To be published at Interspeech 2014
Singapore, September 2014

THE INFLUENCE OF PITCH AND NOISE ON THE DISCRIMINABILITY OF FILTERBANK FEATURES

Malcolm Slaney, Michael L. Seltzer

Microsoft Research, 1065 La Avenida, Mountain View, CA 94043

malcolm@ieee.org, mseltzer@microsoft.com

ABSTRACT

Most features used for speech recognition are derived from the output of a filterbank inspired by the auditory system. The two most commonly used filter shapes are the triangular filters used in MFCC (mel-frequency cepstral coefficients) and the gammatone filters that model psychoacoustic critical bands. However, for both of these filterbanks there are free parameters that must be chosen by the system designer. In this paper, we explore the effect that different parameter settings have on the discriminability of speech sound classes. Specifically, we focus our attention on two primary parameters: the filter shape (triangular or gammatone) and the filter bandwidth. We use variations in the noise level and the pitch to explore the behavior of different filterbanks. We use the Fisher linear discriminant to give us insight about why some filterbanks perform better than others. We observe three things: 1) there are significant differences even among different implementations of the same filterbank, 2) wider filters help remove the non-informative pitch information, and 3) the Fisher criteria helps us understand why. We validate the Fisher measure with speech recognition experiments on the Aurora-4 speech corpus.

Index Terms— Speech Recognition, MFCC, pitch, DNN, deep neural network, mel-frequency cepstral coefficients, Aurora-4

1. OVERVIEW

We started this project to better understand why a slight modification to the ubiquitous MFCC (mel-frequency cepstral coefficients) representation provided a significant improvement over the study's baseline. Instead we found wide variations in the details of MFCC implementations that dramatically affect speech recognition performance. This paper is a cautionary tale, along with a description of a tool to understand the reasons for the performance.

Our initial experiments with PNCC, a new representation based on perception [4], were encouraging and confirmed the paper's conclusions. Yet we were surprised to discover that the PNCC representation did not do better, in a large-scale experiment using a voice-search corpus, than our baseline system. Closer study led us to the observation that the HTK implementation of MFCC used as our baseline was different than the Auditory Toolbox implementation used as a baseline in the PNCC study.

MFCC filterbanks are still important to automatic speech recognition (ASR). State of the art recognizers use deep neural networks (DNNs) and these networks perform best when using a smoothed spectrum as input [9]. Ideally a DNN should give good results with any representation of the input signal, but in practice DNNs have worked best when the input representation if something akin to the MFCC filterbank. Thus we show results of speech-recognition experiments with a DNN-based acoustic model to verify the real-world performance.

There is much interest in designing new features for ASR that better characterize the important phonetic information, and remove the uninformative. One such approach is scattering theory, where a representation is designed to be sensitive to important distinctions, and insenstive to noise [5]. Our investigations into the effect of pitch on recognition accuracy is an experimental step in this direction.

Pitch and speech recognition have a mixed history. The conventional wisdom is that MFCC is a good representation because the cepstral processing, in a process reminiscent of homomorphic filtering, removes the pitch fluctuations in the power spectrum. But recent work has shown that adding pitch information to a recognizer improves performance [7], perhaps because formant positions are modified to fit the pitch harmonics [10][1]. Yet other work shows that estimating the tone in a Chinese utterance can be done *without* measuring the pitch [8]. Thus the influence of pitch and speech recognition is not as clear as we might like. Still pitch is used in this study as a source of variability in the recognition task.

2. TOOLS

In this section we describe the MFCC representation, and a modification using gammatone filters instead of the normal triangle filters. We also describe our diagnostic procedure, which consists of synthetic vowels, and the Fisher linear discriminant.

2.1. MFCC

There are two popular implementations of the MFCC representation. They are the ones found in the Hidden Markov Toolkit (HTK) [13] and the Auditory Toolbox [12]. Both implement the triangular filters that are the hallmark of the MFCC representation, but there are many small differences. The basic processing for MFCC processing consists of several steps: spectral analysis, forming critical-band filters to approximate the perceptual system, a loudness transformation, and then a discrete cosine transform (DCT) to reduce the dimensionality of the signal. Unfortunately this rather loose description allows a number of variations.

