A STUDY OF AUDITORY-FILTERBANK BASED PREPROCESSING FOR THE GENERATION OF WARPING-INVARIANT FEATURES

Jan Rademacher and Alfred Mertins

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

ABSTRACT

Auditory filterbanks have a long history in the preprocessing stage of automatic speech recognition systems, with the most prominent examples being the mel frequency cepstral coefficients (MFCCs). In this paper, we study the usefulness of auditory-filterbank analyses as a preprocessor for the generation of frequency-warping invariant features. The results indicate, that gammatone-filterbank analyses following the equivalent rectangular bandwidth (ERB) scale yield the most robust feature sets. The performance improvements are most significant when the vocal tract lengths in the training and test sets differ, which is important when, for example, children speech is to be recognized with a system that was mainly trained on adult data.

1. INTRODUCTION

Vocal tract length normalization has become an integral part of many automatic speech recognition engines [1, 2]. It is based on the idea that the short-time spectra of two speakers A and B are approximately related as $X_A(\omega) = X_B(\alpha\omega)$, where α is the so-called warping factor. The value of α depends on the ratio of the vocal tract lengths of both speakers and usually lies in the range between 0.8 and 1.2, relative to an average speaker. The value of α is typically selected as the one that yields the highest likelihood scores in a subsequent hidden Markov model (HMM) based recognizer [2, 3]. Recent approaches even normalize the utterances from the same speaker with optimal α on a frame-by-frame basis, in order to better match the standard realizations of the phonemes [3].

Recently, a method for the generation of vocal tract length invariant (VTLI) features has been proposed in [4]. In this 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. 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. In this paper, we therefore study the usefulness of auditory-filterbank motivated preprocessors.

While the MEL scale has been found from the analysis of just noticeable differences in the frequency of sinusoids, the so-called equivalent rectangular bandwidth (ERB) scale has been found from masking experiments and better describes the bandwidths of the auditory filters. Moreover, the analysis of the human auditory system as well as physiological animal experiments have led to an approximation of the cochlear frequency analysis by so-called gammatone filters. Both paradigms can be combined, and gammatone filterbanks can be used with center frequencies and bandwidths that follow the ERB scale. Moreover, it has been suggested in [5] to carry out vocal tract length normalization along the MEL scale.

In this paper we study the usefulness of auditory-motivated gammatone analyses as a preprocessor for generating robust feature sets that are nearly invariant to vocal tract length variations. The paper is organized as follows. In the next section, we briefly introduce the wavelet transform and then describe the gammatone analysis. Section 3 then presents the generation of the proposed, warping-independent VTLI features. 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

2.1. The discrete-time wavelet analysis

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 per octave, and N is the subsampling factor used to reduce the sampling rates in the wavelet subbands. Assuming K octaves, the scaling parameter a takes on values $a_k = 2^{k/M}$, $k = 0, 1, \ldots, MK-1$.

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

The continuous-time wavelet $\psi(t)$, whose samples occur in the sum in (1), is the so-called mother wavelet. For this, in [4] the Morlet wavelet [6] given by $\psi(n) = \exp(j\omega_0 n) \times \exp(-\frac{n^2}{2\sigma_n^2})$ was used with the parameters $\omega_0 = 0.9\pi$ and $\sigma_n^2 = 100$. The transform $w_x(n,k)$ was carried out with $M = 12$ voices per octave and $K = 7$ octaves, resulting in primary feature vectors of length 84.

An important property of the wavelet transform (1) is that it is computed on a regular time grid with the same subsampling factor N applied to all frequency bands, regardless of the bandwidth. This is in contrast to the discrete wavelet transform (DWT), which operates on a dyadic grid and uses different sampling rates in different octaves. Due to the constant sampling rate in all frequency bands, the wavelet transform (1) does not suffer from the shift-invariance problem known from the DWT, provided that the factor N is sufficiently small and chosen in accordance with the number of voices (e.g., such that $N \leq M$).

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

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

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 $h(n)$ like, for example, the Hanning window. In [7], the lowpass filter $h(n)$ was simply a rectangular window of 200 coefficients, and the initial downsampling was set to $N = 1$.

Fig. 1 gives an example of a wavelet analysis. In Fig. 1(b), which shows $|w_x(n,k)|$ as a grey-scale image (i.e., a scalogram), the pitch is visible in most frequency bands. This pitch pattern is no longer visible in Fig. 1(c), which depicts $y_x(n,k)$.

