COMPUTING MEL-FREQUENCY CEPSTRAL COEFFICIENTS ON THE POWER SPECTRUM

Sirko Molau, Michael Pitz, Ralf Schlüter, and Hermann Ney

Lehrstuhl für Informatik VI, Computer Science Department,
RWTH Aachen – University of Technology, 52056 Aachen, Germany
{molau, pitz, schlueter, ney}@informatik.rwth-aachen.de

ABSTRACT

In this paper we present a method to derive Mel-frequency cepstral coefficients directly from the power spectrum of a speech signal. We show that omitting the filterbank in signal analysis does not affect the word error rate. The presented approach simplifies the speech recognizer's front end by merging subsequent signal analysis steps into a single one. It avoids possible interpolation and discretization problems and results in a compact implementation. We show that frequency warping schemes like vocal tract normalization (VTN) can be integrated easily in our concept without additional computational efforts. Recognition test results obtained with the RWTH large vocabulary speech recognition system are presented for two different corpora: The German VerbMobil II dev99 corpus, and the English North American Business News 94 20k development corpus.

1. INTRODUCTION

Most of today's automatic speech recognition (ASR) systems are based on some type of Mel-frequency cepstral coefficients (MFCCs), which have proven to be effective and robust under various conditions. This paper describes an alternative concept to derive MFCCs directly from the power spectrum of the speech signal. A number of subsequent steps of the traditional signal analysis are integrated into the cepstrum transformation, which avoids possible discretization and interpolation errors. The new concept yields equally good recognition performance without a filterbank, thus reduces the number of parameters that need to be optimized.

The remainder of this paper is organized as follows: In the next section we will briefly recapitulate the typical signal analysis procedure. Then we discuss in detail implementational issues of the traditional MFCC computation and present our integrated approach. In section 4 we will demonstrate that frequency warping schemes like VTN can be easily integrated as well. Finally, we will present recognition test results for the VerbMobil II and the North American Business News Corpus, and draw the conclusions of our work.

2. SIGNAL ANALYSIS

Figure 1 shows the signal analysis front end of a typical ASR system. The speech waveform, sampled at 8 or 16 kHz, is first differentiated (preemphasis) and cut into a number of overlapping segments (windowing), each 25 ms long and shifted by 10 ms. A Hamming window is multiplied and the Fourier transform (FFT) is computed for each frame. The power spectrum is warped according to the Mel-scale in order to adapt the frequency resolution to the properties of the human ear. Then the spectrum is segmented into a number of critical bands by means of a filterbank. The filterbank typically consists of overlapping triangular filters. A discrete cosine transformation (DCT) applied to the logarithm of the filterbank outputs results in the raw

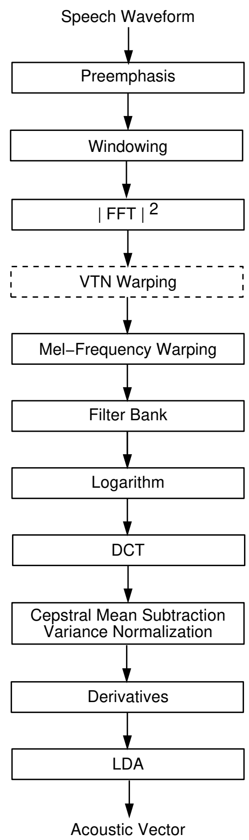

Figure 1: Typical signal analysis front end

MFCC vector. The highest cepstral coefficients are omitted to smooth the cepstra and minimize the influence of the pitch which is irrelevant for the speech recognition process. The mean of each cepstral component is subtracted, and the variance of each component may also be normalized. Finally, the MFCC vector is augmented with time derivatives. Additional transformations like linear discriminant analysis (LDA) may further increase the temporal context and the discriminance of the acoustic vector. As a result signal analysis provides every 10 ms an acoustic vector, which is typically of dimension 25 to 50.

3. COMPUTATION OF MFCCS

We now want to have a closer look at the computation of cepstral coefficients from speech spectra, i.e. the signal analysis steps between FFT and DCT. We will discuss problems of different implementations, and finally present a method to compute MFCCs directly on the power spectrum. Both the traditional and the integrated approach suggested here are depicted in Figure 2.

