FEATURE EXTRACTION WITH A MULTISCALE MODULATION ANALYSIS FOR ROBUST AUTOMATIC SPEECH RECOGNITION

Florian Müller and Alfred Mertins

Institute for Signal Processing
University of Lübeck
Ratzeburger Allee 160, 23562 Lübeck, Germany

ABSTRACT

In this work we present a new feature extraction method that is robust against the effects of varying vocal tract lengths. The principle of the method is based on invariant integration and makes use of a modulation filtering approach, similar to the recently proposed scattering transform. In particular, we show how the transform can be used to obtain features that are robust against variations of the vocal tract length. Phoneme recognition experiments show a clearly increased robustness in case of mismatching average vocal tract lengths.

Index Terms— Automatic speech recognition, feature extraction, robustness, speaker-independence

1. INTRODUCTION

Systems for automatic speech recognition (ASR) have to cope with various sources of variability [1]. Among the inter-speaker variabilities, the vocal tract length (VTL) is one major factor, which has led to different vocal tract length normalization (VTLN) techniques, with frequency warping during the feature extraction being one of these [2, 3]. Another common method to compensate for the effects of different VTLs tries to adapt the acoustic models of the recognition system to the speaker-specific characteristics. Maximum-likelihood linear regression (MLLR) [4] is a typically used method for this approach. Instead of normalizing the features or adapting the acoustic models, a third approach is to extract features that are independent of the effects of different VTLs. Based on an acoustic tube model, the dependency between the resonance frequencies F_i and length l of the tube is given by $F_i = \frac{c}{4l}(2i-1),\ i = 1, 2, 3, \ldots$, where c is the speed of sound [5]. Accordingly, it follows that the spectra S_A, S_B from two speakers A and B that utter the same phone are related by a frequency warping of the form

$$S_A(\omega) = S_B(\alpha\omega). \tag{1}$$

While VTLN methods typically estimate the warping factor α, invariant feature extraction methods try to compute features that yield the same feature values for different α. As an example, the work in [6] made use of the scale transform [7], which leads to coefficient magnitudes that are independent of the warping factor.

Auditory filterbanks that are commonly used within the field of ASR locate the center frequencies of their filters linearly spaced along (quasi-) logarithmic frequency scales. The relation from (1) then becomes

$$S_A(\log\omega) = S_B(\log\alpha + \log\omega), \tag{2}$$

meaning that frequency warping results in a translation along the log-frequency axis. Various feature extraction methods that are invariant to translations were proposed in the recent years [8, 9, 10, 11]. Among these, the invariant-integration features (IIF) [11] have shown superior robustness against the effects of varying VTLs while keeping highest accuracies in ASR. For a given transformation group, the central idea of invariant integration is to compute group averages of nonlinear functions of the input data. For a finite group, as considered in case of the IIFs, it was shown that the use of monomials as nonlinear functions up to a certain order leads to high accuracies in matching as well as in mismatching training-test conditions with respect to the average VTL in each dataset. In general, arbitrary (possibly) nonlinear functions can be used instead of monomials to achieve invariance to the considered transformation group. The separability of the classes of interest, however, has to be evaluated for each specific application.

Recently, the scattering transform has been introduced, and a link to MFCCs has been provided [12]. Given an input signal, the scattering transform computes co-occurrence coefficients for multiple scales that result from cascaded filterbanks. The process is quite similar to a cascaded computation of modulation spectra as in [13], but uses different filters. After a lowpass filtering of modulation spectra, one obtains coefficients (features) that are translation-invariant up to a certain degree. Under certain conditions on the filterbank, it becomes possible to fully recover the input signal from the coefficients [12], at least up to a global translation.

This work has been supported by the German Research Foundation under Grant No. ME1170/4-1.

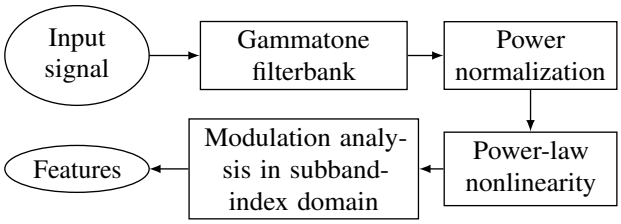

Fig. 1. Processing scheme for feature computation.

