Robust Features for Speaker-Independent Speech Recognition Based on a Certain Class of Translation-Invariant Transformations*

Florian Müller and Alfred Mertins

Institute for Signal Processing, University of Lübeck, 23538 Lübeck, Germany
{mueller,mertins}@isip.uni-luebeck.de

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

Keywords: Speech recognition, speaker-independency, translation-invariance.

1 Introduction

The vocal tract length (VTL) is a source of variability 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, commonly known as "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.

Several techniques for handling this warping effect have been proposed. One group of techniques tries to adapt the acoustic models to the features of the individual speakers, for example, [2,3]. These methods are also known as (constrained) MLLR techniques. Other methods try to normalize the spectral effects of different VTLs at the feature extraction stage [4,5] in order to reduce the mismatch between training and testing conditions. Though both groups of methods

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

J. Solé-Casals and V. Zaiats (Eds.): NOLISP 2009, LNAI 5933, pp. 111–119, 2010.

are working in different stages of an ASR system, it was shown in [3] that they are related by a linear transformation. In contrast to the mentioned techniques, a third group of methods avoids the additional adaption step within the ASR system by generating features that are independent of the warping factor [6,7,8,9].

A known approach for the time-frequency (TF) analysis of speech signals is to equally space the center frequencies of the filters on the quasi-logarithmic ERB scale. This scale approximately represents the frequency resolution of the human auditory system. Within this domain, linear frequency warping can approximately be described by a translation. On basis of the TF-analysis, this effect can be utilized for the computation of translation-invariant features [7,8,9].

Nonlinear transformations that lead to translation-invariant features have been investigated and successfully applied in the field of pattern recognition for decades. Following the concepts of [10], the general idea of invariant features is to find a mapping T that is able to extract features in such a way that they are the same for possibly different observations of the same equivalence class with respect to a group action. Thus, 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. In practical applications, however, usually one has to deal with non-complete transformations.

The idea of the method proposed in this work is to extract features that are robust against VTL changes by using nonlinear transformations that are invariant to translations. Well-known transformations of this type are, for example, the cyclic autocorrelation of a sequence and the modulus of the discrete Fourier transformation (DFT). A general class of translation-invariant transformations was introduced in [11] and further investigated in [12,13] in the field of pattern recognition. The transformations of this class, which will be called $\mathbb{C}T$ in the following, can be computed efficiently.

Different transformations will be investigated in this paper with the aim of obtaining a feature set that has a high degree of completeness under the group action induced by VTL changes. Results will be presented for large-vocabulary phoneme recognition tasks with a mismatch in the mean VTL between the training and testing sets. Experiments show that the individual transformations of the class $\mathbb{C}T$ as well as previously investigated individual transforms [7,8] achieve a recognition performance that is comparable to the one of mel-frequency cepstral coefficients (MFCC). However, combinations of different invariant transformations significantly outperform the MFCCs with respect to the problem of VTL changes.

The next section introduces the class of transformations $\mathbb{C}T$ and explains our method for using these transformations for the extraction of features for speech

recognition tasks. Section 3 describes the experimental setup. Results are presented in Section 4, followed by some conclusions in the last section.

2 Method

2.1 Translation-Invariant Transformations of Class $\mathbb{C}T$

A general class of translation-invariant transformations was originally introduced in [11] and later given the name $\mathbb{C}T$ [12]. Their computation is based on a generalization of the 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 [10], members of the class $\mathbb{C}T$ are defined by the following recursive transformation T with commutative operators $f_1(.,.)$, $f_2(.,.)$:

$$T(\boldsymbol{x}) := (T(f_1(\boldsymbol{x}_1, \boldsymbol{x}_2)),\, T(f_2(\boldsymbol{x}_1, \boldsymbol{x}_2)))\,. \tag{1}$$

Herein, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ denote the first and second halves of the vector $\boldsymbol{x}$, respectively. The recursion ends with $T(x_i) = x_i$. Fig. 1 shows a corresponding signal-flow diagram for $N = 4$.