Our analysis is based on the summary code produced by Dan Ellis [3]. His code implements many popular variations of MFCC and characterizes their differences with algorithmic parameters. The HTK and Auditory Toolbox implementations of MFCC differ in 8 different dimensions! Table 1 shows these differences.

The PNCC system recommends a number of changes to MFCC and suggested significantly better results due to these modifications

[1] Simpson notes that female speakers often have an expanded vocal space, as measured by formant frequencies. This can be interpreted to suggest that speakers with higher pitch enhance the speech they produce to make the formants easier to distinguish, even in the face of wide harmonic spacing.

Label	AT Value	HTK Value
Liftering exponent	0	-22
DCT Type	2	3
Number of bands	40	24
Max Freq	6855	Nyquist
Filterbank Scale	mel	HTK mel
Filter domain	Magnitude	Power
Min Freq	133.33	0
Window Length	16ms	25ms

Table 1. Differences between Auditory Toolbox (AT) and HTK implementations of MFCC (as implemented by Ellis [3])

[4]. We believe the two primary contributions in Kim's paper are the shape of the filters, and how the power is normalized. This paper only addresses the filter-shape issue.

Note, we only work with the filterbank representation in this paper. For an MFCC representation, one reduces the dimensionality of the data using a discrete cosine (DCT) transform, but the effect of this step is orthogonal to the issues described here.

2.2. Gammatone Filters

We want to better understand the effect of bandwidth on performance. The bandwidth of MFCC triangular filters is determined by the spacing of the channels, since MFCC triangle i starts at the center frequency of triangle $i-1$, peaks at CF_i, and falls to zero at the center frequency of channel $i+1$.

PNCC changes the shape of the basic filter from triangular to a gammatone filter. The gammatone filter is a popular model of how auditory "channels" behave in the human auditory system. The filter widths are normally measured with critical band experiments, based on simple masking experiments, and then fit to the gammatone model.

The gammatone function is defined by this time-domain expression

$$h(t) = \begin{cases} ct^{n-1}exp(-2\pi bt)cos(2\pi f_0 t), & t >= 0 \\ 0, & t < 0, \end{cases} \quad (1)$$

where b determines bandwidth of the filter and f_0 is the center frequency. Using Glasberg and Moore's parameters [11] the bandwidth of a filter is a non-linear function of the center frequency (CF), $b = 1.09((CF/9.26449 + 24.7)$. We converted this to a frequency-domain weight using Darling's derivation [1].

These filters are often implemented as frequency-domain filters. It is important to note whether filters are characterized by their response in the magnitude domain, as is often done, or in the power domain (as can be done by HTK.) This distinction is important, for the purposes of this paper, because squaring a gammatone filter so it can be applied in the power domain makes it look more triangular. Or perhaps in a way that is more germane to this discussion, a triangle in the power domain (ala HTK) looks like a gammatone in the magnitude domain (ala the Auditory Toolbox). This is shown in Figure 1.

2.3. Synthetic Vowels

To more easily test the ideas in this paper, we synthesized three static vowels /a/, /i/, and /u/ using the standard formant frequencies in the Auditory Toolbox [12]. To make the task more meaningful we varied the pitch of the vowels over a wide range (100–250 Hz) and added