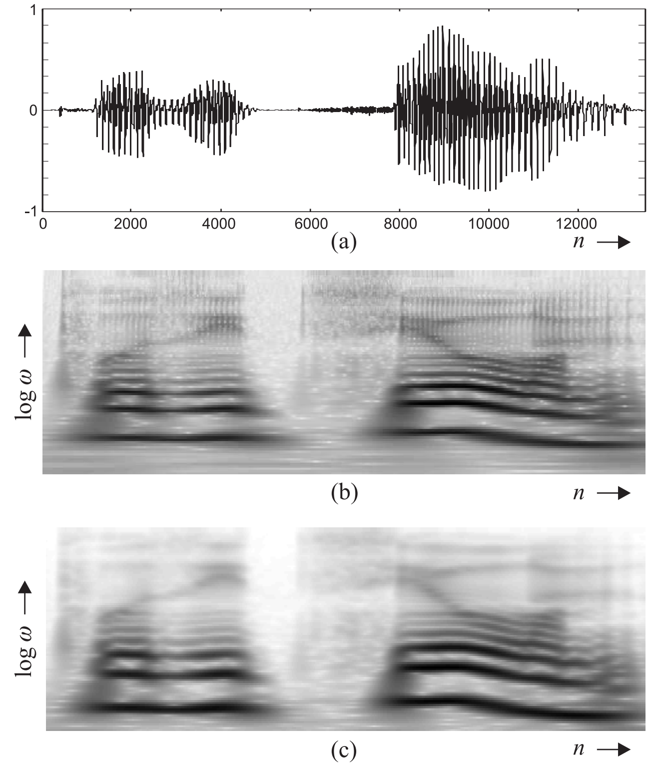

Fig. 1. Example of a wavelet analysis. (a) Time signal. (b) Wavelet transform magnitude $|w_x(n,k)|$ with $k \propto -\log\omega$. (c) Smoothed wavelet analysis $y_x(n,k)$.

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, according to Patterson et al. [8], 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.

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 [9]

$$g_\gamma(n) \quad = \quad n^{\gamma-1} \cdot \tilde{a}^n \quad , \qquad n \geq 0 \tag{3}$$

with

$$\tilde{a} \quad = \quad \lambda \cdot \exp(j\beta)$$

where λ denotes the bandwidth or damping parameter, β determines the center frequency and γ denotes the filter order.

The center frequency f_c of such a filter is parameterized by the phase β of $\tilde{a}$ which takes the value

$$\beta \quad = \quad 2\pi \frac{f_c}{f_s}. \tag{4}$$

Using the following analytical expression for the equivalent rectangular bandwidth (ERB) of auditory filters as a function of the frequency f as given in [10]

$$ERB_{auditory}(f) \quad = \quad 24.7 + \frac{f}{9.265} \tag{5}$$

Patterson et al. show in [8] that the damping parameter λ can be well approximated by

$$\lambda \quad = \quad \exp\left(-\frac{2\pi b}{f_s}\right) \tag{6}$$

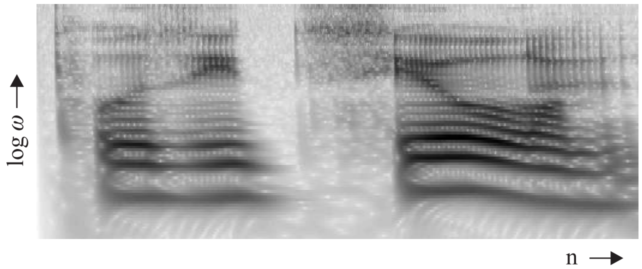

Fig. 2. Gammatone analysis.

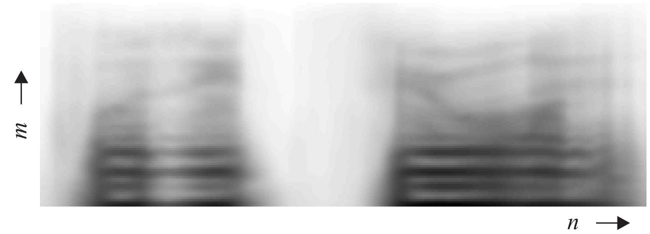

Fig. 3. Autocorrelation features $r_x(n, 0, m)$ for $m \geq 0$.

with

$$\begin{aligned} b &= ERB/a_\gamma, \\ a_\gamma &= \frac{\pi\,(2\gamma-2)! \cdot 2^{-(2\gamma-2)}}{(\gamma-1)!^2} \end{aligned}$$

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

Keeping in mind that a linear frequency warping of the signal by a factor α should yield in a translation in the log frequency domain, the individual filters should be logarithmically spaced. The corresponding representation with logarithmically spaced center frequencies will be denoted as $g_x^{log}(n,k)$ with superscript ”log”. From the physiological point of view, however, the filters should be linearly spaced on the ERB scale, which is only approximately logarithmic and results only in an approximate translation in the log-frequency domain for a linear frequency warping. This different spacing will be labeled by the superscript ERB in the corresponding representation $g_x^{ERB}(n,k)$. Finally, we also used a MEL spacing, labeled by the superscript MEL in the corresponding representation $g_x^{MEL}(n,k)$.

