INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

An alternative to MFCCs for ASR

Pegah Ghahramani[1,*], *Hossein Hadian*[2], *Daniel Povey*[3], *Hynek Hermansky*[1,4], *Sanjeev Khudanpur*[1,4]

[1]Center for Language and Speech Processing, Johns Hopkins University, Baltimore, USA
[2]Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
[3]Xiaomi Corp., Beijing, China
[4]Human Language Technology Center of Excellence, Johns Hopkins University, Baltimore, USA

pegahgh@gmail.com

Abstract

The Mel scale is the most commonly used frequency warping function to extract features for automatic speech recognition (ASR) and is known to be quite effective. However, it is not specifically designed for ASR acoustic models based on deep neural networks (DNN). In this study, we introduce a frequency warping function which is a modified version of Mel scale. This warping function is parameterized using 2 parameters and we use it to propose a new set of features called modified Mel-frequency cepstral coefficients (MFCC), which use cosine-shaped filters. The bandwidths are computed using a new function. By evaluating the proposed features on a variety of ASR data sets, we see consistent improvements over regular MFCCs and (log) Mel filter bank energies.

Index Terms— feature extraction, ASR, Modified Mel

1. Introduction

The Mel scale is a perceptual scale of frequencies which is hand crafted based on physiological models of the human auditory system. This scale is used in the MFCC and log-Mel features, which are perhaps the most commonly used features for ASR. However, they are not guaranteed to work well with the latest ASR models which are all based on DNN.

One approach to extracting features that are more suited to DNNs, is to train the ASR model from the signal domain and let the network craft its own features in a data-driven scheme. This is also known as joint feature extraction and acoustic modeling and has been investigated in a few studies [1, 2, 3, 4].

In particular, in a previous study, we proposed a data-driven feature learning layer (from frequency domain) that can be trained jointly with ASR [5]. We used that to learn new filter banks, outperforming the MFCC-based models. The main drawback was that the network learned data dependent filter banks and could over-fit to the training data.

To address this issue, we proposed a new analytic filter bank which we estimated using the learned data-driven filter banks. We successfully obtained similar results as the learned filter bank using the analytic filters. However, the warping function used in the analytic filter bank is not invertible and cannot be easily scaled to all sampling frequencies (e.g., wide-band).

In this study, we improve upon that work and propose a new alternative for MFCC features: Modified MFCC. Similar to our previous work, these features are based on the filter banks directly learned on different narrow-band and wide-band datasets using the joint feature learning setup proposed in [5].

* work done before the author joining Amazon

At the core of modified MFCC features is an invertible frequency warping function that is used for finding the center frequencies for the filters. This warping function is analogous to the Mel scale. The bandwidths for filters are computed using a combination of two functions: a new function of the center frequencies and a function of the filters' overlap.

The main baselines in this paper are MFCC and Mel filter bank features which are well known and both use the Mel scale [6]. The bark scale is another well-known frequency warping function, which is based on critical bandwidth values [7]. A novel warping function based on high-energy portions of speech signals is proposed in [8]. Other data-driven approaches comparable to our work include [9] which uses linear discriminant analysis to maximize the separability between linguistic classes and [10] which uses discriminative feature extraction for designing warping functions.

The rest of this paper is as follows. In Section 2, we describe the proposed modified MFCC method. The experimental results are presented in Section 3. Finally, the conclusions will appear in Section 4.

2. Proposed Modified MFCCs

In this section, we explain the proposed modified MFCC features. Specifically, we describe the filter shape, the frequency warping function, and finally how we calculate the filter bandwidths. The filters used in modified MFCC are cosine-shaped. The formula used for filter estimation is shown in Equation 1, where each filter is specified using a center frequency f_c and a bandwidth w.

$$\begin{cases} \cos(\frac{\pi(x-f_c)}{w}) & f_c - \frac{w}{2} \leq x \leq f_c + \frac{w}{2} \\ 0 & else \end{cases} \quad (1)$$

It can be shown mathematically that the proposed cosine-shaped filters have the bandwidth of $\frac{w}{2}$ according to the noise-equivalent formula.

2.1. Center frequency approximation

In MFCC, the center frequencies are calculated using the Mel scale, which is basically a frequency warping function. Here we propose a new warping function based on filters learned in the filter bank layer in DNN.

