iv–349–iv–352
20. Peng, H., Long, F., Ding, C.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(8) (Aug. 2005) 1226–1238
21. Lee, K.F., Hon, H.W.: Speaker-independent phone recognition using hidden markov models. IEEE Transactions on Acoustics, Speech and Signal Processing **37**(11) (Nov. 1989) 1641–1648
22. 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.: The HTK Book (for HTK version 3.4). Cambridge University Engineering Department, Cambridge. (Dec. 2006)
23. Patterson, R.D.: Auditory images: How complex sounds are represented in the auditory system. Journal-Acoustical Society of Japan (E) **21**(4) (2000) 183–190
24. Bacon, S.P., Fay, R.R., Popper, A.N.: Compression: from cochlea to cochlear implants. IV. Springer, New York (2004)