ROBUST CONTINUOUS SPEECH RECOGNITION THROUGH COMBINATION OF INVARIANT-FEATURE BASED SYSTEMS

Florian Müller and Alfred Mertins

Institute for Signal Processing, Ratzeburger Allee 160, 23562 Lübeck
{mueller,mertins}@isip.uni-luebeck.de

Abstract: In the recent years, different types of invariant features have been proposed which promise to improve the robustness of speech recognition systems in mismatching training-test conditions with respect to the mean vocal tract lengths. Many state-of-the-art systems make use of system combination. By considering speech recognition systems with different front ends, this paper investigates whether the system combination of standard-feature and invariant-feature based systems with ROVER yields improvements in accuracy. Results show that the combination of the considered systems leads to clear accuracy improvements. An error analysis also shows that the considered invariant features carry different types of information than the standard ones. The performance achieved with our system combination is in the range of what the best systems achieve in literature, although our approach does not yet include discriminative training or heteroscedastic feature transformation.

1 Introduction

Variability is an inherent characteristic of speech signals. One of the many sources of variability comes from the different *vocal tract lengths* (VTLs) of human speakers. Adult VTLs may differ by up to 25 percent [9]. Without further processing, the performance of speaker-independent *automatic speech recognition* (ASR) systems is two to three times lower than the one of speaker-dependent systems [1].

State-of-the-art ASR systems usually combine several methods to enhance their robustness in speaker-independent applications. On the one hand, there are methods that try to adapt the acoustic models to the characteristics of each speaker, which are also known as (constrained) *maximum likelihood linear regression* (MLLR) techniques [6]. 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 [9] belong to this group. In [18] it was shown that VTLN can be seen as a constrained MLLR. A third class of methods tries to directly extract features that are invariant to the spectral effects of VTL changes [11, 12, 13, 14, 15].

The application of auditory motivated scales like the mel [3] or the ERB [17] scale approximately maps the scaling to translation along the frequency axis of TF representations. This effect was used for translation-based VTLN [21], as well as for *Gaussian-mixture-model* (GMM) based features [12]. Recently, different types of translation-invariant transformations were investigated for their applicability in the field of speech recognition. Correlation-based features were proposed in [11]. Besides other invariant-transformation based features [13, 14], so-called *invariant integration features* (IIFs) were described in [15] and showed a clear performance enhancement for speaker-independent ASR systems. The transformations that are used within

these feature extraction methods are mathematically well founded and were successfully used in the field of image analysis.

Another common approach of advanced ASR systems is system combination. Besides cross-adaptation [7] that uses the output of one system as the input of another system, another common method is to take hypotheses of multiple systems and combine them. This could be done, for example, with the *recognizer output voting error reduction* (ROVER) [5] method or *confusion network combination* (CNC) [4]. Compared to the accuracies of single recognition systems, the combining approaches have shown the ability for large performance improvements [22, 20].

In this work, we investigate the combination of the outputs of systems, whose front ends use standard feature extraction methods, as well as invariant feature extract methods as mentioned above. The following section describes four different feature types that use different translation-invariant transformations and were used in recent works. In Section 3.1, the data and modeling of the individual systems is described. Baseline accuracies and an error analysis are made in Section 3.2. The system combinations are investigated in the third part of Section 3. At last, Section 4 summarizes the main contributions of this paper and describes future plans.

2 Invariant Feature Types

As described above, feature extraction methods have been proposed for ASR systems that are based on different nonlinear invariant transformations. All methods have in common that they rely on a TF representation that approximately maps the spectral effects due to VTL differences to translations along the subband-index space. Common first steps of an ASR front-end are the computation of the power spectrum for a given frame followed by a critical-band grouping. All feature types considered in this work initially compute this sort of representation. In the following, we give a brief description of the individual invariant feature types.

The *vocal tract length invariant* (VTLI) features [11] use the auto- and cross-correlation function for feature computation. They comprise three different correlation-based feature types that differ in the parameters for the temporal and spectral lags. In contrast to the other invariant feature types considered in this work, the VTLI features are also supplemented with *mel frequency cepstral coefficients* (MFCCs).