Given a representation $g_x(n,k)$, the final primary representation is then computed as in (2):

$$y_x(n,k) = \sum_\ell h(\ell)\,|g_x(nL-\ell,k)| \tag{7}$$

In contrast to the linear-phase wavelet filters, the gammatone filters are closer to being minimum phase. This implies that, at the same bandwidth and center frequency, onsets will be better represented with the gammatone filters, but at the cost of longer tails in the responses. Fig. 2 illustrates this effect for the same signal as in Fig. 1.

3. WARPING-INVARIANT FEATURES

Due to the nature of $y_x(n,k)$, warping-invariant features can be easily generated by taking the Fourier transform of $y_x(n,k)$ with respect to parameter k and retaining only the magnitudes of the transform coefficients. However, this is only one of several possibilities to obtain warping-invariant features. Any feature extraction strategy that is independent of a translation with respect to k will allow us to achieve this goal.

Other possibilities include, but are not limited to correlation sequences between transform values or nonlinear functions thereof at two time instances n and $n-d$ (correlation with respect to the log-frequency index k). In particular, we here consider

$$r_x(n,d,m) = \sum_k y_x(n,k) y_x(n-d,k+m) \tag{8}$$

and

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

The parameter d is a time lag, and m is the lag for the log-frequency index k. The features $r_x(n,0,m)$ will give information on the signal spectrum in time frame n. For $d \neq 0$ the features $r_x(n,d,m)$ will give information on the development of short-time spectra over time. A feature vector for time index n can contain any collection of the above mentioned features computed for the same time index n. Moreover, 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. 3 shows the features for the signal of Fig. 1. 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 formant-related structures will indeed be independent of the warping factor. For real speech, of course, this is only an approximation [11], but it leads to formant-like structures that are robust to vocal tract length variations.

4. EXPERIMENTAL 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 gammatone filterbank, three approaches were examined:

- 90 logarithmically spaced center frequencies
- 90 ERB spaced center frequencies
- 90 MEL spaced center frequencies

Center frequencies were considered in the range of 40-6700Hz, each with a bandwidth of one ERB. The lowpass filter $h(n)$ was a rectangular window of 200 coefficients.

The original speech signals were sampled at 16 kHz sampling rate, and the final frame rate was set to 10 ms. The following 45 vocal-tract length invariant features (VTLI-F) were used:

- the first 20 coefficients of the discrete cosine transform (DCT) of $\log(r(n, 0, m))$ with respect to parameter m for $m = 0, 1, \ldots, 83$.
- the first 20 coefficients of the DCT of $c(n, 4, m)$ with respect to parameter m with $m = -83, \ldots, 83$.
- $\log(r(n, 4, m))$ 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). Moreover, the first 15 DCT coefficients (DCT with respect to frequency parameter k) of the logarithmized wavelet features $\log(y_x(n, k))$ were used for feature set amendment as well. Finally, for all features, also the delta and delta-delta coefficients were included. Altogether, this makes a total number of 219 features. In a subsequent step, the number of features was reduced, using either feature selection or a linear discriminant analysis (LDA) [12].

The following feature sets were considered, where the factor 3 stands for the inclusion of delta and delta-delta features:

- 3×13 MFCC
- All 219 features, reduced via an LDA to 47 features. In each case, it has been indicated which filterbank and frequency spacing was used. We have
  - WT wavelet-transform
  - GT^{log} log-spaced gammatone filter
  - GT^{ERB} ERB-spaced gammatone filterbank
  - GT^{MEL} MEL-spaced gammatone filterbank
- 3×13 MFCC + 8 VTLI-F. These are the MFCCs, amended with the 8 most important features from the plain VTLI-F setting above.
- 3×13 MFCC + 3×5 VTLI-F. These are the MFCCs, amended with first five DCT coefficients of $\log(r(n, 0, m))$ with respect to the frequency lag m.

Table 1. Accuracies in % for phoneme recognition using a HMM recognizer with eight mixtures and diagonal covariance matrices.