Figure 1 shows center frequency distributions for filter banks learned on Switchboard, AMI-IHM, 8kHz and 16kHz multi-en datasets. We combine multiple English corpora, WSJ [11], Switchboard [12], HUB4, TedLium and Fisher [13]. Switchboard and Fisher are 8kHz and WSJ, TedLium and

http://dx.doi.org/10.21437/Interspeech.2020-2690

HUB4 are 16kHz datasets. To handle different sampling rates for different corpora, all datasets are down-sampled to 8kHz for narrow-band experiments and up-sampled to 16kHz for wide-band experiments.

As shown in Figure 1, the center frequency distribution probability is higher around the average of the first and second formant frequencies. The original Mel scale function has a higher distribution around 700 Hz and the center frequency distribution gradually decreases after 700 Hz.

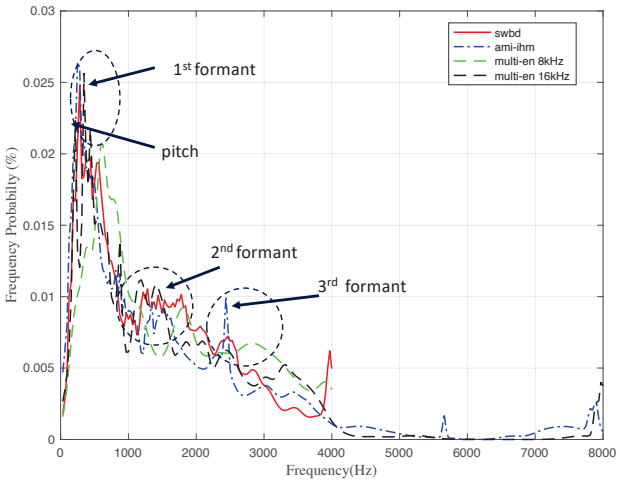

Figure 1: *Center frequency distribution for different datasets*

Equation 2 shows the proposed warping function for modified MFCC. f_{b_1} and f_{b_2} are the adjustable parameters in this function, which control the first and second peaks in the center frequency distribution.

$$g(f) = \log\left(f_{b_1} + f_{b_2}\log(1 + \frac{f}{f_{b_2}})\right) \quad (2)$$
$$f_{b_1} \in [300, 900]\, Hz$$
$$f_{b_2} \in [1500, 3500]\, Hz$$

Figure 2 compares center frequencies for original and modified Mel warping functions. f_{b_1} and f_{b2} used in this figure are 300 Hz and 1500 Hz, respectively.

2.2. Bandwidth approximation

We propose two different approaches for approximating the bandwidths. In the first approach, we use a fixed function to compute the bandwidths while in the second approach, we let the DNN learn the bandwidths. These are described in detail in the following subsections.

2.2.1. Fixed bandwidth approximation

In our previous work [5], the bandwidth is estimated using a piece-wise linear function, that is approximated using filters learned in the filter bank layer in DNN. In that method, the bandwidth does not depend on the number of filters. The main drawback is that some frequency regions, especially in the high-frequency sub-bands, may not be covered by any filters.

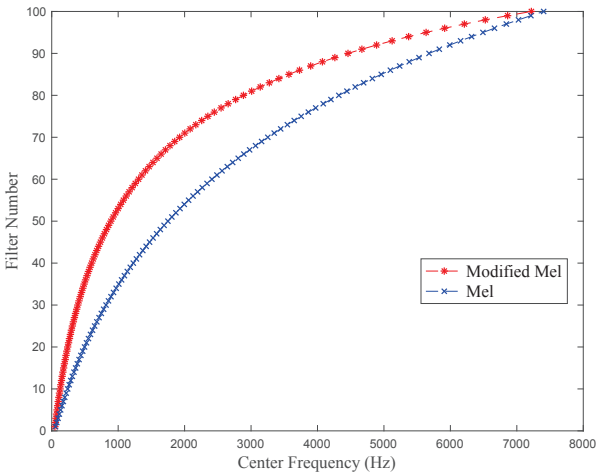

Figure 2: *Center frequency vs. filter index for Mel and Modified Mel ($f_{b_1} = 300$ Hz, $f_{b_2} = 1500$ Hz in modified Mel)*