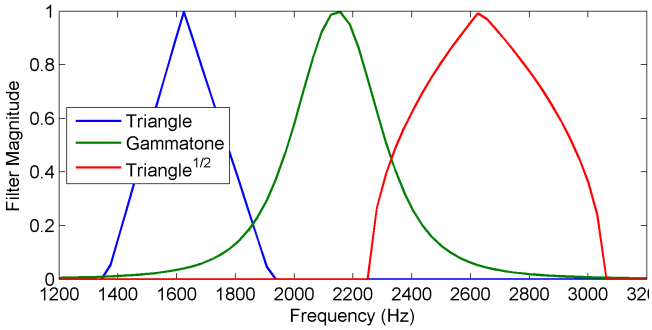

Fig. 1. Three types of filters used in auditory filterbanks. The rightmost filter is the magnitude response for a triangular filter (left) implemented in the power domain and is a better approximation of a gammatone filter (middle).

noise at varying levels. We then converted these audio waveforms into different versions of the MFCC representation and judged their suitability for recognition. This was done by using a measure of discriminability (see Section 2.4 below). A representation is better if there is better discriminability between the acoustic classes. In Section 3.4 we verify these ideas using a full recognition experiment.

2.4. Fisher Linear Discriminant Criteria

We use the criterion from the Fisher linear discriminant to characterize the performance of different representations. While in the end all that matters is speech-recognition performance, we would like to have a better understanding of *why* different representations perform as they do. Thus, we use discriminability as a ruler. Clouds of data that are further apart are more discriminable. Classes with more noise, and thus larger clouds, are more confusable.

Consider a set of data x that contains data belonging to two or more classes $x \in X_i$. The Fisher linear discriminant finds a rotation w of the data to maximize the following two-class criterion function [2]

$$J(w) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2}, \quad (2)$$

where $\tilde{m}_i$ is the mean of the n_i rotated data points wx_i and $1/n_i\tilde{s}_i^2$ is the corresponding variance. The Fisher criteria maximizes the ratio of two quantities: the inter-class and intra-class spreads. We base the intra-class spread, which will become the denominator so we hope it is small, on the spread of each data cluster.

$$S_w = \sum_i \tilde{s}_i^2. \quad (3)$$

The inter-class spread, which is larger for better discriminability, is equal to

$$S_B = \sum_i n_i(\tilde{m}_i - \tilde{m})(\tilde{m}_i - \tilde{m})^t. \quad (4)$$

The discrimination criteria becomes

$$J(w) = \frac{w^t S_b w}{w^t S_w w}. \quad (5)$$

We optimize this criteria, a generalized eigenvalue problem, with respect to w by finding the roots of the characteristic polynomial

$$|S_B - \lambda_i S_w| = 0 \quad (6)$$

and then solving for the first eigenvector w_1 using

$$(S_B \lambda_i S_W) w_i = 0. \quad (7)$$

The optimum value of w allows us to calculate the optimum discriminability. The maxium of this critera (2) is a good tool for explaining why one representation is better than another.

The Fisher criteria has a geometric explanation. The inter-class distance is divided by the average spread (intra-class) of the data, giving a dimensionless measure that is scaled by the data's spread. Thus a Fisher measure of 2 means that the data's centroids are twice the spread of the data, so the data clouds, as measured by their standard deviation, are just touching.

3. RESULTS

To better understand the role of the filterbank representation on ASR performance, we tested many variations of MFCC. We compared the two standard representations in Section 3.1. Sections 3.2 and 3.3 describe the effect of filter bandwidth and pitch on discriminability. Finally, in Section 3.4 we close the loop with full recognition experiments using MFCC filterbanks, DNNs, and the Aurora-4 corpus.

3.1. Effect of MFCC Parameters

As described above, the HTK implementation of MFCC and that found in the Auditory Toolbox differ along 8 dimensions. Figure 2 shows a summary of these differences, with the aim of understanding which dimensions matter for speech recognition. Each panel shows the discriminability of the three vowels as a function of signal to noise ratio. Evidently, the only two dimensions that make a difference are the (temporal) window length and number of filters (channels). Changing these two parameters causes the HTK representation to match the discriminability of the Auditory Toolbox representations, as shown in the lower-right panel of Figure 2.

