INTERSPEECH 2009 BRIGHTON

Invariant-integration method for robust feature extraction in speaker-independent speech recognition

Florian Müller and Alfred Mertins

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

Abstract

The vocal tract length (VTL) is one of the variabilities that speaker-independent automatic speech recognition (ASR) systems encounter. Standard methods to compensate for the effects of different VTLs within the processing stages of the ASR systems often have a high computational effort. By using an appropriate warping scheme for the frequency centers of the time-frequency analysis, a change in VTL can be approximately described by a translation in the subband-index space. We present a new type of features that is based on the principle of invariant integration, and an according feature selection method is described. ASR experiments show the increased robustness of the proposed features in comparison to standard MFCCs.

Index Terms: speech recognition, speaker-independency, invariant integration, monomials

1. Introduction

The vocal tract is a fundamental component of the human speech production system. The gender and age of the individual speakers are two factors that determine the average vocal tract length (VTL) [1]. Besides the vocal tract's shape it is the VTL that determines the location of the resonance frequencies, also known as "formants". The ratio between the VTLs of any two speakers A and B is called "warping factor" (denoted as α here). On a short-time basis the magnitude spectra of the speakers can approximately be related by $S_A(\omega) = S_B(\alpha \cdot \omega)$. In a typical speaker-independent automatic speech recognition (ASR) task the value of α is between 0.8 and 1.2.

This intrinsic variability has a negative effect on the recognition rate of speaker-independent ASR systems. Methods that try to compensate this have become a standard component of today's ASR systems and different types of methods have been proposed. One group of techniques tries to adapt the acoustic models of the recognition system to the individual speakers, e.g, [2]. Other methods try to normalize the spectral effects of different VTLs at the feature extraction stage [3, 4]. The mentioned methods have the drawback that, in general, they have a high computational effort. Thus, a third group of methods tries to generate features that are independent of the warping factor [5, 6, 7].

Methods that extract a time-frequency (TF) representation of an input signal for ASR tasks commonly locate the frequency centers of the analysis filters on auditory motivated scales like the Mel- or ERB-scale. Using these scales, it was shown that VTL changes approximately lead to translations in the subband-index space of these TF representations [7, 8]. This can be utilized for the computation of features that are invariant to translation [6, 7]. The invariance can lead to an increase of robustness against VTL changes.

The determination of invariants is well-founded in the field of mathematics and physics. Practical methods for the retrieval of invariants against rotation and translation were especially applied in the field of pattern recognition. One of these general methods integrates regular nonlinear functions of the features over the transformation group for which an invariance should be achieved. This method is commonly known as "invariant integration". Other methods include translation invariant transformations, such as the modulus of the discrete Fourier transformation, the auto-/cross-correlation function or the (modified) rapid transformation [9]. In this paper the method of invariant integration is used for the computation of features that are robust against VTL changes. Because this method leads to an undesired large set of possible features, an appropriate feature selection method is described. It will be shown in large-vocabulary phoneme recognition experiments that the resulting feature sets lead to better recognition results than the standard mel-frequency cepstral coefficients (MFCC) under matching training and testing conditions and that the proposed features outperform the MFCCs in cases in which training and testing conditions differ with respect to the mean VTL.

The paper is organized as follows. The next section introduces in its first part the method of invariant integration and describes how it can be used to compute features that are robust against VTL changes. An adaption of an iterative feature selection method [10] is described in the second part of Section 2. The experimental setup and results are described Section 3. Discussion with subsequent conclusions follow in the last section.

2. Methods

2.1. Features by invariant integration

In the following, we briefly introduce the basic concepts of invariant integration. A detailed description and applications of this concept can be found in [11, 12].

The features presented in the following are derived from a TF representation $y(n, k)$ of an input signal x, where $1 \leq n \leq N$ is the time index and $1 \leq k \leq K$ is the subband-index. The vector $\boldsymbol{v} = (v_1, v_2, \ldots, v_K) = y(\hat{n}, k)$, $k = 1, 2, \ldots, K$, containing all spectral values for any time index $\hat{n}$ is called "frame". The corresponding frequency centers of the subbands are assumed to be equally distributed on an auditory motivated frequency scale (such as the ERB-scale) that maps the spectral effects of VTL changes to translations in the subband-index space. Thus, the relation between two given representations y_A and y_B of the same utterance, but with different

VTLs can then be described framewise by

$$v_k^A = v_{k+t}^B, \tag{1}$$

where $t \in \mathbb{Z}$ is the shift. The scale factor that corresponds to a translation by one depends on the properties of the TF representation. For example, if we take 90 samples in the log-warped domain corresponding to frequencies between 50 and 6700 Hz, then each sample translation corresponds to a change of the warping factor value by about 0.05. Thus, a warping factor α with range from 0.8 to 1.2 corresponds to translations in the range between -4 and +4 sample translations.

