VOCAL TRACT LENGTH INVARIANT FEATURES FOR AUTOMATIC SPEECH RECOGNITION

Alfred Mertins and Jan Rademacher

Signal Processing Group
University of Oldenburg, Institute of Physics, 26111 Oldenburg, Germany
Email: {alfred.mertins, jan.rademacher}@uni-oldenburg.de

ABSTRACT

The effects of vocal tract length (VTL) variation are often approximated by linear frequency warping of short-time spectra. Based on this relationship, we present a method for generating vocal tract length invariant features. These new features are computed as translation invariant, correlation-type features in a log-frequency domain. In phoneme classification experiments, their discrimination capabilities turned out to be considerably better than for Mel-frequency cepstral coefficients (MFCCs). The best results are obtained when VTL-invariant (VTLI) features and MFCCs are combined. The superiority of the combined feature set and its resilience to VTL variations is also shown for word recognition, using the TIDIGITS corpus and the HTK recognizer.

1. INTRODUCTION

Vocal tract length normalization [1, 2] has become an integral part of many automatic speech recognition 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). The factor α usually lies in the range between 0.8 and 1.2, relative to an average speaker. More recent approaches even normalize the utterances from the same speaker with optimal α on a frame-by-frame basis, in order to even better match the standard realizations of the phonemes [3]. The value of α is often selected as the one that yields the highest likelihood scores in a subsequent hidden Markov model (HMM) based recognizer, when testing a number of given values in the above mentioned range [2, 3]. However, determining the optimal α is, in general, a computationally expensive task.

Besides warping of short-time spectra, also the computation of warping-invariant features has been proposed in form of the scale transform [4]. For the scale transform, the magnitude spectra of two signals $x(t)$ and $\frac{1}{\sqrt{\alpha}}x(t/\alpha)$ are the same. In this paper, we also aim at producing warping-invariant features. However, in contrast to [4], we base our analysis on the wavelet transform, which naturally represents a signal with respect to a logarithmized frequency axis. The initial frequency resolution of the wavelet transform used in this paper is much higher than the resolution obtained in typical Mel filterbanks or that of the scale transform, as computed in [4]. This allows us to obtain highly selective, warping independent features in form of correlation sequences or nonlinear functions thereof. These features will be referred to as vocal tract length invariant (VTLI) features henceforth.

Experimental results for different recognition and classification tasks show that the produced features are robust and complementary to standard MFCCs, so that both sets can be combined in order to obtain highly selective and yet robust feature sets. For frame-wise phoneme classification using simple linear classifiers, the results for the combined feature set are significantly better than for MFCCs. Also in digit recognition, especially when the training data does not match the test conditions or in the presence of background noise, the combined set is significantly superior to the MFCCs alone.

The paper is organized as follows. In the next section, we discuss the scale and the wavelet transforms and their capabilities of producing warping-independent features. Section 3 then presents the proposed features that are computed as functions of the wavelet coefficients. In Section 4 we describe the experimental setup and the method of feature combination. Experimental results on phoneme classification and word recognition are given in Section 5. Finally, Section 6 gives some conclusions.

This work was supported by the EU DIVINES Project under Grant No. IST-2002-002034.

0-7803-9479-8/05/$20.00 © 2005 IEEE

2. TRANSFORMS THAT LEAD TO WARPING-INVARIANT FEATURES

In this section, we discuss two signal representations that naturally enable the extraction of features which are robust to vocal tract length variations. The first one is the scale transform, introduced by Umesh et al. [4] in order to generate features that are independent of linear frequency warping and thus to vocal tract length variations. The second one is the integral wavelet transform, implemented in its discretized version.

The scale transform is defined as

$$D_x(c) = \int_0^\infty X(f)\, \frac{e^{-j2\pi c\, \ln f}}{\sqrt{f}}\, df \qquad (1)$$

where $X(f)$ is the signal spectrum with f being the frequency in Hz and c is the scale parameter. This transform exhibits the interesting property that the scale transform of a frequency warped signal $\sqrt{\alpha}X(\alpha f)$ is given by $D_x^{(\alpha)}(c) = e^{j2\pi c \ln \alpha} D_x(c)$, so that its magnitude is independent of the warping parameter α.

In addition to the scale transform itself, also a scale cepstrum was introduced in [4]. This is defined as

$$D_s(c) = \int_0^\infty \log|S(f)|\, \frac{e^{-j2\pi c \ln f}}{\sqrt{f}}\, df \qquad (2)$$

where $S(f)$ is the Fourier transform of a short-time autocorrelation estimate $r_{xx}(m)$. Again, the magnitude of the scale cepstrum is invariant to linear frequency warping.