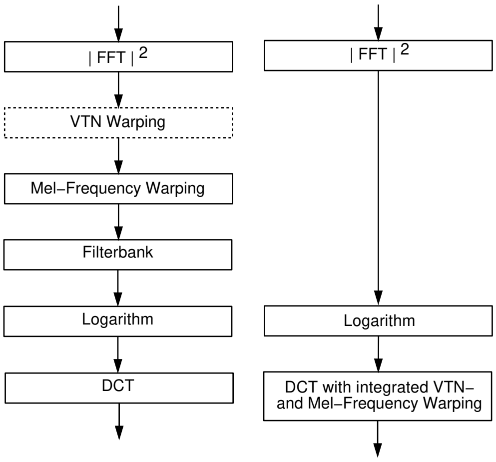

Figure 2: Comparison of the traditional MFCC computation (left) with the integrated approach (right) investigated here.

3.1. Traditional Filterbank Approach

Mel-frequency warping and the filterbank can be implemented easily in the frequency domain (see Figure 3). One method is to transform the power spectrum, i.e. to compute a Mel-warped spectrum by interpolation from the original discrete-frequency power spectrum. The advantage is that the following triangular filters all have the same shape and can be placed uniformly at the Mel-warped spectrum. On the other hand, the discretization may be especially critical due to the large dynamic range of the power spectrum.

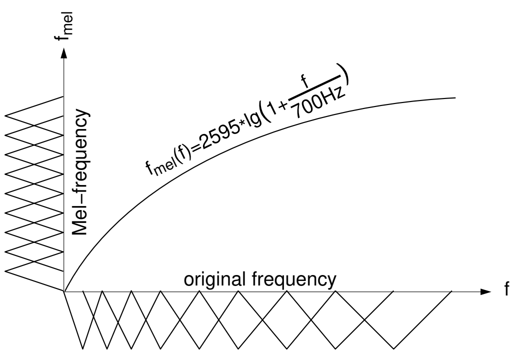

Figure 3: Schematic plot of different triangular filterbank implementations. The filters are either uniformly distributed at the Mel-warped spectrum, or non uniformly at the original spectrum. In the latter case, they should be asymmetric as well.

Another way is to place the triangular filters non uniformly at the unwarped spectrum and thereby implicitely incorporate Mel-frequency scaling [1]. However, discretization errors may then occur if the spectral resolution is not appropriate. The lowest filters could be placed at a very few spectral lines only, and the maximum of one of the filters may fall just inbetween two spectral lines. In addition, the filters should not be triangular and symmetric anymore, but bend according to the shape of the Mel-function at the position of the filter.

Last but not least it is not clear how many filters are required and which filter shape is optimal. Triangular filters are occasionally replaced by trapezoidal or more complex shaped ones derived from auditory models, and we sometimes observed better word error rates when using filters with cosine shape.

In all cases the logarithm of the filterbank output is cosine transformed to obtain MFCCs.

3.2. Computing MFCCs Directly On The Power Spectrum

We have investigated an alternative method to compute Mel-frequency warped cepstral coefficients directly on the power spectrum and thereby avoid possible problems of the standard approach.

Ignoring any spectral warping for a moment, cepstral coefficients c_k can be derived by Eq. (1):

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega \, \lg |X(e^{j\omega})| \cdot e^{j\omega k} \tag{1}$$

Depending on whether a filterbank is used or not, $|X(\cdot)|$ stands for either the filterbank outputs or the power spectrum.

The sequential application of a monotone invertible frequency warping function $g : [-\pi, \pi] \to [-\pi, \pi]$ and DCT can be expressed as follows:

$$\begin{aligned} \omega &\to \tilde{\omega} = g(\omega) \\ c_k &= \frac{1}{2\pi} \int_{-\pi}^{\pi} d\tilde{\omega} \, \lg |X(e^{jg^{-1}(\tilde{\omega})})| \cdot e^{j\tilde{\omega}k} \end{aligned} \tag{2}$$