Following the notion of [11] this translation effect can be traced back to the action of a group G on the input signal space S. In our case G is the group of translations. Since we have a finite number of subbands in the TF representation all indices are understood modulo K. These periodic boundary conditions are used throughout the following theoretical explanations. In practice, however, we set $v_k = 0 \quad \forall k \notin \{1, \ldots, K\}$. We say that two frames $\boldsymbol{v}, \boldsymbol{w}$ contain the same information relevant to classification (i.e., $\boldsymbol{v}$ and $\boldsymbol{w}$ are "equivalent") if there exists an element $g \in G$ with $\boldsymbol{v} = g\boldsymbol{w}$. Now, a feature is given by a map $A : S \to F$ which maps for a given $\boldsymbol{v} \in S$ the set $\{\, g\boldsymbol{v} \mid g \in G \,\}$ onto a single point in the feature space F. This means that a feature is invariant with respect to the action of the transformation group on the input signals, i.e.,

$$A(g\boldsymbol{v}) = A(\boldsymbol{v}) \;\; \forall g \in G. \tag{2}$$

Two input signals $\boldsymbol{v}, \boldsymbol{w} \in S$ that are not equivalent, that is $\boldsymbol{v} \neq g\boldsymbol{w}$ for all $g \in G$, should be mapped into different points in the feature space F.

Therefore, features describe properties which are common among all equivalent input signals. This gives reason to compute invariant features by an appropriate averaging. Hurwitz invented the principle of integrating over the transformation group for constructing invariant features in 1897 [13]. A group average for G as a finite group of order $|G|$ is given by

$$A_f(\boldsymbol{v}) := \frac{1}{|G|} \sum_{g \in G} f(g\boldsymbol{v}), \tag{3}$$

where f is a given regular function on S. The question arises how to define the function f. Noether's theorem [14] can be used to construct a basis that spans the invariant feature space F by computing the group averages with choosing f out of the set of the monomials m of $\boldsymbol{v}$,

$$m(\boldsymbol{v}; i) := \prod_{k=1}^{K} v_{k+i}^{b_k}, \tag{4}$$

where $b_k \in \mathbb{N}_0$ and $i \in \mathbb{Z}$ as a translation parameter used in the following. The order of a monomial is defined as the sum of its exponents. Noether showed that for input signals of dimensionality K and finite groups with $|G|$ elements, the group averages of monomials with degree less or equal $|G|$ form a generating system of the pattern space. Such a basis has at most $\binom{|G|+K}{K}$ elements. It has to be pointed out that this is an upper bound. It was shown that in many applications the number of needed functions is considerably smaller [11, 15, 16]. We assume that the translations that occur as effects of VTL changes in the subband-index space only stretch across a limited interval within the subband-index space. This further restricts the number of elements within the considered transformation group. Now, an "invariant integration feature" (IIF), as a group average on the basis of monomials, can be defined as

$$A(\boldsymbol{v}) := \frac{1}{2W+1} \sum_{i=-W}^{W} m(\boldsymbol{v}; i), \tag{5}$$

with $W \in \mathbb{N}_0$ being the "window size". As an example, we consider the monomial $m(\boldsymbol{v}; i)$ of order 2 with $b_k = 0$ for all $k \in \{1, 2, \ldots, K\} \backslash \{k_1, k_2\}$ and $b_k = 1$ for $k \in \{k_1, k_2\}$. Note, that k corresponds to the subband indices of the underlying spectro-temporal representation. The corresponding group average with window size $W = 1$ (cf. (5)) is then given by

$$A(\boldsymbol{v}) = \frac{1}{3} \left(v_{k_1-1} v_{k_2-1} + v_{k_1} v_{k_2} + v_{k_1+1} v_{k_2+1} \right). \tag{6}$$