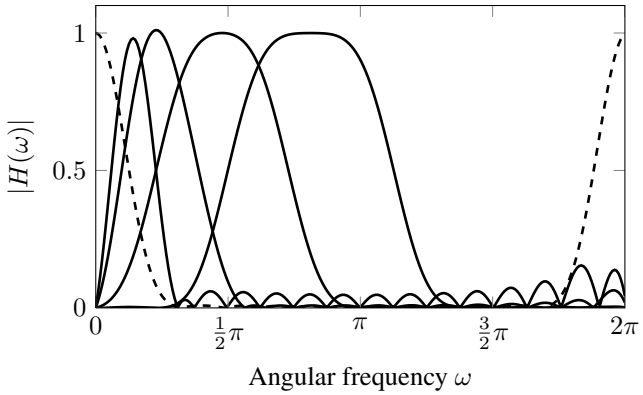

Fig. 2. Magnitudes of the lowpass (dashed) and bandpass (solid) filters of the frequency responses of $h_j(k)$, $j = 0, \ldots, 4$.

In this work, we present a method that follows the idea of invariant integration to achieve vocal tract length invariance. In contrast to previous works that used monomials, a modulation filtering approach, similar to the scattering transform, but without the strict derivation of filters from a mother wavelet as in [12], is used. Different from [12, 13], the modulation filtering is not carried out with respect to time, but with respect to the subband index in a gammatone filterbank analysis. In the following section, the feature extraction is described. We show how the scattering transform fits into the scheme of invariant integration and discuss implementation aspects. The proposed method is evaluated in Section 3. Conclusions and an outlook on future work are given in the last section.

2. MULTISCALE MODULATION ANALYSIS FOR ROBUST FEATURE EXTRACTION

2.1. Overall Extraction Scheme

The proposed feature extraction method makes use of a repeated modulation-filter analysis (similar to the scattering transform) to compute features that are robust against the effects of different VTLs. An overview of the proposed feature extraction approach that we call "scale-translation invariant features" (STIF) in the following is given in Fig. 1. As done for the extraction of IIFs, the first step for the computation of the features involves a time-frequency (TF) analysis. This is done with a gammatone filterbank in this work, with center frequencies of the filters equally spaced on the equivalent rectangular bandwidth (ERB) scale [14].

Let $s_n(k)$ denote the magnitude of the gammatone-filter based TF representation of a speech signal, where n is the time index with $1 \leq n \leq N$, and k is the subband index with $1 \leq k \leq K$. Since the ERB scale is almost logarithmic, the translation in (2) approximately results in a shift of the subband indices: $s_n(k) \rightarrow s_n(k + k_\alpha)$ with some k_α depending on the warping factor α and k_α not necessarily being an integer. Thus, the spectral effects due to different VTLs can be modelled as a finite group $\mathcal{G}$ of translations along the subband index space of a time-frequency analysis with logarithmically spaced filters, which is nearly the case for the above mentioned gammatone filterbank. Now let a frame for a time index n be given by $\boldsymbol{s}_n = (s_n(1), s_n(2), \ldots, s_n(K))$. According to the principle of invariant integration [15], features that are invariant with respect to the group $\mathcal{G}$ can be computed with

$$T_f(\boldsymbol{s}_n) = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} f\left(g\boldsymbol{s}_n\right), \tag{3}$$

where $f(\cdot)$ is a kernel function with the frames of the TF representation as input. One way for obtaining a transformation

$$T\left(\boldsymbol{s}_n\right) = \left(T_{f_1}(\boldsymbol{s}_n), T_{f_2}(\boldsymbol{s}_n), \ldots, T_{f_F}(\boldsymbol{s}_n)\right)^\top \tag{4}$$

that is complete in the sense that it is only invariant to the group $\mathcal{G}$ and no other operation, is to consider only monomials for the kernel functions [15]. This choice was made for the IIFs in [11]. The modulation analysis considered in this work can be seen as another kind of nonlinear kernel function that is used instead of monomials.