To incorporate warping directly into the cosine transformation, we change the integration variable and use the derivative of the warping function $d\tilde{\omega}/d\omega$ (Eq. 3). The continuous integral is later approximated in the standard way by a discrete sum (Eq. 4):

$$c_k = \frac{1}{2\pi} \int_{-\pi}^{\pi} d\omega \, \lg |X(e^{j\omega})| \cdot e^{jg(\omega)k} \cdot g'(\omega) \tag{3}$$

$$\cong \frac{1}{N} \sum_{n=0}^{\frac{N}{2}-1} \left\{ \lg |X(e^{j\frac{2\pi n}{N}})| \cdot \cos[g\left(\frac{2\pi n}{N}\right) k] \cdot g'\left(\frac{2\pi n}{N}\right) \right\} \tag{4}$$

One specific type of frequency warping is the Mel-frequency scaling $\mu(\cdot)$, which is usually carried out according to formula (5) with the sampling frequency f_s [6]:

$$\mu(\omega) = 2595 \cdot \lg\left(1 + \frac{\omega f_s}{2\pi \cdot 700Hz}\right) \tag{5}$$

For integration into the cosine transformation, the Mel-warping function needs to be normalized in order to meet the criterion $\tilde{\mu}(\pi) = \pi$.

$$\begin{aligned} \tilde{\mu}(\omega) &= \frac{\pi}{\mu(\pi)} \cdot \mu(\omega) \\ &= d \cdot \lg\left(1 + \frac{\omega f_s}{2\pi \cdot 700Hz}\right) \end{aligned} \tag{6}$$

with

$$d = \frac{\pi}{\lg\left(1 + \frac{f_s}{2 \cdot 700Hz}\right)}.$$

Replacing $g(\cdot)$ in Eq. (4) by $\tilde{\mu}(\cdot)$ leads to a compact implementation of MFCC computation with only a few lines of code. A look-up table for constants like the derivative and the cosine term can be precomputed, all that remains is a matrix multiplication on the logarithm of the power spectrum. Figure 4 shows the effect of the modified signal analysis on two cepstrum coefficients for a test sentence from the VerbMobil II corpus. Whereas the lower order coefficients are almost identical, the difference increases with higher coefficient orders due to the discarted filterbank.

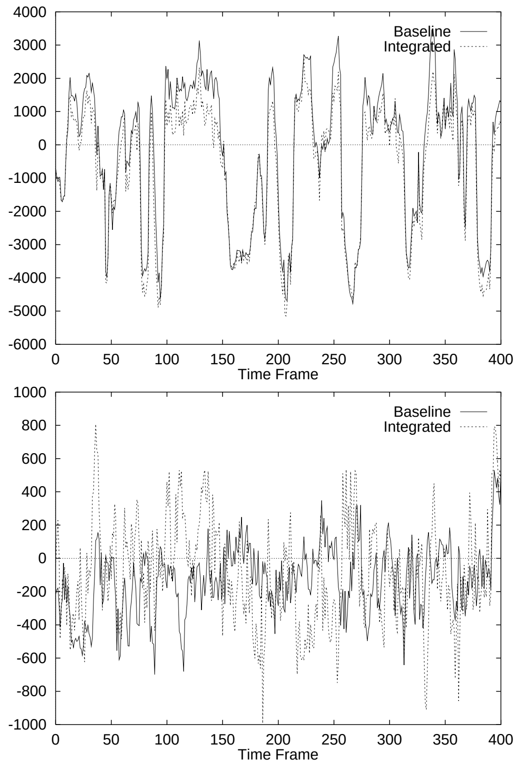

Figure 4: Comparison of cepstrum coefficients 1 (upper curve) and 15 (lower curve) for a test sentence from the VerbMobil II corpus (baseline: traditional filterbank approach; integrated: DCT with integrated Mel-frequency warping).

4. INTEGRATION OF VTN

Vocal tract length normalization (VTN) is a speaker normalization scheme that also relies on warping the power spectrum. The idea is to compensate for the shift of formants in speech spectra caused by the speaker-specific length of the vocal tract.

It has been shown before that one possible VTN implementation is to modify the location of filters in the filterbank just as for Mel-frequency scaling [2]. From what we have presented in the previous section it is clear, however, that VTN can also be fully integrated into the cepstrum transformation.