2.2. Feature selection

According to (4) and (5) the parameters to be defined for one feature are the exponents b_k and the window size W. Because of the large number of possible parameter combinations, an appropriate feature selection method has to be used. The method should be adaptable with respect to its selection criteria and should work efficiently due to the high amount of data. In this work a modified version of the feature-finding neural network (FFNN) [10] is used. This approach was successfully applied in other fields of speech recognition [10, 17]. The feature selection process of the FFNN works iteratively with a single-layer perceptron as its basis. The capability of a fast training of the linear classifier is crucial for this method. It consists of the following steps S1–S4:

- S1) Start with a set of M features which are randomly chosen.
- S2) Use the linear classifier for computing the relevance of each feature.
- S3) Remove the feature with the least relevance.
- S4) If a stopping criterion (e.g., the total number of iterations) is not fulfilled, add a new, randomly generated feature to the current feature set and go back to step S2), otherwise stop.

The linear classifier performs frame-wise phoneme classification experiments in each iteration. In the basic version of the algorithm the relevance of each feature i is based on the root mean square (RMS) classification error of the classifier when feature i is left out of the feature set using the whole training and testing data. The feature with the least increase of RMS after removal is assumed to be the least relevant one in the current feature set. Note that this approach gives an implicit order for the features which will be taken up in the experiments later on. Since the proposed IIFs should be robust against VTL changes the basic computation of the relevance is accordingly selected in our work: The considered features defined in (5) are translation-invariant. Now, we seek a feature set that has a high degree of separability with respect to the phonetic classes. Therefore, we consider three training-testing scenarios which differ by the mean VTL and decide for features with a high degree of relevance in all three scenarios. Because of the translational effect of VTL changes in the subband-index space and the translation invariance of the considered features, we assume that such a feature set exists.

First, the training and testing data were divided into female and male utterances. Then the RMS errors of three training-testing scenarios were computed. These scenarios were training

on both male and female data (FM-FM), training on female and testing on male data (F-M), and training on male and testing on female data (M-F). For each feature, the maximum of the three RMS errors was taken into account in steps S2 and S3.

3. Experimental setup and results

As data base, the TIMIT corpus was used with a sampling rate of 16 kHz. The "SA" sentences have not been used to avoid an unfair bias for certain phonemes [18]. We chose to use a complex-valued Gammatone filterbank with 90 filters equally-spaced on the ERB scale as TF analysis method. This setup was chosen to allow for a comparison with previous works (cf. [6]) and other choices for K could also be taken here. The magnitudes of the subband signals were lowpass filtered in order to decrease the time resolution to $20\,\mathrm{ms}$. These filtered magnitudes were then subsampled to obtain a final frame rate of the time-frequency representations of one frame every $10\,\mathrm{ms}$. 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.

3.1. Feature selection

The adapted feature selection method as described in Section 2.2 was performed with the TIMIT data on a frame-by-frame basis with 48 phonetic classes. Because of the large amount of data, only every 10th frame of the filterbank results was taken for feature selection. In order to allow for a systematic analysis of the presented IIFs, we constricted the parameter space for the generation of randomly chosen features. Besides the size of the feature set, the parameters that had to be optimized were the exponents b_k and the window size W for each IIF of the set (cf. (4) and (5)). As indicated in Section 2.1, Noether's theorem [14] gives an upper bound for the maximum order of the monomials that are needed to construct a complete feature set. This upper bound is given by the number of elements within the considered finite transformation group which can be related to the maximum number of translations that can occur when any two speakers with different VTLs are observed. With respect to the parameters of the chosen TF analysis this number was assumed to be 6 in our experiments, and we examined feature sets of five categories. The maximum order of the monomials of the features in each category were 2, 3, 4, 5 and 6, respectively. The integration over the whole frame corresponds to setting $W = \lfloor K/2 \rfloor$. Thus, the randomly chosen window size for each feature was constrained to $0 \leq W \leq \lfloor K/2 \rfloor$.