2.2. Cascaded Modulation Analysis

In this work, we propose to use a set of bandpass filters combined with the modulus operator as kernel functions. The integration over the group $\mathcal{G}$ from (3) is realized with a finite impulse response (FIR) lowpass filter $h_0(k)$ of length 16. We designed $J = 4$ complex, nearly analytic FIR bandpass filters $h_j(k)$, $1 \leq j \leq J$. The bandpass filters approximately have a constant-Q characteristic, as far as the short filter length and the fact that filters are supposed to be analytic permits. Furthermore, all bandpass filters have a zero mean. The magnitudes of the frequency responses are shown in Fig. 2.

After normalization and power-law compression, the values of each frame are processed by bandpass filters, the nonlinear modulus operator, and the lowpass filter, as illustrated in Fig. 3. Importantly, these filter operations are carried out with respect to the subband index k, and not the time

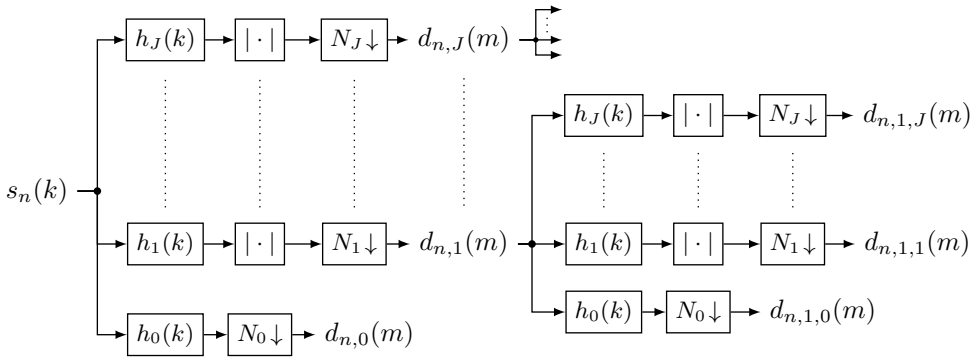

Fig. 3. Two stages of the processing scheme.

index n. In each stage of the cascade, the input signal is passed to a lowpass filter $h_0(k)$ as well as to J bandpass filters $h_1(k), \ldots, h_J(k)$. The lowpass output is downsampled by a factor N_0 leading to the signal $d_{n,0}(m)$. It forms the first part of the feature vector. The signals that were filtered with the bandpass filters $h_j(k)$ are passed to the modulus operator and downsampled by a factor N_j afterwards, which leads to the signals $d_{n,1}(m), \ldots, d_{n,J}(m)$. First-order scattering coefficients are computed by lowpass filtering and downsampling these signals and by concatenating the resulting signals to the feature vector. Higher-order scattering coefficients are obtained by iterating this scheme: For example, second-order coefficients are obtained by passing each of the signals $d_{n,1}(m), \ldots, d_{n,J}(m)$ to the filters $h_1(k), \ldots, h_J(k)$, followed by applying the modulus operator and by downsampling and so on. As mentioned in [12] and proved in [16], the energy of the scattering coefficients of order q decreases to zero as q increases if the filterbank has a certain contractive behavior. However, for the filters used in this work, this property does not exactly hold.

The number of extracted features for a given input signal depends on the length of the input signal, the lengths of the filters, the number of stages, the downsampling rate between the individual stages of the transform, and the selected subset of coefficients. Given 26 spectral input coefficients for a given time frame, in our implementation, the dimensionality of the feature vectors ranges between 30 and 270 features, depending on the scattering order and downsampling factors. In order to reduce the dimensionality and to better fulfill the assumptions made by diagonal covariance modeling in the recognizer, the resulting features together with the corresponding dynamic features were passed to a linear discriminant analysis (LDA) followed by a maximum-likelihood linear transform (MLLT) [17]. Since a specific selection of individual features has proved beneficial in comparison to a linear transform like the LDA in various application (see, e.g., [11, 18]), we also evaluated the performance of the proposed features in combination with the general feature selection approach that was used in [11].