The wavelet transform of a continuous-time signal $x(t)$ is given by

$$\mathcal{W}_x(t,a) = |a|^{-\frac{1}{2}} \int_{-\infty}^{\infty} x(\tau)\, \psi^*\Big(\frac{\tau - t}{a}\Big)\, d\tau \qquad (3)$$

where $\psi(t)$ is the so-called mother wavelet, a is the scaling parameter, and the asterisk * denotes complex conjugation. By varying a, the center frequency, bandwidth, and effective time-width of $\psi(t/a)$ are changed according to the scaling theorem of the Fourier transform.

In our context, the wavelet $\psi(t)$ is assumed to be analytic, which means that it satisfies $\Psi(\omega) = 0$ for $\omega \leq 0$ where $\Psi(\omega)$ is the Fourier transform of $\psi(t)$. Such wavelets can also be seen as impulse responses of analytic bandpass filters.

To see the effect of frequency warping, we consider the computation of $\mathcal{W}_x(t,a)$ from $X(\omega)$ (the Fourier transform of $x(t)$), in the form [5]

$$\mathcal{W}_x(t,a) = |a|^{\frac{1}{2}}\, \frac{1}{2\pi} \int_{-\infty}^{\infty} X(\omega)\, \Psi^*(a\omega)\, e^{j\omega t} d\omega. \qquad (4)$$

From this expression, we see that the wavelet transform $\mathcal{W}_{x_\alpha}(t,a)$ of a normalized, linearly frequency warped signal $x_\alpha(t) = \frac{1}{\sqrt{\alpha}} x(\frac{t}{\alpha})$, $\alpha > 0$, with spectrum $X_\alpha(\omega) = \sqrt{\alpha}X(\alpha\omega)$ is related to $\mathcal{W}_x(t,a)$ as

$$\mathcal{W}_{x_\alpha}(t,a) = \mathcal{W}_x\Big(\frac{t}{\alpha}, \frac{a}{\alpha}\Big) \qquad (5)$$

The scaling of the time axis in (5) is inherent to frequency warping and also applies to the scale transform. It is of no concern here, as we are only interested in the short-time behavior of signal spectra. The scaling of the parameter a shows that a linear frequency warping of the signal by a factor of α results in a translation of the wavelet transform by $\log\alpha$ in the $(\log a)$-domain. This is important, because the wavelet transform is naturally computed for equally spaced values of $\log a$.

Now let us take the Fourier transforms of $\mathcal{W}_x(t,a)$ and $\mathcal{W}_{x_\alpha}(t,a)$ with respect to the parameter $\nu = \log a$, considering the relationship (5):

$$F(t,\mu) = \int_{-\infty}^{\infty} \mathcal{W}_x\Big(t, e^{\nu}\Big)\, e^{-j\mu\nu}\, d\nu, \qquad (6)$$

$$F_\alpha(t,\mu) = \int_{-\infty}^{\infty} \mathcal{W}_x\Big(\frac{t}{\alpha}, e^{\nu - \log\alpha}\Big)\, e^{-j\mu\nu}\, d\nu. \qquad (7)$$

Hence,

$$F_\alpha(t,\mu) = e^{-j\mu\log\alpha} F\Big(\frac{t}{\alpha}, \mu\Big) \qquad (8)$$

Thus, ignoring the time scaling, we see that the magnitudes of the Fourier transforms are the same. Therefore, one obtains features that are invariant to linear frequency warping. The transforms $F_\alpha(t,\mu)$ and $D_x(c)$, although having similar frequency-warping properties, are very different in their time-frequency resolution. While $F_\alpha(t,\mu)$ inherently has the zoom-in effect of the wavelet transform, the transform $D_x(c)$ has, due to the way it is computed, inherited the time resolution of the short-time Fourier transform.

Note that taking the magnitude of the $F(t,\mu)$ is only one of several possibilities to obtain features that are not affected by linear frequency warping. More possibilities will be discussed in the next section.