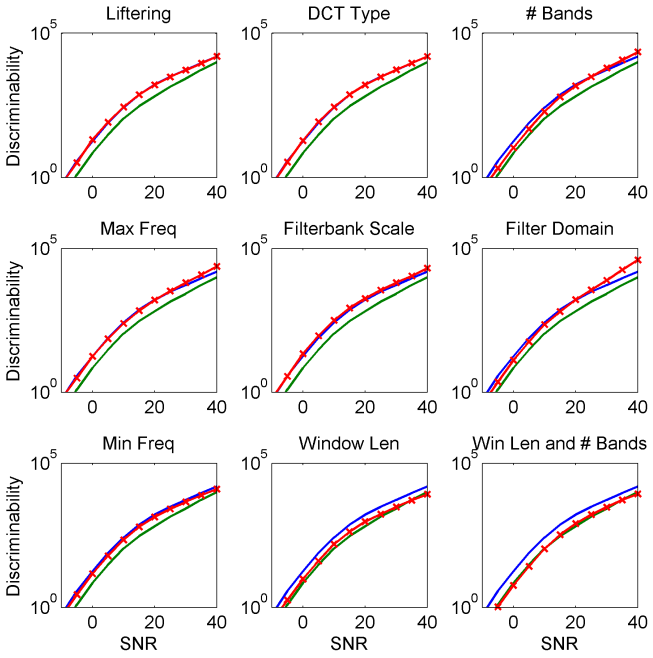

Fig. 2. The performance of 8 different variations of MFCC parameters, as measured by the Fisher discrimination critera. The upper line indicates the performance of the HTK MFCC, and the lower line is for the Auditory Toolbox performance. The line marked with an "x" are the result of changing the HTK parameters in one (or two) dimensions at a time, in order to match the Auditory Toolbox parameters. The first 8 plots show the disciminability by changing one parameter at a time. The last plot shows the effect of changing the only two parameters that matter.

3.2. Effect of Filter Bandwidth on Discrimination

Figure 3 shows the discriminability of the MFCC representations from the HTK and Auditory Toolbox using triangle and gammatone filters. As described in the original PNCC paper, the switch to gammatone filters gives a significant difference in discriminability over the original Auditory Toolbox filters. But this switch is not enough to exceed the performance of the HTK filters, using either gammatone or triangle filters. Perhaps there is something else that matters more than the filter shape

The curves of Figure 3 are a summary of the discriminability over the range of SNRs shown in Figure 2. To reduce the complexity of the graph we plot the average difference between the measured discriminability and the response for the standard HTK filter with a bandwidth multiplier of 1. Thus the standard HTK filter has a relative discriminability of 1 and is shown with an X on the plot.

3.3. Effect of Pitch on Discriminability

We started our work with a hypothesis that filter shape had an effect on the representation of a harmonic (voiced) speech signal. A representation with sharp filters, like that of a triangle, might be more sensitive to pitch because the exact location of the peak might or might not line up exactly with a harmonic. Small changes in the pitch would move a harmonic through the triangles peak, giving relatively large changes in the filterbank representation. Next we concentrate on the relative filter bandwidth, since the default bandwidth of the triangle (CF-CF) and gammatone (critical band) filters are different.

To explore the connection between the sensitivity of the representations to pitch, we synthesized vowels with a range of low (100–158Hz) and high (158–250Hz) pitches. We also varied the filter bandwidth by linearly scaling the filter width with a filter bandwidth multiplier. A multiplier of 1 means that the triangle filters extend from CF to CF, as originally specified, while gammatone filters have a nominal width of one critical band.

Figure 4 shows the effect of a filter bandwidth multiplier on discriminability. In general the discriminability of the low-pitch vowels is better than the high-pitch vowels (especially for high-bandwidth multipliers). This is (probably) because the high-pitch sounds have more intra-class variability. On the left with narrow filters, the gammatone filters are broader so the discriminability of the gammatone filters is better than the mel filters.

