Improved Warping-Invariant Features for Automatic Speech Recognition

Jan Rademacher, Matthias Wächter and Alfred Mertins

Signal Processing Group
Dept. of Physics, University of Oldenburg
26111 Oldenburg, Germany
jan.rademacher@uni-oldenburg.de

Abstract

In this paper, we extend a previously introduced method for the generation of vocal tract length invariant (VTLI) features. The novelty is a reduction of the number of obtained invariances to the more desired ones, which results in a significant improvement of recognition rates. In experiments on the TIMIT database, the enhanced discrimination capabilities and robustness to mismatches between training and test conditions are shown.

Index Terms: Automatic speech recognition (ASR), feature extraction, vocal tract length normalization (VTLN), vocal tract length invariance (VTLI), warping-invariant features, gammatone analysis .

1. Introduction

The variation of the vocal tract length from speaker to speaker leads to shifts in the frequency of the prominent spectral peaks (formants) of speech, negatively affecting the performance of automatic speech recognition (ASR) systems. For this reason, vocal tract length normalization (VTLN) [1, 2] has become an integral part of many ASR engines. The background behind the normalization is basically the fact that the short-time spectra of two speakers A and B, when uttering the same vowel, are approximately related as $X_A(\omega) = X_B(\alpha\omega)$, where α is related to the vocal tract length ratio of both speakers. The frequency warping itself is typically carried out by warping the Mel filters when producing Mel-frequency cepstral coefficients (MFCCs). Determining the optimal α is, in general, a computationally expensive task, which is one of the main drawbacks of the method.

Besides warping of short-time spectra, also the computation of warping-invariant features has been proposed. The methods include the scale transform [3] and a more general technique for the generation of vocal tract length invariant (VTLI) features that was introduced by Mertins and Rademacher in [4, 5]. In the latter method, the wavelet transform was used as a preprocessor that produces a time-frequency analysis in which linear frequency warping results in a translation with respect to a log-frequency parameter. In a second step, VTLI features were generated by analyzing the wavelet representations in a translation-invariant manner. The methods studied in [4, 5] include the auto- and cross-correlations of local wavelet spectra magnitudes as well as linear and nonlinear transforms thereof.

The work of [4, 5] was extended in [6] by considering auditory-system motivated primary frequency analyses. While a strict wavelet analysis with logarithmically spaced center frequencies exactly carries out the conversion of linear frequency warping of sinusoidal inputs into a translation in the log-frequency domain, it does not exactly match the frequency analysis that is carried out in the human auditory system. A more human-auditory-system motivated frequency analysis is obtained with so-called gammatone filterbanks, which have been found from physiological animal experiments as well as from mathematical analyses of cochlear mechanics. The frequency resolution of the human auditory system is best represented with the so-called equivalent rectangular bandwidth (ERB) scale, which has been found from masking experiments.

The method in [4, 5, 6] not only yields invariance with respect to frequency translation, but also with respect to other operations. This undesired side effect may reduce the discrimination capabilities of the features. In this paper, we show a method of how to reduce the number of invariances while retaining the desired one of translation invariance. This allows us to further increase the performance of VTLI features, especially when the distribution of vocal tract lengths in the training set does not match the one in the test set.

The paper is organized as follows. In the next section, we shortly introduce the wavelet transform as it is used in this paper and then describe the gammatone analysis. Section 3 then presents the generation of the proposed, warping-independent VTLI features and the reduction of associated invariances. In Section 4 we describe the experimental setup and present results on phoneme recognition experiments. Section 5 gives some conclusions.

2. Primary time-frequency analysis

In this section, we discuss two signal representations that naturally enable the extraction of features in such a way that linear frequency warping is transformed into a simple translation with respect to a specific frequency scale. The first one is the integral wavelet transform, implemented in its discretized version. The second one is the auditory motivated gammatone analysis.