We now consider the computation of the wavelet transform for a discrete-time signal $x(n)$. We assume K octaves, using M voices per octave, which means that the scaling parameter a takes on values $a_k = 2^{k/M}$, $k = 0, 1, \ldots, MK-1$. Moreover, we consider the computation of the wavelet transform with time shifts of N. By discretizing (3) we then obtain the values

$$w_x(n,k) = 2^{-k/(2M)} \sum_m x(m)\, \psi^*\Big(\frac{m - nN}{2^{k/M}}\Big) \qquad (9)$$

Due to the constant sampling rate in all frequency bands, the wavelet transform (9) does not suffer from the same shift-invariance problem as the discrete wavelet transform (DWT). Rather than implementing (9) directly, which means a significant computational load, one may use the à trous algorithm [6], implemented separately for each of the M voices.

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. For the wavelet transform, the final primary features will then be of the form

$$y_x(n,k) = \sum_{\ell} h(\ell)\, |w_x(nL-\ell,k)| \qquad (10)$$

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 $h(n)$ like, for example, the Hanning window.

3. GENERATION OF WARPING-INVARIANT FEATURES

From the discussion in the previous section it is evident that a linear frequency warping leads to a complex, unit-magnitude prefactor for the scale transform and a translation for the wavelet transform. Therefore, for the wavelet transform, any translation-invariant features will automatically be invariant to linear frequency warping.

In the following, we will consider the primary features $y_x(n,k)$, which already occur in the final frame rate, in order to generate warping-invariant features. Taking the magnitude of the Fourier transform with respect to frequency parameter k has already been mentioned as an example in Section 2.

Other possibilities 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 $n-d$. In particular, we here consider

$$r_x(n,d,m) = \sum_{k} y_x(n,k) y_x(n-d,k+m) \qquad (11)$$

and

$$c_x(n,d,m) = \sum_{k} \log(y_x(n,k)) \cdot \log(y_x(n-d,k+m)). \qquad (12)$$

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.

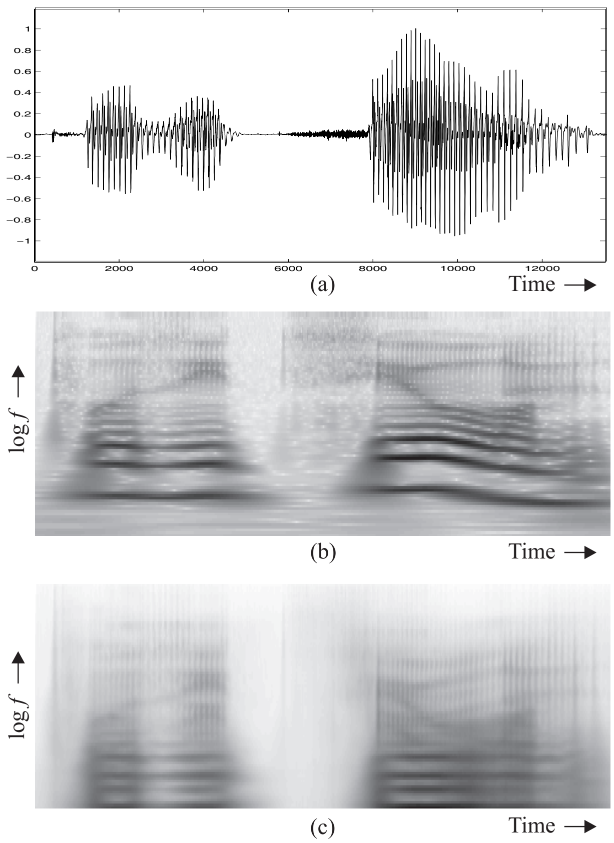

Fig. 1. Example of wavelet analysis and autocorrelation features. (a) Time signal. (b) Wavelet spectrum $y_x(n,k)$. (c) Autocorrelation features $r_x(n,0,m)$ for $m \geq 0$.

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

To give an illustration of the properties of the correlation-based features, we consider the set $r_x(n,d,m)$ for $d=0$ (i.e., autocorrelation features). Fig. 1 shows an example in which the waveform $x(n)$, the spectra $y_x(n,k)$ and the autocorrelation $r_x(n,0,m)$ are plotted. It is interesting to see that the autocorrelation, although it is in some sense phase-blind, still retains the formant structure. This is due to the fact that noticeable correlation values are achieved when the high-energy pitch component is shifted and multiplied with the formant components during the correlation operation. Under the assumption that the linear warping model is true for vocal tract length variations, these format-related structures will indeed be independent of the warping factor. For real speech, of course, this is only an approximation [7], but it leads to formant-like structures that are robust to vocal tract length variations.