The second feature type uses an invariant transformation that originates from the field of image analysis and is known as the class of translation-invariant transformations $\mathbb{C}T$ [2]. Transformations of this class can efficiently be computed with a complexity of $\mathcal{O}(N\log(N))$ and can be understood as a generalization of the linear, fast Walsh-Hadamard transformation. Parameters are the choice of two (arbitrary) commutative mappings used within the $\mathbb{C}T$-transformations. In [14], different combinations of $\mathbb{C}T$-based transformations were applied on multiple scales of the TF representation and supplemented in the same feature vector with a subset of the VTLI features. In the following these features are referred to as CT features.

The third invariant feature type considered in this work is based on the class of transformations known as *generalized cyclic transformations* (GCT) [10]. The computation of features of this type involves two steps: first, the input signal is projected on basis vectors given by a matrix A, which is the matrix product of a so-called *generalized characteristic matrix* (GCM) and the transformation matrix of the modified Walsh-Hadamard transformation. The choice of coefficients for the GCM yields a high degree of freedom. In a second step, either an absolute-value spectrum or an extended group spectrum can be computed as translation-invariant features. In [13], different GCT-based features were computed from a series of sub-frames and combined in the same feature vector with a subset of the VTLI features.

The *invariant integration features* (IIFs) [15] are the fourth invariant feature type considered in this work. They are based on the idea of constructing invariant features by integrating nonlinearly transformed input signals over a finite transformation group. In the context of designated translation invariance, monomials up to a certain order are a good choice as nonlinear functions. The parameter space of the IFFs is very large and an appropriate feature selection has do be done. Experiments showed that, in contrast to the CT and GCT features, the IIFs do not need any supplementary features in order to perform best. It is shown in [15] that the IIFs can outperform MFCCs in matching as well as in mismatching scenarios.

A common disadvantage of the described invariant feature types is the high dimensionality of the resulting feature vectors. Therefore, a dimensionality reduction usually followed by a decorrelating method is applied. A detailed description of the experimental settings is given in the next section.

3 Experiments

In the first part of the experiments, six individual ASR systems were built. Each system used one of the described feature extraction methods in its front end. The acoustic and language modeling as well as the adaptation components in the back end of the system were the same for all systems. No attempt to optimize the back-end parameters, for example, word insertion probability or grammar scale factor, for the individual feature types was made.

3.1 Data and Modeling

Phoneme recognition experiments have been conducted within this work. The TIMIT corpus with a sampling rate of 16 kHz was used. The standard NIST training set consists of 3696 utterances from 462 female and male speakers. The complete test set without "SA" sentences consists of 1344 utterances from 162 speakers. Two training-test scenarios have been defined; the first includes female and male utterances in the training and test set. The second scenario simulates a mismatch in the mean VTL between training and test data. In practice, this situation arises, for example, in case of children using an ASR system that was trained only on adult utterances. Therefore, the training set of the second scenario includes only male utterances from the original training set, while the test set includes only female utterances from the complete test set. Following the standard approach [8], an initial set of 48 phonemes was used for training acoustic models. This set was collapsed to 39 phonemes for testing purposes.

In all systems triphone context-dependent *hidden Markov models* (HMMs) were trained with a left-to-right 3-state topology with no skip states. The output distributions were modeled with diagonal covariance matrices. Decision-tree clustering was applied for state-tying and a bigram language model was used. Each system applied VTLN. Here, the systems based on invariant feature types used cepstral coefficients that were based on the TF representation used by the individual feature types. In case of the MFCC- and PLP-based system, scaling was used for the warping of the frequency axis, while the invariant-feature based systems used a translational VTLN. *Speaker-adaptive training* (SAT) with CMLLR and speaker-adaptation with a combination of CMLLR and MLLR during testing were adopted.

Two systems with standard feature types were considered for the experiments. For the first system, standard MFCCs with 12 coefficients were used. The second system computed 12 *perceptual linear prediction* (PLP) coefficients. All systems used a frame length of 20 ms and a frame shift of 10 ms and appended log-energy together with first and second order derivatives to the feature vectors.