The size M of the target feature set was chosen as 90 features. Initially selecting this large number of features together with the order given by the adapted relevance criterion allows an analysis of subsets consisting of only a certain number of the most-relevant features. These results will be shown in Section 3.2. However, M could be chosen even larger, which should go along with an increased number of feature selection iterations. Here, as stopping criterion the total number of 750 iterations was chosen. Examining the mean classification result of the linear classifier during the feature selection processes for the five categories has led to the results shown in Figure 1 in our experiments. It can be seen that the increase of the mean classification rate converges to zero. Thus, a total number of 750 feature selection iterations seems to be adequate within this setup. We further noticed that the accuracy rates of feature sets with different random starts did not vary much. A deeper analysis will be part of further work.

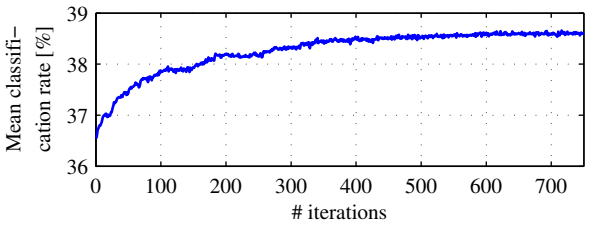

Figure 1: *Development of mean classification rate of the linear classifier during the feature selection process.*

Table 1: *Accuracies of experiments. N IIF(O) denotes the N most-relevant invariant integration features of the set with maximum order of O.*

Feature set	FM-FM	M-F	F-M
MFCC	66.57	55.00	52.42
90 IIF(5)	66.15	62.31	61.95
70 IIF(4)	66.65	61.87	61.50
20 IIF(2)	66.46	60.53	59.08
90 IIF(5)+ACF+CCF	65.76	62.03	61.30
70 IIF(4)+ACF+CCF	65.99	61.63	61.00
20 IIF(2)+ACF+CCF	65.48	60.18	59.71
90 IIF(5)+MFCC	65.95	61.72	62.13
70 IIF(4)+MFCC	66.13	61.88	61.89
20 IIF(2)+MFCC	66.35	60.81	59.90

3.2. Speech recognition with invariant integration features

For the evaluation of the feature sets found by the feature selection method, phoneme recognition experiments have been conducted. To simulate diversity between training and testing conditions, the training and testing data were each split into male and female subsets. These subsets were combined as described above to the three scenarios FM-FM, M-F, and F-M. According to [18], 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). Monophone models with three states per phoneme, 8 Gaussian mixtures per state and diagonal covariance matrices were used together with bigram statistics. MFCCs were used to obtain baseline recognition accuracies. The MFCCs were calculated by using the standard HTK setup which yields 12 coefficients for each frame.

Different combinations of previously presented VTL invariant features [6] together with the IIFs have also been considered. These features are based on the discrete cosine transform (DCT) of the cross-correlation sequence of logarithmized spectral values (CCF) and on the DCT of the logarithmized auto-correlation sequence of spectral values (ACF). Furthermore, the described feature sets have been combined with MFCCs.

All feature sets were amended by the logarithmized energy of the original frames together with delta and delta-delta coefficients. A linear discriminant analysis (LDA) was performed after the computation of the features. The reduction matrices of the LDA were based on the 48 phonetic classes contained in both, the male and female utterances. The recognition accuracies of the MFCCs and notable results of the described feature sets are summarized in Table 1 and are described in the following. The feature sets with IIFs of order 3 and 6 did not have remarkably different properties and are not further described.

The first experiment with the selected feature sets investigated the dependency between the number of the most-relevant features of each set and the recognition accuracy. Overall, the recognition accuracies of the five categories of different orders were very similar. The experiments have shown that a feature set consisting of at least 20 IIFs results in accuracies that are comparable to those of the MFCCs in the FM-FM scenario and that are superior to MFCCs in the M-F and F-M scenarios. Furthermore, the results yielded the following:

- The best mean accuracy in all three scenarios was obtained when using all 90 features of the set with a maximum order of 5. This choice yields the following accuracies for the three scenarios: FM-FM: 66.15%, M-F: 62.31% and F-M: 61.95%. Though slightly lower than the MFCCs in the FM-FM case, this choice shows an increase of accuracy of about 7% and 10% in the two scenarios M-F and F-M, respectively.
- The highest result in the FM-FM scenario was achieved with the 70 most-relevant features of the set with a maximum order of 4. Here, the accuracy is 66.65% in the FM-FM scenario, and it is 61.87% and 61.5% in the M-F and F-M cases, respectively. It is notable that the FM-FM accuracy is higher than the one of the MFCCs.
- The feature set with the highest accuracy in relation to its size consists of the 20 most-relevant features with monomial order less or equal 2, which results in the accuracies FM-FM: 66.46%, M-F: 60.53% and F-M: 59.08%.