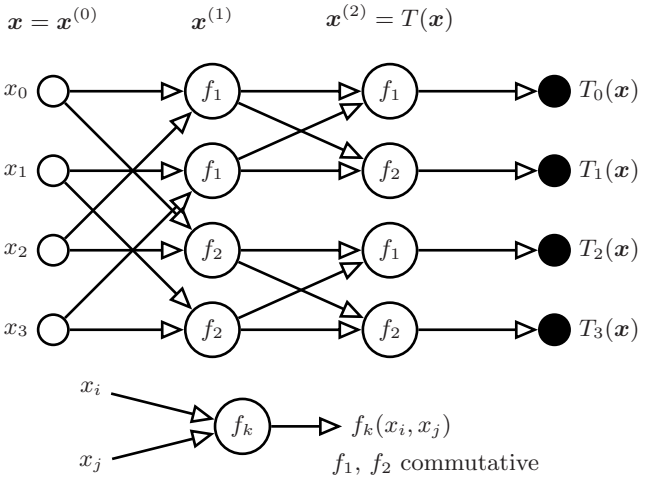

Fig. 1. Signal-flow diagram for transformations of the class $\mathbb{C}T$ for $N = 4$

The pairs of commutative operators that are examined in this work have found applications in pattern recognition tasks [11,12,14]. One representative of the class $\mathbb{C}T$ is the *rapid transformation* (RT) which has found a notably wide application [10,15,16]. In comparison to the RT, it was shown in [11] that taking the min(.,.) and max(.,.) functions as f_1 and f_2, respectively, can lead to better separability properties. The transformation with this pair of functions is denoted as "MT". It was shown in [12] that the power spectrum of the modified

Table 1. Common pairs of commutative operators

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

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 summarized in Table 1.

In [13], a preprocessing operator for the RT was presented that destroys the unwanted property of invariance under reflection of the input data, and, thus, increases the separation capability of the RT. This operator, denoted as b, works element-wise and is defined as

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

This particular preprocessing followed by the RT is called *modified rapid transformation* (MRT).

2.2 Translation-Invariant Feature Candidates for ASR

The translation-invariant features for an input signal $\boldsymbol{x}$ are computed on the basis of the magnitude of a TF-analysis. The TF representation 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 subband index with $0 \leq k \leq K-1$. The transformations RT, MRT, MT, and QT are applied frame-wise to the primary features. In addition to the transformations described above, individual translation-invariant features from previous work [7,8] are also considered in this study. These are based on the logarithmized correlation sequences of spectral values,

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

and on the correlation sequences of logarithmized spectral values,

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

Besides using the TF analysis $y_{\boldsymbol{x}}(t,k)$ directly as input for the transformations, we also consider multi-scale representations of it. The method of multi-scale analysis has been successfully applied to various fields of speech processing [17,18,19,20]. Multiple scales of spectral resolution for each frame were computed. The length of a frame on scale n is half 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 [21]. The feature-selection method uses a mutual information criterion, and the resulting feature sets are denoted with the subscript "Scales-50".

3 Experimental Setup

On the basis of the feature types described in Section 2, 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 [22]. Training and testing sets were both split into female and male utterances. This was used to simulate matching and mismatching training and testing conditions with respect to the mean VTL. Three different training and testing scenarios were defined: Training and 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 [22], 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) [23]. Monophone models with three states per phoneme, 8 Gaussian mixtures per state and diagonal covariance matrices were used together with bigram language statistics.

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 [4] was used.