The VTN warping function $\nu_\alpha : [0, \pi] \rightarrow [0, \pi]$ needs to be monotone and invertible as well. A simple choice is a piece-wise linear warping function as shown in Figure 5. The inflexion frequency ω_0 at which the slope of the function changes depends on α:

$$\omega_0 = \begin{cases} \frac{7}{8}\pi & \alpha \leq 1 \\ \frac{7}{8 \cdot \alpha}\pi & \alpha > 1 \end{cases}$$

In order to avoid complicated case distinctions for different warping factors and frequencies, we write the warping function $\omega \rightarrow \nu_\alpha(\omega)$ in the following convenient form

$$\nu_\alpha(\omega) = \beta_\omega \omega + \gamma_\omega \tag{7}$$

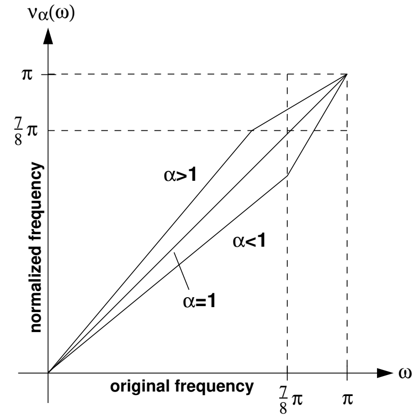

Figure 5: Warping function for piece-wise linear VTN

with parameters β_ω and γ_ω. Although these parameters formally depend on ω, they can take on only two values:

$$\beta_\omega = \begin{cases} \alpha & \omega \leq \omega_0 \\ \frac{\pi - \alpha\omega_0}{\pi - \omega_0} & \omega > \omega_0 \end{cases}$$

$$\gamma_\omega = \begin{cases} 0 & \omega \leq \omega_0 \\ (\alpha - 1) \cdot \frac{\pi \cdot \omega_0}{\pi - \omega_0} & \omega > \omega_0 \end{cases}$$

Mel-warping is applied after the spectra are scaled according to VTN. Hence, the combination $\chi(\cdot)$ of Mel- and VTN warping becomes

$$\begin{aligned} \chi(\omega) &= \tilde{\mu}(\nu_\alpha(\omega)) \\ &= d \cdot \lg\left(1 + \frac{[\beta_\omega \omega + \gamma_\omega] \cdot f_s}{2\pi \cdot 700Hz}\right) \end{aligned} \tag{8}$$

with the derivative:

$$\chi'(\omega) = \frac{d \cdot \beta_\omega \cdot f_s}{(2\pi \cdot 700Hz + [\beta_\omega \omega + \gamma_\omega] \cdot f_s) \cdot \ln(10)} \tag{9}$$

Cepstrum coefficients with integrated VTN and Mel-frequency warping are obtained by replacing $g(\cdot)$ in Eq. (4) by $\chi(\cdot)$.

5. RECOGNITION TESTS

To evaluate the proposed signal analysis approach, we performed recognition tests with the RWTH large vocabulary speech recognition system (see [3] , [4], and [5] for detailed system descriptions) on two different corpora. The VerbMobil II task (VM II) is German spontaneous speech with a 10k-word vocabulary, and the North American Business News task (NAB) is clean read speech of Wall Street Journal texts with a recognition vocabulary of 20k. Details of the training and test corpora are given in Table 1.

Table 1: Statistics of the training and test corpora

Corpus	VerbMobil II		Wall Street Journal	
	Training CD1-41	Test DEV99	Training WSJ0+1	Test DEV-94
Duration	61.5h	1.6h	81.4h	0.8h
Sil. Portion	13%	11%	27%	19%
# Speakers	857	16	284	20
# Sent.	36,015	1,081	37,571	310
# Words	701,512	14,662	649,624	7,378
Trigram PP.	-	62.0	-	126.6