4. EXPERIMENTAL SETUP AND FEATURE COMBINATION

In our experiments, we used the linear-phase wavelet transform based on the Morlet wavelet [5] given by

$$\psi(n) = e^{j\omega_0 n}\, e^{-\frac{n^2}{2\sigma_n^2}} \qquad (13)$$

with $\omega_0 = 0.9\pi$ and $\sigma_n^2 = 100$. The transform was carried out for $M = 12$ voices per octave and $K = 6$ octaves for data sampled at 16kHz and $K = 5$ octaves for data sampled at 8 kHz. This yields 72 and 84 wavelet coefficients at sampling rates of 8 and 16 kHz, respectively. The initial downsampling factor N was chosen as $N = 10$. The low-pass filter $h(n)$ was designed as a Hanning window, and the final downsampling was done to obtain a frame every 12.5 ms.

The following warping-invariant features were used:

- the first 20 coefficients of the discrete cosine transform (DCT) of $\log(r(n, 0, m))$ with respect to parameter m for $m = 0, 1, \ldots, 84$.
- the first 20 coefficients of the DCT of $c(n, 2, m)$ with respect to parameter m with $m = -84, \ldots, 84$.
- $\log(r(n, 2, m))$ for $m = -2, -1, \ldots, 2$

Because the warping-invariant features are mainly of interest for the classification of vowels, they were also amended with 13 classical MFCC features, produced with the same frame rate and a frame length of 25 ms. Moreover, the first 15 DCT coefficients of the logarithmized wavelet features $\log(y_x(n, k))$ were used for feature set amendment as well (DCT with respect to frequency parameter k).

For all static features, also the delta and delta-delta coefficients were computed.

To reduce the size of the feature vectors, the collected features (maximally 219 in our case, when all the above mentioned features were used) were fed into a linear discriminant analysis (LDA) [8] that was set up to deliver reduced feature vectors which yield the best results for free phoneme classification on the basis of individual frames, using a linear classifier. Thus, a given feature vector $\boldsymbol{X}$, containing the above mentioned features, was transformed into a new vector $\boldsymbol{x} = \boldsymbol{U}^T\boldsymbol{X}$ where the columns of matrix $\boldsymbol{U}$ are the eigenvectors of a matrix $\boldsymbol{S} = [\boldsymbol{S}_w^{-1}\boldsymbol{S}_b]$, where $\boldsymbol{S}_w$ is the within-class scatter matrix, averaged over all phonemes under consideration, and $\boldsymbol{S}_b$ is the between-class scatter matrix.

5. EXPERIMENTAL RESULTS

In this section we present results for two different tasks. The first one is phoneme classification where decisions are made on the basis of single feature vectors. The second one is word recognition. In all experiments, the sampling rate for the speech waveforms was 8 kHz.

Table 1. Accuracies in % for frame-wise phoneme classification. In all cases, delta and delta-delta features were included prior LDA. ”ST”, ”WT”, and ”VTLI-F” stand for scale transform, wavelet transform, and VTLI features, resp.

Original features	number of used features	Training set	Test set
13 MFCC	39	34.37	34.66
45 VTLI-F	39	39.59	39.36
45 VTLI-F, 13 MFCC	39	43.05	42.95
45 VTLI-F, 13 MFCC, 15 WT	39	44.10	44.01
45 VTLI-F	55	40.01	39.64
45 VTLI-F, 13 MFCC	55	44.00	43.64
45 VTLI-F, 13 MFCC, 15 WT	55	**45.19**	**44.75**
128 ST	55	32.30	31.51
128 ST, 13 MFCC	55	40.42	39.19

For the LDA-based feature combination and subsequent phoneme classification, the TIMIT corpus was used. By merging differently labeled types of silence and removing unused phone labels, the original 62 labels were mapped onto 56 different possible phoneme labels. The LDA was carried out to find the P best features for linear phoneme classification. Frames for which the 25 ms window for MFCC calculation covered two differently labeled sections were not considered.

The value of P was chosen as 39 and 55, respectively.[1] For phoneme classification, the classifier was a single-layer perceptron [9]. Such a simple classifier cannot deliver recognition results as good as a Gaussian mixture model (GMM) based classifier or a complete HMM-based phoneme recognizer, but the results still give an indication of the quality of a feature set. In frame-wise phoneme classification, especially confusion between long and short versions of the same phoneme have to be expected, as the differences cannot be seen from a single frame.