The second experiment combined the IIFs with the ACF and CCF features from [6]. In this previous work the combination of different invariant feature types lead to an increase of robustness against the effects of VTL changes. The results as shown in Table 1 indicate that the combination of ACF, CCF and IIFs does not yield any improvements. The third experiment amended the MFCCs to the previously used invariant feature sets. This has been necessary to boost the performance with the method in [6]. The results show no significant enhancement of accuracy when combining the IIFs with MFCCs here.

An implementation of the VTLN technique described in [4] leads to the following accuracies: FM-FM: 68.61%, M-F: 64.02%, F-M: 63.39%. Clearly, the VTLN method outperforms the IIFs by about two percent. The drawback, however, is the much higher computational complexity of the VTLN method which involves the feature computation and decoding for several warped versions of one and the same utterance.

4. Discussion and Conclusions

We have presented a method for feature extraction that is based on the theory of invariant integration. The derived "invariant integration features" exploit the translational effect of VTL changes in the subband-index space of certain TF representations. A feature-selection approach was described that finds a feature set under given constraints. Phoneme-recognition experiments have shown a superior performance of the invariant integration features in comparison to the MFCCs especially in diverse training-testing conditions with resepct to the mean VTL.

Further investigations will be directed toward the incorporation of adjoining frames in the definition of the invariant integration features. Additionally, recognition experiments with corpora containing adult and children speech will give further insights. Also, the robustness against noise has to be examined as well as possibilities that improve the feature selection process. Supplementary material of the described experiments can be found at http://www.isip.uni-luebeck.de/index.php?id=481.

5. Acknowledgments

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

6. References

[1] L.-J. Boë, J. Granat, P. Badin, D. Autesserre, D. Pochic, N. Zga, N. Henrich, and L. Ménard, "Skull and vocal tract growth from newborn to adult," *Proceedings of the 7th ISSP*, Dec. 2006.

[2] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, Apr. 1995.

[3] 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.

[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] S. Umesh, L. Cohen, N. Marinovic, and D. J. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 40–45, Jan. 1999.

[6] J. Rademacher, M. Waechter, and A. Mertins, "Improved warping-invariant features for automatic speech recognition," *Proc. Int. Conf. Spoken Language Processing*, pp. 1499–1502, Sept. 2006.

[7] 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. Acoust. Soc. Am.*, vol. 123, no. 5, pp. 3066–3066, Jul. 2008.

[8] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Frequency-warping in speech," *Proc. Int. Conf. Spoken Language Processing*, vol. 1, pp. 414–417, 1996.

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

[10] T. Gramss, "Word recognition with the feature finding neural network (FFNN)," *Proc. IEEE Int. Workshop Neural Networks for Signal Processing*, pp. 289–298, Oct. 1991.

[11] H. Schulz-Mirbach, "On the existence of complete invariant feature spaces in pattern recognition," *Proc. Int. Conf. Pattern Recognition*, vol. 2, pp. 178–182, Aug. 1992.

[12] ——, "Algorithms for the construction of invariant features," in *Mustererkennung 1994, 16. DAGM Symposium*, no. 5. Wien: W. G. Kropatsch and H. Bischof, 1994, pp. 324–332.

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

[14] E. Noether, "Der Endlichkeitssatz der Invarianten endlicher Gruppen," *Mathematische Annalen*, vol. 77, no. 1, pp. 89–92, Mar. 1915.

[15] H. Schulz-Mirbach, "Invariant features for gray scale images," in *Mustererkennung 1995, 17. DAGM-Symposium*. London, UK: Springer-Verlag, 1995, pp. 1–14.

[16] S. Siggelkow, "Feature histograms for content-based image retrieval," Ph.D. dissertation, Albert-Ludwigs-Universität Freiburg, Breisgau, 2002.

[17] M. Kleinschmidt and D. Gelbart, "Improving word accuracy with Gabor feature extraction," *Proc. Int. Conf. Spoken Language Processing*, pp. 25–28, 2002.

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