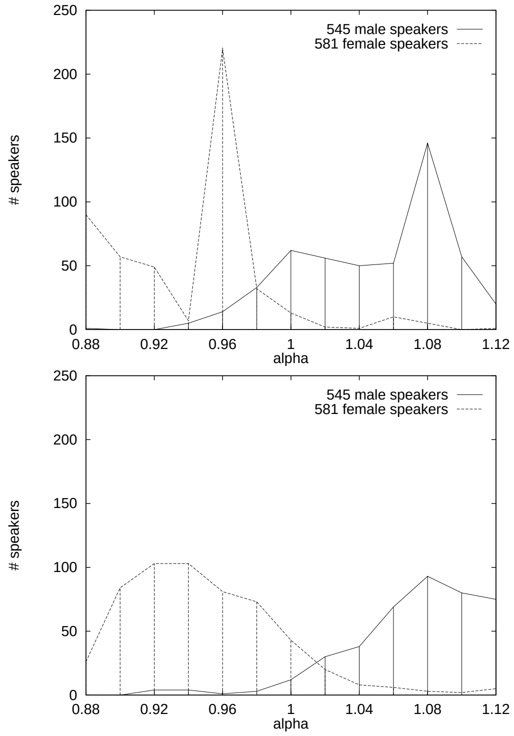

Figure 6: Warping factor distribution of the VM II training speakers. The upper histogram was obtained with Mel- and VTN warped spectra obtained by linear interpolation, the lower histogram with integrated Mel-frequency and VTN warping.

The first result of using the integrated approach in VTN training was a much smoother distribution of warping factors. Figure 6 shows the corresponding histograms for the VM II training corpus. A closer inspection revealed that linear interpolation of spectral lines when transforming the power spectrum for VTN warping was the main reason for the erratic distribution observed before. It turned out, however, that the word error rate (WER) was only marginally affected by this difference.

Next, we compared the recognition performance of the traditional signal analysis approach (baseline) with the integrated MFCC computation. Additional tests were carried out with two-pass and fast VTN as described in [4]. The best results of each setup are summarized in Table 2.

Table 2: Recognition test results for the VM II and the NAB corpus applying no VTN, two-pass, and fast VTN (baseline: traditional filterbank approach; integrated: DCT with integrated frequency warping).

Corp.	VTN	Cepstrum	#Dns [k]	Overall [%] Del - Ins	WER
VM II	no	Baseline	455	4.9 - 4.8	25.7
		Integrated	457	5.0 - 4.4	25.3
	2-Pass	Baseline	450	4.4 - 4.3	23.8
		Integrated	451	4.9 - 4.1	24.0
	Fast	Baseline	450	4.5 - 4.5	23.8
		Integrated	451	5.0 - 4.1	24.0
NAB	no	Baseline	596	1.5 - 2.3	12.5
		Integrated	599	1.5 - 2.3	12.4
	2-Pass	Baseline	563	1.4 - 2.4	11.8
		Integrated	591	1.4 - 2.2	11.7
	Fast	Baseline	563	1.4 - 2.3	11.9
		Integrated	591	1.5 - 2.2	11.8

We found that the recognition performance of both methods is similar. In most cases the integrated approach performed almost as good or slightly better than the traditional sequential analysis with a filterbank. A similar behaviour was found on smaller German and Italian telephone speech corpora (VerbMobil and EuTrans).

6. CONCLUSIONS

In this paper we have presented an alternative signal analysis approach that merges a number of subsequent analysis step into one. Omitting the filterbank and integrating Mel-frequency warping into the cepstrum transformation simplifies the signal analysis (no filterbank parameters need to be optimized), avoids possible interpolation and discretization problems, and leads to a compact implementation of the MFCC front end. We have shown that concepts like VTN that rely on warping speech spectra can be easily integrated as well. Recognition tests on the VerbMobil II and the North American Business News corpus revealed that the new approach performs as good as the traditional signal analysis.

7. REFERENCES

[1] S. B. Davis and P. Mermelstein: "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", *IEEE Transactions on Acoustic, Speech, and Signal Processing*, Vol. 28, No. 4, August 1980, pp. 357–366.

[2] L. Lee and R. Rose: "Speaker normalization using efficient frequency warping procedures", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, Mai 1996, pp. 353–356.

[3] H. Ney, L. Welling, S. Ortmanns, K. Beulen, and F. Wessel: "The RWTH Large Vocabulary Continuous Speech Recognition System", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, pp. 853–856.

[4] A. Sixtus, S. Molau, S. Kanthak, R. Schlüter, and H. Ney: "Recent Improvements of the RWTH Large Vocabulary Speech Recognition System on Spontaneous Speech", *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000, pp. 1671–1674.

[5] L. Welling, N. Haberland, and H. Ney: "Acoustic Front-End Optimization for Large Vocabulary Speech Recognition", *Proc. EUROSPEECH*, Rhodes, Greece, September 1997, pp. 2099–2102.

[6] S. J. Young: "HTK: hidden markov model toolkit V1.4", User Manual, Cambridge University Engineering Department, Cambridge, England, February 1993.