3.4. ASR on Aurora-4

To connect this analysis to speech recognition performance, we performed a series of experiments using the Aurora-4 corpus [9]. Aurora-4 is a medium vocabulary task based on the Wall Street Journal (WSJ0) corpus. We performed the experiments with the 16 kHz multi-condition training set consisting of 7137 utterances from 83 speakers. One half of the utterances were recorded by the primary Sennheiser microphone and the other half were recorded using one of a number of different secondary microphones. Both

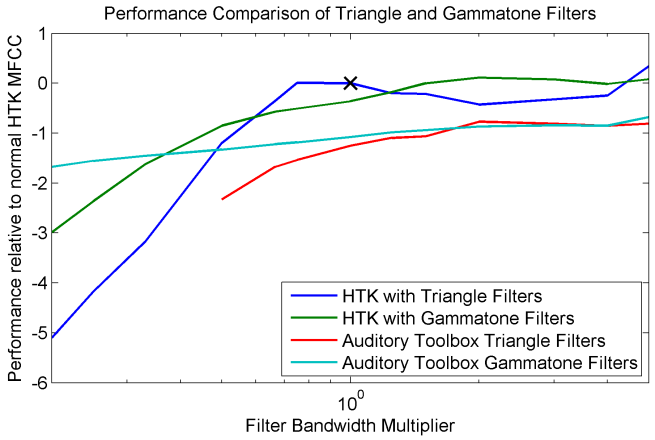

Fig. 3. Overall performance (averaged over all SNRs for HTK and Auditory toolbox implementations of MFCC, with triangular and gammatone filter shapes. This graph is a function of an extra bandwidth multiplication factor, over and above the normal bandwidth: CF_{i-1} to CF_{i+1} for the triangles, and critical bands for the gammatone. The bandwidth multiplier ranges from 0.2 to 5.

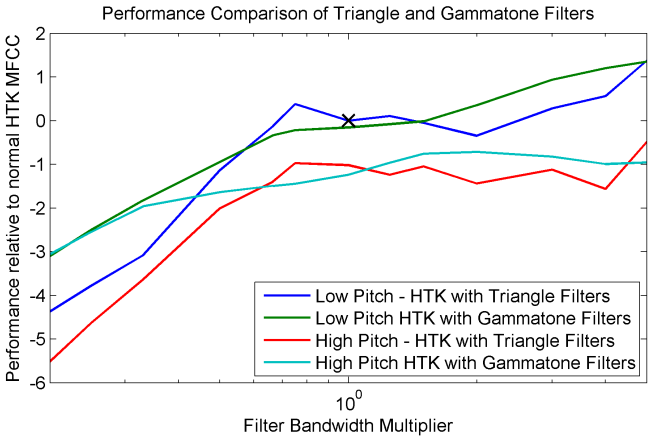

Fig. 4. Filterbank performance as a function of extra bandwidth, for low and high pitch speech sounds. The bandwidth multiplier ranges from 0.2 to 5.

halves include a combination of clean speech and speech corrupted by one of six different noises (street traffic, train station, car, babble, restaurant, airport) at 10–20 dB SNR. The evaluation set is derived from the WSJ0 5K-word closed-vocabulary test set which consists of 330 utterances from 8 speakers. This test set was recorded by the primary microphone and a secondary microphone. These two sets are then each corrupted by the same six noises used in the training set at 5–15 dB SNR, creating a total of 14 test sets. These 14 test sets can then be grouped into 4 subsets: clean, noisy, clean with channel distortion, noisy with channel distortion.

We performed speech recognition using a hybrid DNN-HMM acoustic model. We trained the DNN using a cross-entropy objective function with labels generated by a forced alignment of the training data to the 3202 senones of a conventional GMM–HMM recognizer. We decoded the speech with the task-standard WSJ0 bigram language model.