We chose to use a complex-valued Gammatone filterbank [24] 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 the previous works [7,8]. The magnitudes of the subband signals were low-pass 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. A power-law compression [25] 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 [7,8] 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 $d = 0$ (denoted as "ACF") have been used, as well as the first 20 coefficients of the DCT of the term (4) with $d = 4$ (denoted as "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. In

order to limit the size of the resulting feature vectors, 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 [23]. 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, all of the previously described feature types were tested individually in the three scenarios. The resulting accuracies 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 mismatching scenarios, but perform worse in the matching 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 mismatching 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 mismatching scenarios the "Scales-50" versions perform worse than the "Scales" version. The correlation-based features perform similar to the $\mathbb{C}T$-based ones. In comparison, the cross correlation features CCF perform better than the ACF features. This indicates the importance of contextual information for ASR.

As a further baseline performance, the VTLN method [4] using MFCCs as features 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.

As the results show, the combination of two well-selected feature sets leads to an accuracy that is comparable to the MFCCs in the matching case. However, in contrast to the MFCCs, feature-type combinations lead to an accuracy that was 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 that contained in the $\mathbb{C}T$-based features. Also the MRT and MT features seem to contain complementary information. The observation 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

Table 2. Accuracies of individual feature types

Feature type	FM-FM	M-F	F-M
MFCC	66.57	55.00	52.42
RT	58.39	55.30	51.99
MRT	57.90	53.88	50.75
QT	53.00	48.03	46.12
MT	59.96	56.53	54.45
RT_{Scales}	64.29	57.36	56.67
MRT_{Scales}	64.27	58.90	58.42
QT_{Scales}	62.64	56.75	55.34
MT_{Scales}	64.05	58.79	58.02
$RT_{Scales\text{-}50}$	64.47	55.49	54.28
$MRT_{Scales\text{-}50}$	64.08	55.66	54.03
$QT_{Scales\text{-}50}$	62.25	53.07	52.15
$MT_{Scales\text{-}50}$	64.19	53.77	52.38
ACF	58.85	46.97	48.76
CCF	62.46	54.54	53.41

Table 3. Highest accuracies for feature sets with different sizes and energy amendment

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

lower accuracy for the gender separated scenarios or by the assumption that the RT, MRT and QT do contain 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 [7,8]. The results of the experiment are shown in Table 4. It is notable that the MFCCs do not increase the accuracies significantly in the matching scenarios and increase only slightly in the mismatching scenarios. This means that the MFCCs do not carry much more additional discriminative information compared to the feature set combinations that solely consist of translation-invariant features.

Using the best feature set presented in the previous work [8] leads to the following accuracies: FM-FM: 65.70%, M-F: 60.75% and F-M: 59.90%. As expected, 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 and do not rely on the MFCCs.

Table 4. Highest accuracy for feature type combinations with different sizes and MFCC amendment

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

5 Conclusions

Vocal tract length changes approximately lead to translations in the subband-index space of time-frequency representations if an auditory-motivated filterbank is used. 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 robust to the effects of VTL changes. We showed that combining certain types of translation-invariant features leads to accuracies that are similar to those of MFCCs in case of matching training and testing conditions with respect to the mean VTL. For mismatched training and testing conditions, the proposed features significantly outperform the MFCCs. 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 on nonlinear feature-extraction methods. Also the combination with other invariant feature types will be investigated. 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.

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), 763–786 (2007)
2. Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language 12(2), 75–98 (1998)
3. Pitz, M., Ney, H.: Vocal tract normalization equals linear transformation in cepstral space. IEEE Trans. Speech and Audio Processing 13(5 Part 2), 930–944 (2005) (ausgedruckt)
4. Welling, L., Ney, H., Kanthak, S.: Speaker adaptive modeling by vocal tract normalization. IEEE Trans. Speech and Audio Processing 10(6), 415–426 (2002)
5. Lee, L., Rose, R.C.: A frequency warping approach to speaker normalization. IEEE Trans. Speech and Audio Processing 6(1), 49–60 (1998)
6. Umesh, S., Cohen, L., Marinovic, N., Nelson, D.J.: Scale transform in speech analysis. IEEE Trans. Speech and Audio Processing 7, 40–45 (1999)

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