2.1. The wavelet transform

The discrete-time wavelet transform of a signal $x(n)$ can be computed as

$$w_x(n,k) = 2^{-k/(2M)} \sum_m x(m)\, \psi^*\Big(\frac{m-nN}{2^{k/M}}\Big), \qquad (1)$$

where M is the number of voices (subbands) per octave, and N is the subsampling factor used to reduce the sampling rate in the wavelet subbands. Assuming K octaves, the scaling parameter of the wavelet transform takes on values $a_k = 2^{k/M}$, $k = 0, 1, \ldots, MK-1$. The continuous-time wavelet $\psi(t)$, whose samples occur in the sum in (1), is the so-called mother wavelet.

The wavelet analysis will have better time resolution at higher frequencies than needed for producing feature vectors every 5 to 15 ms. Direct downsampling of features will therefore introduce aliasing artifacts. Since we are mainly interested in the signal-energy distribution over time and frequency, we may take the magnitude of $w_x(n,k)$ and filter it with a lowpass filter in time direction before final downsampling. The final primary features will then be of the form

$$y_x(n,k) = \sum_{\ell} h(\ell)\, |w_x(nL-\ell,k)| \qquad (2)$$

where $h(\ell)$ is the impulse response of the lowpass filter, L is the downsampling factor introduced to achieve the final frame rate $f_s/(N \cdot L)$, and f_s is the sampling frequency. To avoid that the filtered values $y_x(n,k)$ can become negative, we assume a strictly positive sequence $h(n)$.

2.2. The gammatone analysis

The wavelet transform described above is a true constant-Q analysis with the same relative bandwidth in all frequency bands. However, the assumption of constant relative bandwidths as well as the strict logarithmical frequency-spacing as mentioned before, does not correspond with the filtering process in the human auditory system [7]. The impulse responses of the filters in the auditory system can be approximated by the following sampled impulse response of a complex analog gammatone filter [8]

$$p_k(n) \quad = \quad n^{\gamma-1} \cdot \lambda^n \exp(j\beta_k n) \quad , \qquad n \geq 0 \qquad (3)$$

where λ denotes the bandwidth or damping parameter, β_k determines the center frequency of the kth filter, and γ denotes the filter order. The center frequency f_k of such a filter is parameterized by the angle β_k which takes the value $\beta_k = 2\pi f_k/f_s$.

Using the following analytical expression for the ERB of auditory filters as a function of the frequency f as given in [9],

$$ERB_{auditory}(f) \quad = \quad 24.7 + \frac{f}{9.265}, \qquad (4)$$

Patterson et al. showed in [7] that the damping parameter λ can be well approximated by

$$\lambda \quad = \quad \exp\left(-\frac{2\pi b}{f_s}\right), \quad b = ERB \cdot \frac{(\gamma-1)!^2}{\pi\,(2\gamma-2)! \cdot 2^{-(2\gamma-2)}} \quad (5)$$

leading to an auditory motivated, constant bandwidth on the ERB scale.

Given a representation

$$g_x(n,k) = \sum_{m} x(n-m) p_k(m), \qquad (6)$$

the final primary representation $y_x(n,k)$ is then computed as in (2) with $w_x(n,k)$ being replaced by $g_x(n,k)$.

3. Warping-invariant features

Based on a time-frequency analysis $y_x(n,k)$ of a signal $x(n)$ in which linear frequency warping results in a translation of $y_x(n,k)$ with respect to k, it is possible to obtain VTLI features by analyzing the representations in a translation-invariant manner. There are different methods to achieve translation invariance.

First, let us consider the Fourier transforms of $y_x(n,k)$ and $y_x(n,k-k_0)$ with respect to the parameter k. We obtain

$$\begin{aligned} Y_x(n,e^{j\nu}) &= \sum_k y_x(n,k)\, e^{-j\nu k} \\ Y_{x,k_0}(n,e^{j\nu}) &= \sum_k y_x(n,k-k_0)\, e^{-j\nu k} \quad = e^{-j\nu k_0} Y_x(n,e^{j\nu}) \end{aligned} \qquad (7)$$

Thus, the magnitude of $Y_{x,k_0}(n,e^{j\nu})$ is independent of k_0, resulting in a VTLI feature set.