Another approach for filter bandwidth estimation would be to enforce specific overlaps between neighboring filters. Using this approach, the bandwidth would strongly depend on the number of filters. However, [5] shows this behavior is not seen in low-frequency sub-bands for filters learned in the filter bank layer in DNN and the bandwidth in some sub-band regions (e.g., in proximity of 1^{st} and 2^{nd} formant frequencies) are independent of the number of filter banks. In this study, the goal is to combine the benefits of both methods in low and high-frequency sub-bands and combine the bandwidth using a bandwidth estimation formula. Equation 3 shows the formula used to approximate bandwidth in the modified Mel filter banks. The bandwidth $bw(i)$ for filter i is a combination of $bw_{lin}(i)$ and $bw_{op}(i)$, where $bw_{lin}(i)$ is estimated using a linear function and is not a function of number of filters (i.e., $bw_{min} \leq bw_{lin}(i) \leq bw_{min} + s_{bw}$). bw_{min} is minimum bandwidth for linear bandwidth and $bw_{min} + s_{bw}$ is the maximum value. Overlap-based bandwidth, $bw_{op}(i)$, is computed based on an overlap op to satisfy a minimum overlap between adjacent filters.

$$bw_{lin}(i) = bw_{min} + s_{bw}(\frac{f_c(i)}{f_c(i) + f_{b_1}}) \quad (3)$$
$$bw_{min} \in [30, 100]\, Hz$$
$$s_{bw} \in [30, 100]\, Hz$$
$$bw_{op}(i) = (f_c(i) - f_c(i-1))(1 + op)$$
$$op \in [0.0, 0.5]$$
$$bw(i) = \sqrt{bw_{lin}(i)^2 + bw_{op}(i)^2}$$

2.2.2. Trainable DNN-based bandwidth approximation

As an alternative approach, we propose to use the DNN to learn the bandwidths in the cosine filter banks, instead of estimating them beforehand as explained in the previous section. We use the frequency-domain setup proposed in [5] with parametric filters in the filter bank layer. The filters in the filter bank layer

(i.e., Equation 4) are cosine-shaped with fixed center frequencies, set using the proposed formula in Section 2.1. Therefore, the filter bandwidths $w^{'}$ are the only parameters in the cosine-shaped filters that are learned using the DNN. It should be noted that [14] also uses parametric sinc function (i.e., SincNet) to implement band-pass filters in speaker recognition from raw waveform.

Square function is used to impose positivity for $w^{'}$, where $w^{'} = w^2$ and the parameter w is computed using DNN. Also, a single layer with per-filter scale parameter is also added after the filter bank layer to learn the scaling values for each filter.

$$\begin{cases} \cos(\frac{\pi(x-f_c)}{w^{'}}) & f_c - \frac{w^{'}}{2} \leq x \leq f_c + \frac{w^{'}}{2} \\ 0 & else \end{cases} \quad (4)$$

3. Results

3.1. Effect of different parameters on modified MFCC

3.1.1. Effect of f_{b_1} and f_{b_2} in center frequency approximation

Table 1 shows the effect of two cutoffs f_{b_1} and f_{b_2} in modified Mel warping function on Switchboard, AMI-SDM [15]. The bandwidth in the filter banks is approximated using Equation 3, where bw_{min} and s_{bw} are at 30 and 60 Hz. Log and DCT transform are applied on the modified Mel filter banks and modified MFCC (80-dim) are used in all experiments.

Table 1: *Effect of f_{b_1} and f_{b_2}*

(f_{b_1}, f_{b_2})	SWBD		AMI-SDM
	rt03	eval2000	eval
$(300, 1500)$	**17.3**	14.6	40.5
$(500, 1500)$	17.4	14.6	40.6
$(300, 2000)$	17.4	14.6	40.5
$(300, 2000)^1$	17.6	14.6	40.6
$(300, 3500)$	17.5	14.5	40.5
$(500, 3500)$	17.4	**14.5**	40.5
$(900, 3500)$	17.4	14.7	**40.3**

1: 40-dim modified MFCC

3.1.2. Effect of bw_{min}, s_{bw} and overlap for bandwidth approximation

Table 2 compares the results using different values of bw_{min} and s_{bw}. The result shows that WER is less sensitive to these parameters. In this table, f_{b_1} and f_{b_2}, at 300 and 1500 Hz, are used in all experiments.