3. EXPERIMENTS

3.1. Data and Setup

In a first step, experiments were conducted on the TIMIT corpus with a sampling rate of 16 kHz. The SA sentences were excluded. The training and test sets consisted of 3696 and 1344 utterances from female and male speakers, respectively. To evaluate the robustness of the proposed features against mismatching average vocal tract lengths in training and test data, two training-test scenarios were defined: The matching scenario used the standard training and test sets, which contain female as well as male utterances. The mismatching scenario used only the utterances from male adults of the original training set and only the utterances from female adults of the original test set for training and testing, respectively.

For comparison, standard MFCC features were computed which consist of 12 cepstral coefficients concatenated with the log-energy feature and the corresponding dynamic features, so that, overall, MFCC feature vectors consisted of 39 components in total. Furthermore, 30 IIFs were computed with parameters that led to highest accuracies in [11]. For another baseline, cepstral coefficients were computed based on a 26-channel gammatone filterbank with a minimum center frequency of 100 Hz and a maximum center frequency of 7800 Hz. The frame length was set to 20 ms, and a frame shift of 10 ms was chosen. Moreover, a power-function nonlinearity with an exponent of 0.1 was applied on the spectral values in order to resemble the nonlinear compression found in the human auditory system. This filterbank was also used for the extraction of the features that are based on the nonlinear feature transform as proposed in this work.

Phone recognition experiments were conducted with a recognizer that is based on HTK [19]. State-clustered, cross-word triphone models with diagonal covariance modeling were used. The acoustic models had three emitting states with a left-to-right topology. Depending on the amount of available training data, up to 16 mixtures were used for the Gaussian mixture models. A bigram language model based on the TIMIT training data was used. For the evaluation of the recognition ac-

Table 1. Baseline accuracies [%] for mel frequency cepstral coefficients (MFCC), gammatone cepstral coefficients (GTCC) and invariant-integration features (IIF) on TIMIT for matching and mismatching training-test conditions.

Features	matching	mismatching
MFCC	73.2	56.0
GTCC	73.4	54.3
IIF	75.5	61.4

Table 2. Accuracies [%] for features based on scattering transform for the TIMIT corpus with the maximum scale indicated as subscript.

Features	matching	mismatching
STIF_1	72.3	55.4
STIF_2	71.4	58.7
STIF_3	70.6	61.6
$\text{STIF}_{3,\text{selected}}$	73.2	65.5

curacies, the final recognition results were folded to 39 final classes [20]. To concentrate on the properties of the feature vectors, no VTLN or MLLR were applied within the ASR systems in this work.

For the invariant feature types, an LDA followed by an MLLT was used to reduce the number of dimensions to 55 and to decorrelate the features.

3.2. Baseline Accuracies

The baseline accuracies with the described ASR system are shown in Table 1. It can be seen that the commonly used MFCC features yield a similar performance in case of matching training-test conditions compared to gammatone cepstral coefficients (GTCC). For the mismatching training-test scenario, in which the average VTL is different in the training and the test set, the accuracies of both MFCC and GTCC decrease by around 18 percentage points with the MFCCs having a slightly higher accuracy. In the matching as well as the mismatching training-test scenario, the IIFs yield a higher accuracy than the MFCC and GTCC features. While the accuracy is about two percentage points higher for the matching scenario, for the mismatching scenario the IIFs show a much higher robustness against varying VTLs and perform around five percentage points better than MFCCs.

3.3. Experiments with Extracted Features

We considered scattering coefficients of up to third order within our experiments. In the first part of the experiments, we computed STIF coefficients as described in Section 2.2 and added information about the temporal context by means of conventional delta features as commonly used. The resulting accuracies for these features are shown in the first three lines of Table 2. With respect to the matching scenario, which consists of the standard training and test sets of TIMIT, it can be seen that the accuracy of the STIF coefficients with maximum order of one yield the highest accuracy in comparison to the STIF coefficients of higher order. When looking at the accuracies for the mismatching scenario, however, the accuracies increase with increasing maximum coefficient order. While the accuracy for coefficients up to order one is comparable to that of the MFCC and GTCC features, the STIF coefficients up to order three achieve an accuracy that is similar to the one of the IIFs.