Table 1 - *Baseline phoneme error rates (PER) of individual systems.*

Front-end type	Scenario accuracy [%] matching	mismatching
MFCC (39)	24.0	30.3
PLP (39)	23.4	30.0
GCT (55)	25.1	30.4
VTLI (55)	25.0	33.2
CT (55)	27.0	31.6
IIF (60)	**22.5**	**27.4**

The settings for the individual front-ends with invariant feature types were adapted to the settings as presented in the works [15, 11, 13, 14]. The feature types that yielded the highest accuracies within the individual works were taken. In case of VTLI-, GCT- and CT-based features, a *linear discriminant analysis* (LDA) was used to reduce the dimension of the feature vectors to 55. In case of IIFs, 30 features of order one were selected. The final dimensionality of the IIF vectors after applying LDA was 60.

3.2 Baseline Error Rates and Error Analysis

The *phoneme error rates* (PER) of the individual systems for both scenarios are shown in Table 1. The upper part of the table shows the results of the systems which use the standard MFCC- and PLP-based front ends. It can be seen that the PLP-based system performs slightly better than the MFCC-based system in both scenarios. The lower part of Table 1 shows the accuracies of the systems whose front ends are based on invariant feature types. The highest accuracies, which are also higher than the ones of the PLPs, are achieved with the IIFs. In contrast, the three other invariant feature types yield accuracies that are lower than the accuracies of the non-invariant MFCC- and PLP-based systems. For the mismatching scenario, we expected that the accuracies of the invariant-feature based systems would be higher than the ones of the systems based on standard features. A reason for this shortfall may be the fact, that the back ends were not individually optimized to the feature types. Another reason is that VTLN and MLLR do a good job especially with the non-invariant features.

For a successful combination of system outputs, the ASR systems preferably use different kinds of knowledge and, thus, make different types of errors [22]. First, we analyze the substitution errors for the best performing system. Figure 1 shows a confusion matrix of substitution errors for each phoneme of the IIF-based system. Here, approximately 75% of the confusions occur within the same phoneme class. Therefore, it can be assumed that accuracy improvements within individual phoneme classes will lead to improvements of the overall performance.

In a second step, the contributions of each phoneme class c to the total PER E were analyzed for each ASR system,

$$E_c := \frac{D_c + S_c + I_c}{N} \times 100\%, \tag{1}$$

where D_c is the number of deletions, S_c is the number of substitutions, and I_c is the number of insertions within class c. Furthermore, N is the total number of phonemes within the transcription. The matching scenario was considered here. The error contributions E_c are listed in Table 2. Though lower in overall accuracy, it can be observed that the GCT, VTLI, and CT features perform equally good or slightly better within the class of strong fricatives compared

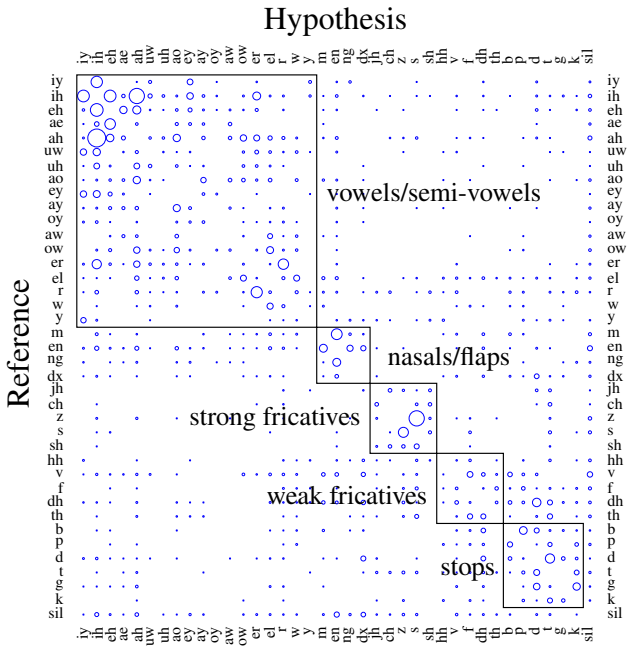

Figure 1 - *Confusion matrix of substitution errors with radii linearly proportional to the error. The largest bubble represents 4.9% of the total error.*

to MFCCs and PLPs. This observation can also be made for VTLI features in the "silence" phoneme class.