Table 2: *Effect of bw_{min} and s_{bw} for bandwidth approximation in modified Mel filter banks*

(bw_{min}, s_{bw})	SWBD		AMI-SDM
	eval2000	rt03	eval
$(30, 60)$	14.6	17.3	40.5
$(80, 30)$	**14.5**	17.2	**40.3**
$(50, 50)$	14.6	17.2	40.4
$(60, 50)$	14.6	**17.1**	40.6
$(80, 100)$	14.5	17.4	40.4

We also used two different methods to combine linear bandwidth bw_{lin} and overlap-based bandwidth bw_{op}. To combine bandwidths, two functions were used: $g_1 = \sqrt{bw_{lin}^2 + bw_{op}^2}$ and $g_2 = \sqrt{bw_{lin} bw_{op}}$, where two bandwidth values bw_{lin} and bw_{op} are lower bounds for g_1 and upper and lower bounds for g_2.

Table 4 compares the results using two combination methods, and bw_{min} and s_{bw} are at 30 and 60 Hz in these experiments. As shown, g_1 produces better results especially on AMI-SDM.

Also, we experimented with different overlap values. Table 3 compares the effect of different overlap values, 0.0, 0.2 and 0.3. As can be seen, an *op* value of 0.3 results in the best performance on 300 hours Switchboard and AMI-SDM. In this table, f_{b_1} and f_{b_2} at 300 and 2000 Hz, are used in all experiments. Comparing row 1 with rows 2 and 4 shows that computing bandwidth using the linear bandwidth formula (i.e., bw_{lin} in Equation 3) results in $0.3 - 0.5\%$ absolute WER degradation, especially in the 16kHz AMI-SDM dataset. Also, comparing rows 2 and 3 and rows 4 and 5 shows that considering the linear part in bandwidth computation produces $0.1 - 0.3\%$ absolute WER improvement.

Table 3: *Effect of overlap in bandwidth computation*

op	(bw_{min}, s_{bw})	SWBD		AMI-SDM
		eval2000	rt03	eval
0	$(60, 50)$	14.9	17.5	40.9
0.2	$(60, 50)$	14.6	17.3	40.4
	$(0, 0)$	14.8	17.3	40.5
0.3	$(60, 50)$	**14.5**	17.4	**40.2**
	$(0, 0)$	14.6	17.4	40.5

Table 4: *Effect of linear and overlap-based bandwidth combination methods*

Method	SWBD		AMI-SDM
	eval2000	rt03	eval
g_1	14.6	17.4	40.5
g_2	14.7	17.5	40.8

3.2. DNN-based bandwidth approximation

In this section, the filter banks are set using the proposed method in Section 2.2.2. The center frequency for the filter banks in the filter bank layer are set using Equation 2 and they are fixed during training. The bandwidths are initialized using bw_{op} with overlap 0.2 in Equation 3 and the bandwidths are trained using DNN during training. Table 5 shows WER results on Switchboard and AMI-SDM using different f_{b_1} and f_{b_2} to approximate the center frequencies. Filter banks with 50-dim are used in the filter bank layer. The results show that the learning bandwidth using proposed parametric filter banks results in nice improvement on Switchboard and AMI-SDM. The computation cost in this approach is small and the number of parameters in the filter bank layer is equal to the number of filters (i.e., 50). The first row in Table 5 shows WER using the proposed frequency setup in [5]. Each filter in the filter bank layer in the old approach contains $\frac{N}{2}$ parameters, where N is the FFT dimension. As can be seen, 3000 and 500 as f_{b_1} and f_{b_2} offers considerable improvement compared to the proposed frequency domain setup with full filter banks. One reason is that learning the filter bank is sensitive to background noises and reverberations. Specifically, the parameters in high frequency regions are

not trained properly in the frequency-domain setup [5]. Learning filter banks using proposed DNN-based bandwidth approximation is less sensitive to different noise conditions and the parametric filters are a regularized version of full filter bank layer.

Table 5: *Effect of f_{b_1} and f_{b_2} on parametric cosine-based filter banks*

(f_{b_1}, f_{b_2})	SWBD		AMI-SDM
	eval2000	rt03	eval
Frequency-domain setup	14.5	17.0	40.2
$(300, 1500)$	14.5	17.0	40.3
$(300, 2000)$	**14.3**	16.9	40.5
$(500, 3000)$	14.4	**16.8**	**39.7**
$(500, 3500)$	14.3	17.0	39.8
$(900, 3000)$	14.4	17.1	40.1

50-dim filter banks are used.