Features	Train.	Test	Acc.
3×13 MFCC	M+F	M+F	69.19
VTLI-WT-F+MFCC+WT	M+F	M+F	67.84
VTLI-GT^{log}F+MFCC+GT^{log}	M+F	M+F	68.15
VTLI-GT^{ERB}F+MFCC+GT^{ERB}	M+F	M+F	68.82
VTLI-GT^{MEL}F+MFCC+GT^{MEL}	M+F	M+F	68.45
3×13 MFCC + 8 VTLI-WT-F	M+F	M+F	**70.74**
3×13 MFCC + 8 VTLI-GT^{ERB}-F	M+F	M+F	70.38
3×13 MFCC + 3×5 VTLI-WT-F	M+F	M+F	69.33
3×13 MFCC + 3×5 VTLI-GT^{log}-F	M+F	M+F	67.69
3×13 MFCC + 3×5 VTLI-GT^{ERB}-F	M+F	M+F	68.02
3×13 MFCC + 3×5 VTLI-GT^{MEL}-F	M+F	M+F	67.96
3×13 MFCC	M	F	56.84
VTLI-WT-F+MFCC+WT	M	F	63.56
VTLI-GT^{log}F+MFCC+GT^{log}	M	F	62.49
VTLI-GT^{ERB}F+MFCC+GT^{ERB}	M	F	63.15
VTLI-GT^{MEL}F+MFCC+GT^{MEL}	M	F	62.22
3×13 MFCC + 8 VTLI-WT-F	M	F	62.27
3×13 MFCC + 8 VTLI-GT^{ERB}-F	M	F	61.53
3×13 MFCC + 3×5 VTLI-WT-F	M	F	59.38
3×13 MFCC + 3×5 VTLI-GT^{log}-F	M	F	58.47
3×13 MFCC + 3×5 VTLI-GT^{ERB}-F	M	F	59.76
3×13 MFCC + 3×5 VTLI-GT^{MEL}-F	M	F	59.04
3×13 MFCC	F	M	55.53
VTLI-WT-F+MFCC+WT	F	M	62.98
VTLI-GT^{log}F+MFCC+GT^{log}	F	M	62.15
VTLI-GT^{ERB}F+MFCC+GT^{ERB}	F	M	63.00
VTLI-GT^{MEL}F+MFCC+GT^{MEL}	F	M	62.61
3×13 MFCC + 8 VTLI-WT-F	F	M	60.79
3×13 MFCC + 8 VTLI-GT^{ERB}-F	F	M	60.09
3×13 MFCC + 3×5 VTLI-WT-F	F	M	59.13
3×13 MFCC + 3×5 VTLI-GT^{log}-F	F	M	57.48
3×13 MFCC + 3×5 VTLI-GT^{ERB}-F	F	M	58.49
3×13 MFCC + 3×5 VTLI-GT^{MEL}-F	F	M	57.75

We present results for phoneme recognition on the TIMIT corpus (including the SA files). 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 [13], 48 phonetic models were trained, and the classification/recognition results were folded to yield 39 final phoneme classes that had to be distinguished. The LDA was based on the 48 phonetic classes.

Table 1 contains the results for HMM-based phoneme recognition using monophone models, three states per phoneme, eight Gaussian mixtures per state, and diagonal covariance matrices. The recognizer was based the Hidden-Markov-Toolkit (HTK).

For the M+F setting, where both male and female data was used during training and test, we see that all examined feature sets yield almost the same performance as the MFCCs, best performance is achieved by the 3×13 MFCC + 8 VTLI-WT-F setup. However, when only male or only female data is used for training, the degradation for the linear-phase wavelet based feature sets as well as for the gammatone based feature sets are far less than for the MFCCs. In contrast to the M+F setting and albeit the nature of preprocessing, the best performances are achieved when VTLI features, preprocessing features and MFCCs are combined via an LDA to a final number of 47 features. This combined feature set is also the most robust one when the training and test conditions are different. A closer examination of these results for the different preprocessing steps shows that only the incorporation of all mentioned audiology aspects can slightly enhance the detection rates. Using the presented approach incorporating both ERB-based bandwidth and ERB-based frequency scaling (GT^{ERB}) best recognition rates were achieved although the center frequencies are not strictly logarithmically spaced. Interestingly, the GT^{log} case leads to lowest recognition rates of all three approaches. The MEL spacing performs slightly better than the logarithmic one, but it cannot reach the performance obtained with the ERB scale.

As the results show, for the GT^{ERB} feature set, the accuracy for the M+F condition is nearly the same as for the MFCCs, and at the same time, it is significantly better for all other conditions: When training on male and testing female data, the accuracy is about 6% better than for MFCCs. When training on female and testing male data, it is even 7.5% better than for MFCCs.

5. CONCLUSIONS

We have proposed a technique for the extraction of vocal tract length invariant features with an auditory-filterbank based preprocessing. The performance of the new features has been demonstrated for phoneme recognition tasks. The results have shown that the incorporation of knowledge about the human auditory system can lead to an enhancement of recognition rates and to more robustness. The optimal choice of the primary frequency analysis and the best feature selection depend on the task at hand and are still open issues.

6. 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] 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] S. Umesh, L. Cohen, and D. Nelson, "Frequency warping and the Mel scale," *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 104–107, Mar. 2001.

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

[7] A. Mertins and J. Rademacher, "Frequency-warping invariant features for automatic speech recognition," in *ICASSP*, accepted 2006.

[8] 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.

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

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

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

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

[13] Kai-Fu Lee and Hsiao-Wuen Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, no. 2, pp. 1641 – 1648, Nov. 1989.