Table 3 shows the relative differences of the number of substitution errors of each feature type compared to the corresponding number of substitution error of the IIF-based system. Since the IIFs perform best in all categories, these values are all positive. Again, it can be seen that the worse performing systems (GCT-, VTLI-, CT-based) make in some phoneme classes less substitution errors than the better performing systems (PLP- and MFCC-based).

3.3 System Combination

The outputs of the individual systems have been combined with the ROVER approach: after the alignment of the 1-best hypotheses into a single phoneme transition network, a subsequent

Table 2 - *Contribution of phoneme classes to the total phoneme error rate (including insertions, deletions, and substitutions) in the matching scenario.*

	PER [%]					
Phoneme class	**MFCC**	**PLP**	**GCT**	**VTLI**	**CT**	**IIF**
vowels	12.5	12.4	13.3	13.3	14.5	12.2
nasals	2.1	2.1	2.8	2.6	3.1	1.9
strong fricatives	1.7	1.6	1.6	1.5	1.6	1.4
weak fricatives	2.4	2.2	2.3	2.4	2.4	2.1
stops	2.9	2.9	2.9	2.9	3.0	2.6
silence	2.3	2.2	2.3	2.2	2.4	2.4
$\sum$	24.0	23.4	25.1	25.0	27.0	**22.5**

Table 3 - *Relative differences of numbers of* substitution errors *of each feature type compared to the number of substitution errors of the IIFs.*

	Rel. difference to #IIF subst. errors [%]				
Phoneme class	**PLP**	**MFCC**	**GCT**	**VTLI**	**CT**
vowels	5.1	4.8	7.9	9.3	18.1
nasals	10.7	7.0	41.9	31.9	54.1
strong fricatives	21.6	15.6	8.1	5.0	11.3
weak fricatives	10.9	5.0	3.6	9.9	5.4
stops	13.1	13.1	11.6	12.1	16.5
silence	10.3	5.1	16.0	15.4	31.4

Table 4 - *Phoneme error rates (PER) of system combinations with different sizes. The systems combined with each other are marked with* •.

MFCC	**PLP**	**GCT**	VTLI	**CT**	**IIF**	PER [%] for scenario	
(24.0)	(23.4)	(25.1)	(25.0)	(27.0)	(22.5)	**matching**	**mismatching**
	•				•	22.0	27.1
		•			•	22.5	27.3
	•		•		•	20.7	26.2
			•	•	•	22.1	26.8
•	•	•			•	20.6	25.6
		•	•	•	•	21.5	26.2
•	•	•	•		•	20.4	**25.3**
•	•	•	•	•	•	**20.3**	25.8

module processes the network and selects the word with the best score [5]. Within this work, the ROVER implementation of the NIST *Scoring Toolkit* (SCTK) [16] was used.

All possible combinations of the individual systems as described above have been considered for this part of the experiments. Table 4 shows a selection of the results. This contains the combinations of size two, three, four, five, and six with highest accuracy, as well as the best combinations of different sizes when only invariant feature types are used. Generally, it can be observed that an output combination of the considered ASR systems leads to performance improvements. The combination of systems based on non-invariant features with systems based on invariant feature types yields higher performance improvements than combinations of only invariant-feature based systems. It can be observed for the matching scenario that the accuracy increases with an increasing number of systems combined with each other. For the mismatching scenario, the combination of five systems is slightly better than that of six systems. Compared to the baseline IIF system, the error rate is reduced by 2.2% in the matching and 2.1% in the mismatching case. This means a relative error rate reduction of 11% for the matching and 6% for the mismatching scenario. The error rates reported in this work are higher than the lowest reported phoneme recognition rate of 19% on TIMIT [19]. However, the ASR system used in this work does not yet include discriminative training or heteroscedastic feature transformation.