Other possibilities to achieve translation invariance include, but are not limited to, correlation sequences with respect to the log-frequency index k, between transform values or nonlinear functions thereof at two time instances n and $n-d$. A straight forward set, which is still related to the Fourier transform mentioned above, is given by

$$r_{yy}(n,d,m) = \sum_k y_x(n,k) y_x(n-d,k+m) \qquad (8)$$

Clearly, also the logarithm of the correlation sequence is translation invariant. Using $\log r_{yy}(n,d,m)$ as features then simulates the compressive nonlinearity of the hearing system, similar to the log operation in the generation of classical MFCCs.

The correlation of logarithmized spectral values yields another possibility:

$$c_{yy}(n,d,m) = \sum_k \log(y_x(n,k)) \cdot \log(y_x(n-d,k+m)). \qquad (9)$$

A feature vector for time index n can then contain any collection of the above mentioned features computed for the same index n. For $d = 0$ these features will give information on the signal spectrum in time frame n. For $d \neq 0$ they will give information on the development of short-time spectra over time.

Any linear or nonlinear combination and/or transform or filtering of $r_{yy}(n,d,m)$ and $c_{yy}(n,d,m)$, including taking derivatives (i.e., delta and delta-delta features) will also yield warping invariant features.

One drawback of techniques for invariant-feature generation proposed in [3, 4, 5, 6] is the fact that the generated features are invariant with respect to more operations than desired. For example, the correlation sequences $r_{yy}(n,d,m)$ and $r_{\widetilde{yy}}(n,d,m)$ of $y_x(n,k), 0 \leq k \leq MK-1$ and $\widetilde{y}_x(n,k) = y_x(n, MK-1-k)$ are the same. More general, arbitrary zeros of the z-polynomial

$$Y_x(n,z) = \sum_k y_x(n,k) z^{-k}$$

can be inverted without affecting the magnitude of $Y_x(n,e^{j\nu})$, and thus, without affecting the corresponding correlation sequences. This unwanted side effect has the potential to reduce the discrimination capabilities of the generated features for the task of speech recognition.

In the following, we present a new, extended approach for the generation of invariant features that reduces the number of related invariances but keeps the desired invariance with respect to frequency warping. In the extended method, we first convert the real-valued primary feature set $y_x(n,k)$ into a complex-valued one, denoted by $u_x(n,k)$, in which the values $y_x(n,k)$ are encoded in both the magnitude and the phase. The proposed generation of $u_x(n,k)$ is as follows:

$$u_x(n,k) = y_x(n,k) \cdot \exp\left(j \left(\frac{y_x(n,k)}{\sqrt{\sum_k |y_x(n,k)|^2}} \right)^{\kappa} \cdot \frac{\pi}{4} \right), \quad (10)$$

where κ denotes a scaling exponent. The normalization of the phase term ensures that the angle does not exceed $\pi/4$. All remaining processing steps are the same as before, using $u_x(n,k)$ instead of $y_x(n,k)$. The correlation sequence

$$r_{uu}(n,d,m) = \sum_k u_x^*(n,k) u_x(n-d,k+m) \quad (11)$$

is then generally complex, where the superscript * denotes complex conjugation. The phase of $r_{uu}(n,d,m)$ provides additional information that reduces the class of invariances, and at the same time, keeps the desired invariance to vocal tract length variations.

4. Experimental setup and results

In our experiments, different setups using the linear-phase wavelet transform described in Section 2.1 and the nonlinear-phase, auditory-system motivated gammatone filterbank according to Section 2.2 were used. For the wavelet transform, we used the linear-phase Morlet wavelet given by

$$\psi(n) = e^{j\omega_0 n} \, e^{-\frac{n^2}{2\sigma_n^2}} \quad (12)$$

with $\omega_0 = 0.9\pi$ and $\sigma_n^2 = 100$. The initial downsampling factor N was chosen as $N = 1$. The transform was carried out for $M = 12$ voices (subbands) per octave and $K = 7$ octaves yielding 84 wavelet coefficients. For the gammatone filterbank, an ERB based approach with 90 ERB spaced center frequencies was examined. Center frequencies were considered in the range of 40-6700 Hz, each with a bandwidth of one ERB. For both the wavelet and the gammatone features, the lowpass filter $h(n)$ was a rectangular window of 200 coefficients, and the frame rate was set to one frame every 10 ms.