3.3. Performance on various datasets

In this section, we evaluate the proposed modified MFCC features on various databases, namely TedLium [16], Switchboard [12], AMI-IHM and AMI-SDM [15]. The results are shown in Table 6. We used 80-dim modified MFCC features, where 300 and 1500 Hz were used as f_{b_1} and f_{b_2} in computing warping function and 80 and 30 Hz represented function g_1 with an *op* value of 0.2 used to compute bandwidth.

Table 6: *Performance on various datasets*

Database	Test set	MFCC	Modified MFCC
Switchboard	eval2000	14.8	14.5
	rt03	17.8	17.2
AMI-IHM	eval	20.3	20.2
AMI-SDM	eval	41.2	40.3

4. Conclusion

In this paper, we proposed a modified version of Mel filter banks. In the new modified MFCC, the warping function is a modified version of Mel warping function with an extra hyper parameter, which controls center frequency distribution. The filter bank's bandwidth is also computed using a combination of a linear and overlap-based bandwidth formula. We also propose a method which consists of parametric cosine filters with trainable bandwidth. The result on different datasets shows a $0.2-0.5\%$ WER improvement using new modified filter banks.

5. References

[1] Tara N Sainath, Brian Kingsbury, Abdel-rahman Mohamed, and Bhuvana Ramabhadran, "Learning filter banks within a deep neural network framework," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 297–302.

[2] Tara N Sainath, Ron J Weiss, Andrew Senior, Kevin W Wilson, and Oriol Vinyals, "Learning the speech front-end with raw waveform cldnns," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] Pegah Ghahremani, Vimal Manohar, Daniel Povey, and Sanjeev Khudanpur, "Acoustic modelling from the signal domain using CNNs.," in *Interspeech*, 2016, pp. 3434–3438.

[4] Neil Zeghidour, Nicolas Usunier, Iasonas Kokkinos, Thomas Schaiz, Gabriel Synnaeve, and Emmanuel Dupoux, "Learning filterbanks from raw speech for phone recognition," in *2018 IEEE international conference on acoustics, speech and signal Processing (ICASSP)*. IEEE, 2018, pp. 5509–5513.

[5] Pegah Ghahremani, Hossein Hadian, Hang Lv, Daniel Povey, and Sanjeev Khudanpur, "Acoustic modeling from frequency domain representations of speech," *Proc. Interspeech 2018*, pp. 1596–1600, 2018.

[6] Stanley Smith Stevens, John Volkmann, and Edwin B Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[7] Eberhard Zwicker, "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. 33, no. 2, pp. 248–248, 1961.

[8] Kuldip Paliwal, Benjamin Shannon, James Lyons, and Kamil Wójcicki, "Speech-signal-based frequency warping," *IEEE signal processing letters*, vol. 16, no. 4, pp. 319–322, 2009.

[9] Lukáš Burget and Hynek Heřmanskỳ, "Data driven design of filter bank for speech recognition," in *International Conference on Text, Speech and Dialogue*. Springer, 2001.

[10] Alain Biem and Shigeru Katagiri, "Filter bank design based on discriminative feature extraction," in *Proceedings of ICASSP'94. IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 1994, vol. 1, pp. I–485.

[11] Douglas B Paul and Janet M Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[12] John J Godfrey, Edward C Holliman, and Jane McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*. IEEE Computer Society, 1992, vol. 1, pp. 517–520.

[13] Christopher Cieri, David Miller, and Kevin Walker, "The fisher corpus: a resource for the next generations of speech-to-text.," in *LREC*, 2004, vol. 4, pp. 69–71.

[14] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[15] Iain McCowan, Jean Carletta, W Kraaij, S Ashby, S Bourban, M Flynn, M Guillemot, T Hain, J Kadlec, V Karaiskos, et al., "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, 2005, vol. 88, p. 100.

[16] Anthony Rousseau, Paul Deléglise, and Yannick Esteve, "Ted-lium: an automatic speech recognition dedicated corpus.," in *LREC*, 2012, pp. 125–129.