For a first experiment, the TIMIT corpus was divided into two equally sized portions. Only one of them was used for training the LDA and the linear classifier. Results for different feature selections are listed in Table 1. From these results we see that the warping-invariant features alone are already better than MFCCs. The combination of both sets yields an additional improvement, and the best results are obtained when all wavelet, MFCC, and invariant features are linearly combined to a final feature set of 55 features.[2] These results also show the complementariness of invariant

[1] Using more than 55 features after LDA is not useful, because the rank of $\boldsymbol{S}_b$ can only be 55 when 56 classes are used.

[2] The fact that the error rates on the training and test sets are similar shows that no overfitting of the classifier has occurred.

Table 2. Accuracies in % for frame-wise phoneme classification. The training was done on male data only. In all cases, delta and delta-delta features were included prior LDA.

Original features	number of used features	Male	Female
13 MFCC	39	36.93	28.08
128 ST	55	37.58	27.27
128 ST, 13 MFCC	55	42.77	31.00
45 VTLI-F	39	41.45	32.10
45 VTLI-F, 13 MFCC, 15 WT	55	**47.45**	**36.38**

features and classical ones like MFCCs. A small degradation is seen when only 39 instead of 55 combined features are used. The scale transform yields about the same performance as the MFCCs, and in combination with MFCCs, the performance is comparable to that of our invariant features alone.

In a second experiment, the TIMIT corpus was split into male and female recordings. The training was done only on the male data, and the tests were performed on both sets. Table 2 shows the results for various feature selections. In all cases we can observe a degradation for the female data. However, the results for female tests using the proposed combination of 55 features are even better than those for the MFCCs in mixed training in Table 1, and they are comparable to the MFCC male results. Again, the scale transform performs comparable to the MFCCs, and in combination with MFCCs, it can improve slightly.

In addition to phoneme classification, the proposed features have been tested on a word recognition task in a setting where the training conditions do not match the test conditions. For this, we have taken ”man” and ”woman” data from the TIDIGITS corpus for training a word recognizer based the hidden-Markov-Toolkit (HTK). Tests were then performed on ”man” and ”woman” data that was not seen in the training as well as on the ”boy” and ”girl” data contained in TIDIGITS. The features used in this experiments were MFCCs and MFCCs together with the first five DCT coefficients of $\log(r(n, 0, m))$, respectively. In both cases, the delta and delta-delta coefficients of the static features were added. The results of the experiment are listed in Table 3. We can clearly see that the inclusion of the warping-invariant features significantly improves the robustness of the recognizer. For the ”girl” data, the error rate approximately halves due to the inclusion of the new features.

Table 3. Word recognition accuracies in % for the TIDIGITS corpus. The training was done on 847 male and 924 female files. The invariant features are the first five coefficients of the DCT of $\log(r(n, 0, m))$ with respect to the frequency lag m. In all cases, delta and delta-delta features were included.

	13 MFCC	13 MFCC + 5 VTLI-F
Man	98.08	98.39
Woman	99.19	99.31
Boy	94.47	**96.62**
Girl	91.29	**95.41**

6. CONCLUSIONS

We have proposed a technique for the extraction of features which are independent of linear frequency scaling and thus robust to vocal tract length variations. The performance of the new features has been demonstrated in both phoneme and word recognition tasks. 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 which are robust to speaker variations, especially when the training conditions do not match the test conditions. Future work will be directed toward investigating the noise robustness of the proposed features, taking more context into account during the feature extraction, and optimizing the primary time-frequency (i.e., wavelet) analysis.

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] A. Miguel, E. Lleida, R. Rose, L. Buera, and A. Ortega, “Augmented state space acoustic decoding for modeling local variability in speech,” in *Proc. Interspeech 2005, Lisbon, Portugal*, in press, 2005.

[4] 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.

[5] M. Vetterli and J. Kovačević, *Wavelets and Subband Coding*, Prentice-Hall, Englewood Cliffs, NJ, 1995.

[6] M. J. Shensa, “The discrete wavelet transform: Wedding the à trous and Mallat algorithms,” *IEEE Trans. Signal Processing*, vol. 40, no. 10, pp. 2464–2482, Oct. 1992.

[7] G. Fant, “A non-uniform vowel normalization,” *Speech Transmssion Lab. Rep., Royal Inst. Technol.,* Stockholm, Sweden, vol. 2-3, pp. 1–19, 1975.

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

[9] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Prentice Hall, Upper Saddle River, NJ, USA, 1999.