A number of phoneme recognition tests were performed with these setups using the TIMIT database (including the SA files) with a sampling rate of 16kHz. The training and test sets were both split into male and female subsets in order to allow for training and testing under different conditions. In the following, M+F, M, and F denote training/test on male+female, male, and female data, respectively. Following the procedure in [10], 48 phonetic models were trained, and the recognition results were folded to yield 39 final phoneme classes that had to be distinguished.

Phoneme recognition tests were performed by a HMM-based phoneme recognizer using monophone models, bigram statistics, three states per phoneme, 8 Gaussian mixtures per state, and diagonal covariance matrices. The recognizer was based on the Hidden-Markov-Toolkit (HTK) [11].

Besides well known MFCCs, vocal tract length normalized MFCCs (VTLN-MFCC) were investigated. The acoustic likelihood for each utterance warped by different warping factors were calculated by a Viterbi decoder. From this, the optimal warping factors were derived for each speaker by an exhaustive search and the models were re-trained using features warped by those factors. This procedure was applied twice to the training features and the corresponding HMMs. For the features of the test set, optimal warping factors were obtained for each utterance in the same way. After warping the test features using the optimal factors, the final recognition was performed.

Warping invariant features based on the real-valued primary features $y_x(n,k)$ and the correlation terms (8) and (9) were generated for both the wavelet and the gammatone analysis. These features will be referred to as "VTLI-WT-F" and "VTLI-GT-F", respectively, in the following. The feature sets consist of the following selection:

- the first 20 coefficients of the discrete cosine transform (DCT) of $\log(r_{yy}(n,0,m))$ with respect to parameter m
- the first 20 coefficients of the DCT of $c_{yy}(n,4,m)$ with respect to parameter m
- $\log(r_{yy}(n,4,m))$ for $m = -2,-1,\ldots,2$

The warping-invariant features were also amended with classical MFCC features. For this, the 12 MFCCs and the single energy feature of the standard HTK setup were used (denoted by 13 MFCC in the following), produced with the same frame rate and a frame length of 20 ms. Moreover, the first 15 DCT coefficients of the logarithmized features $\log(y_x(n,k))$ were used for feature set amendment as well (DCT with respect to frequency parameter k) and denoted as "WT" or "GT" respectively. After the amendment of these feature sets by their delta and delta-delta features, reduced feature sets of 47 features were generated by linear discriminant analyses (LDA) [12]. The LDAs were based on the 48 phonetic classes mentioned above.

Complex-valued correlation features were calculated according to (10) and (11), with $y_x(n,k)$ based on the ERB-spaced gammatone filterbank and the parameter κ set to κ =0.2. Here, the following VTLI feature set was generated:

- the first 20 coefficients of the DCT of $\log(|r_{uu}(n,0,m)|)$ with respect to parameter m
- the first 20 coefficients of the DCT of $\arg(r_{uu}(n,0,m))$ with respect to parameter m
- the first 20 coefficients of the DCT of $c_{yy}(n,4,m)$ with respect to parameter m
- $\log(r_{yy}(n,4,m))$ for $m = -2,-1,\ldots,2$

These warping-invariant features were then amended with MFCCs and the features $\log(y_x(n,k))$ as well as all delta and delta-delta features and then reduced to a subset of 47 features in a similar manner as mentioned before. The corresponding feature set will be labeled with the subscript "u".

Altogether, the following feature sets were considered for experiments on the TIMIT corpus:

- 3×13 MFCCs: 13 MFCCs and the corresponding delta and delta-delta features.
- 3×13 VTLN-MFCCs: 13 VTLN-MFCCs and the corresponding delta and delta-delta features.
- VTLI-WT-F+MFCC+WT.
- VTLI-GT-F+MFCC+GT.
- VTLI-GT$_u$-F+MFCC+ GT$_u$.