4 Conclusions

In this work, we have considered six different ASR systems that differed in the feature extraction method within the front end. Four different invariant feature types were considered in combination with standard MFCC and PLP features. The back ends of all systems were kept the same. We have shown that the (substitution) errors within certain phoneme groups made by the individual systems are differently distributed. Furthermore, the combination of 1-best hypotheses of the systems with ROVER yields improvements in performance.

In the future, we are interested in improving the accuracy of the individual systems. This could be done, for example, by fine-tuning the individual back ends to each feature type. Also using discriminative training, other dimensionality reduction methods, or more sophisticated methods for the computation of a TF representation could further increase the accuracies.

5 Acknowledgements

This work has been supported by the German Research Foundation under Grant No. ME1170/2-1. The software used for the computation of the IIFs within this work can be downloaded from `www.isip.uni-luebeck.de/downloads`.

References

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

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

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

[4] G. Evermann and P. C. Woodland. Posterior probability decoding, confidence estimation and system combination. In *Proc. Speech Transcription Workshop*, College Park, USA, 2000.

[5] J. G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proc. IEEE Int. Workshop Automatic Speech Recognition and Understanding (ASRU)*, pages 347–354, Santa Barbara, CA, USA, Dec. 1997.

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

[7] M. J. F. Gales, X. Liu, R. Sinha, P. C. Woodland, K. Yu, S. Matsoukas, T. Ng, K. Nguyen, L. Nguyen, J.-L. Gauvainy, L. Lamely, and A. Messaoudi. Speech recognition system combination for machine translation. In *Proc. Int. Conf. Audio, Speech, and Signal Processing*, Honolulu, Hawaii, Apr. 2007.

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

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

[10] V. Lohweg and D. Müller. Nonlinear generalized circular transforms for signal processing and pattern recognition. In *Proc. IEEE Workshop Nonlinear Signal and Image Processing*, Baltimore, Jun. 2001.

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

[12] 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. *J. Acoustical Society of America*, 123(5):3066–3066, Jul. 2008.

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

[14] F. Müller and A. Mertins. Nonlinear translation-invariant transformations for speaker-independent speech recognition. In J. Sole-Casals and V. Zaiats, editors, *Advances in Nonlinear Speech Processing*, volume 5933 of *LNAI*, pages 111–119, Heidelberg, Germany, Feb. 2010. Springer.

[15] F. Müller and A. Mertins. Contextual invariant-integration features for improved speaker-independent speech recognition. *Speech Communication*, 53(6):830 – 841, 2011.

[16] National Institute of Standards and Technology (NIST). Speech Recognition Scoring Toolkit (SCTK) Version 2.4.0. web resource: http://www.itl.nist.gov/iad/mig/tools, Jan. 2010.

[17] R. D. Patterson, K. Robinson, J. Holdsworth, D. McKeown, C. Zhang, and M. Allerhand. Complex sounds and auditory images. In Y. Cazals, L. Demany, and K. Horner, editors, *Auditory Physiology and Perception. Advanced Bioscience*, volume 83, pages 429–446, Pergamon, Oxford, 1992.

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

[19] T. N. Sainath, B. Ramabhadran, D. Nahamoo, D. Kanevsky, and A. Sethy. Sparse representation features for speech recognition. In *Proc. Interspeech 2010*, pages 2254–2257, Makuhari, Japan, Sept. 2010.

[20] R. Schlüter, L. Bezrukov, H. Wagner, and H. Ney. Gammatone features and feature combination for large vocabulary speech recognition. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, volume 4, Montreal, Canada, May 2007.

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

[22] S. Stüker, C. Fügen, S. Burger, and M. Wölfel. Cross-system adaptation and combination for continuous speech recognition: The influence of phoneme set and acoustic front-end. In *Proc. Interspeech'06 (ICSLP)*, pages 512–524, Pittsburgh, USA, Sept. 2006.