In the second part of the experiments we used the feature selection method as used in [11] to select STIF coefficients from within a temporal window of length 130 ms as final feature vector. 30 STIF coefficients up to order three were considered during the selection, amended by delta features, and reduced to 55 features by an LDA. The results for this experiments are shown in the last line of Table 2. For both training-test scenarios, an increase in accuracy can be observed. While the accuracy for the matching scenario reaches that of the MFCC and GTCC features, the STIF coefficients with maximum order of three show an accuracy for the mismatching scenario that is even more than four percentage points higher than the corresponding accuracy for IIFs. This shows that the proposed feature extraction process yields features that are extremely robust against mismatches in vocal tract length.

4. CONCLUSIONS

In this work we have proposed a new method for the extraction of robust features with respect to the VTL as variability. The method is based on the idea of invariant integration and makes use of a scattering operator to describe the spectral characteristics of individual phonemes. Two training-testing scenarios for phone recognition on the TIMIT corpus were considered to evaluate the robustness of the features for matching as well as for mismatching average VTLs. The experiments showed that for a matching scenario, the features achieve accuracies that are comparable to those of MFCC and GTCC features. The robustness of the proposed extraction method against the effects of varying VTLs is clearly shown with the mismatching scenario. The resulting accuracy of the proposed features is about ten percentage points higher than for MFCCs and four percentage points higher than for the recently presented IIFs. Future work will further investigate the optimization of the filterbank. Furthermore, the performance of the proposed method on larger tasks and in combination with normalization and adaptation methods will be investigated.

5. 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*, vol. 49, no. 10-11, pp. 763–786, Oct.-Nov. 2007.

[2] Li Lee and Richard Rose, "Speaker normalization using efficient frequency warping procedures," in *Proc. Int. Conf. Audio, Signal, and Speech Processing*, Atlanta, USA, May 1996, vol. 1, pp. 353–356.

[3] L. Welling, R. Haeb-Umbach, X. Aubert, and N. Haberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Seattle, USA, May 1998, vol. 2, pp. 797–800.

[4] 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.

[5] J. R. Deller, J. G. Proakis, and J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan, New York, 1993.

[6] S. Umesh, Leon Cohen, Nenad Marinovic, and Douglas J. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 1, pp. 40–45, Jan. 1999.

[7] L. Cohen, "The scale representation," *IEEE Trans. Signal Processing*, vol. 41, no. 12, pp. 3275–3292, Dec. 1993.

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

[9] Florian Müller, Eugene Belilovsky, and Alfred Mertins, "Generalized cyclic transformations in speaker-independent speech recognition," in *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Merano, Italy, Dec. 2009, pp. 211–215.

[10] Florian Müller and Alfred Mertins, "Robust features for speaker-independent speech recognition based on a certain class of translation-invariant transformations," in *Advances in Nonlinear Speech Processing*, Jordi Sole-Casals and Vladimir Zaiats, Eds., Heidelberg, Germany, Feb. 2010, vol. 5933 of *LNAI*, pp. 111–119, Springer.

[11] Florian Müller and Alfred Mertins, "Contextual invariant-integration features for improved speaker-independent speech recognition," *Speech Communication*, vol. 53, no. 6, pp. 830 – 841, 2011.

[12] J. Andén and S. Mallat, "Multiscale scattering for audio classification," in *Proc. of ISMIR Conference*, Miami, Florida, USA, Oct. 2011.

[13] Torsten Dau, Birger Kollmeier, and Armin Kohlrausch, "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," *Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 2892–2905, 1997.

[14] B.C.J. Moore and B.R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *Journal of the Acoustical Society of America*, vol. 74, pp. 750–753, 1983.

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

[16] S. Mallat, "Group invariant scattering," *Communications in Pure and Applied Mathematics*, vol. 65, no. 10, pp. 1331–1398, Oct. 2012.

[17] R. A. Gopinath, "Maximum likelihood modeling with Gaussian distributions for classification," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Seattle, USA, May 1998, pp. 661–664.

[18] Christos Koniaris, Saikat Chatterjee, and W. Bastiaan Kleijn, "Selecting static and dynamic features using and advanced auditory model for speech recognition," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Dallas, USA, Mar. 2010, pp. 4342–4345.

[19] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying (Andrew) Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland, *The HTK Book (for HTK Version 3.4.1)*, Cambridge University Engineering Department, Cambridge, UK, 2009.

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