Table 1 contains the results for HMM-based phoneme recognition using the above mentioned feature sets. The standard MFCCs yield good results in the M+F setting, but their performance significantly deteriorates when training and test conditions differ. The best performance is achieved with MFCCs and VTLN, as one might expect. An almost equally good performance, however, is achieved with the proposed complex-correlation based feature set VTLI-GT$_u$-F+MFCC+GT$_u$. In the M+F setting, these features are as good as the MFCCs, and in the mismatch conditions, they almost reach the performance of VTLN, without the additional need of determining the warping factor explicitly.

Table 1: Phoneme recognition accuracies in % for TIMIT.

Features	Train Cond.	Test Cond.	Accuracy
3×13 MFCC	M+F	M+F	69.19
3×13 VTLN-MFCC	M+F	M+F	68.97
VTLI-WT-F+MFCC+WT	M+F	M+F	67.84
VTLI-GT-F+MFCC+GT	M+F	M+F	68.82
VTLI-GT$_u$-F+MFCC+GT$_u$	M+F	M+F	69.20
3×13 MFCC	M	F	56.84
3×13 VTLN-MFCC	M	F	65.74
VTLI-WT-F+MFCC+WT	M	F	63.56
VTLI-GT-F+MFCC+GT	M	F	63.15
VTLI-GT$_u$-F+MFCC+GT$_u$	M	F	64.92
3×13 MFCC	F	M	55.53
3×13 VTLN-MFCC	F	M	66.94
VTLI-WT-F+MFCC+WT	F	M	62.98
VTLI-GT-F+MFCC+GT	F	M	63.00
VTLI-GT$_u$-F+MFCC+GT$_u$	F	M	64.71

5. Conclusions

We have proposed a technique for the extraction of vocal tract length invariant features with an auditory-filterbank based preprocessing. The number of unwanted invariances that occur as a side effect of an invariance transform are minimized by encoding the information in both the magnitude and phase of temporary feature vectors prior to the invariance transform. This leads to an enhancement of recognition rates and to more robustness. The results have shown that the new features are complementary to the well-known MFCCs and that they can be used to construct combined feature sets that are robust to speaker variations, especially when the training conditions do not match the test conditions. The more complex VTLN technique, however, still gives slightly better results. Future work will be directed toward fine tuning the processing steps and the feature selection, to close the gap of recognition accuracy between VTLI features and the VTLN method.

6. Acknowledgements

This work was supported by the European Commission 6th Framework Program project DIVINES under contract number FP6-002034.

7. References

[1] A. Andreou, T. Kamm, and J. Cohen, "Experiments in vocal tract normalization," in *Proc. CAIP Workshop: Frontiers in Speech Recognition II*, 1994.

[2] 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.

[3] S. Umesh, L. Cohen, N. Marinovic, and D. Nelson, "Scale transform in speech analysis," *IEEE Trans. Speech and Audio Processing*, vol. 7, no. 1, pp. 40–45, Jan. 1999.

[4] A. Mertins and J. Rademacher, "Vocal tract length invariant features for automatic speech recognition," in *Proc. 2005 IEEE Automatic Speech Recognition and Understanding Workshop*, San Juan, Puerto Rico, Nov. 27 -Dec. 1 2005, pp. 308–312.

[5] ——, "Frequency-warping invariant features for automatic speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, vol. V, Toulouse, France, May 14-19 2006, pp. 1025–1028.

[6] J. Rademacher and A. Mertins, "A study of auditory-filterbank based preprocessing for the generation of warping-invariant features," in *Proc. Speech Recognition and Intrinsic Variation Workshop*, Toulouse, France, Mai 20 2006.

[7] R. D. Patterson, J. Nimmo-Smith, J. Holdsworth, and P. Rice, "An efficient auditory filterbank based on the gammatone function," in *Proc. Meeting of the IOC Speech Group on Auditory Modelling at RSRE*, December 14-15 1987.

[8] V. Hohmann, "Frequency analysis and synthesis using a Gammatone filterbank," *Acta Acustica United with Acustica*, vol. 88, pp. 433–442, 2002.

[9] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," in *Hearing Research*, vol. 47, 1990, pp. 103–138.

[10] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 2, pp. 1641 – 1648, Nov. 1989.

[11] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, Cambridge University, 1995.

[12] K. Fukunaga, *Introduction to Statistical Pattern Recognition*. New York: Academic Press, 1972.