In the experiments performed, we used different variations of HTK filterbank features as input to the DNN. In all cases, we normalized the utterance-level means and used first- and second-order derivative features. We formed the input layer from a context window of 11 frames. The DNNs had 7 hidden layers with 2048 hidden units in each layer and the final soft-max output layer had 3202 units, corresponding to the senones of the HMM system. We initialized the networks using layer-by-layer generative pre-training and then discriminatively trained using twenty-five iterations of back propagation. We used a learning rate of 0.16 for the first 15 epochs and 0.004 for the remaining 10 epochs, with a momentum of 0.9. Seltzer et al. published more details of the training procedure [9].

We experimented with 6 different filterbanks. In all the cases, the we applied the filterbank's magnitude response to the power spectrum of the input signal and then applied a natural logarithm. We experimented with both the mel-frequency triangular filters and the gammatone filter responses and in each case, evaluated filters with the standard bandwidth, half the bandwidth, and twice the bandwidth. Figure 5 shows the results. The normal triangle filterbank has its best performance at the default width, while shrinking the bandwidth provides better results for the gammatone filterbank.

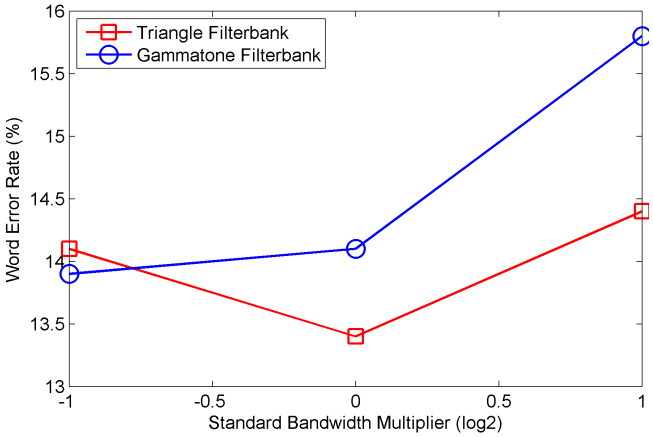

Fig. 5. ASR performance on the Aurora-4 corpus with HTK MFCC using triangular and gammatone filters, as a function of extra bandwidth (a multiplicative factor).

4. CONCLUSIONS

This paper describes the effect of bandwidth and filter shape on modern ASR systems. This topic is important because MFCC filterbanks are often used as a baseline for ASR experiments, and there are significant differences in performance with seemingly small changes in implementation. We investigated these differences by looking at the system's response to noise and pitch variations. We used a ruler based on the Fisher linear discriminant criteria to measure how different filterbank parameters affect recognition performance, and then validated our hypothesis using the Aurora-4 corpus. While both bandwidth and filter shapes represent linear rotations of the original spectral slice, and DNNs should be immune to such changes, they do have a significant impact on system performance. The discriminability experiments suggest that wider bandwidths are best in the face of pitch variations, but this must be tempered with narrow filters that can represent many different phonetic classes, as found in the Aurora-4 corpus. These countervailing factors give the minimum error shown in Figure 5.

5. REFERENCES

[1] A. M. Darling. Properties and Implementation of the Gammatone Filter: A Tutorial. In *Speech Hearing and Language, Work in Progress,* University College London, Department of Phonetics and Linguistics, pp. 43–61, 1991.

[2] R.O. Duda, P.E. Hart und D.G. Stork. *Pattern Classification,* Second Edition, Wiley, 2001.

[3] Daniel P. W. Ellis. PLP and RASTA (and MFCC, and Inversion) in Matlab. `http://www.ee.columbia.edu/%7Edpwe/resources/matlab/rastamat/`, 2005.

[4] C. Kim, and R. M. Stern. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing,* Kyoto, Japan, March 2012.

[5] Stephane Mallat. *Can Signal Classification Speak Mathematics?* Plenary talk, ICASSP, Kyoto, Japan, March 2012.

[6] N. Parihar and J. Picone. Aurora working group: DSR front-end LVCSR evaluation AU/384/02. Tech. Rep., Inst. for Signal and Information Process, Mississippi State University, 2002.

[7] Pegah Ghahremani, Bagher BabaAli, Daniel Povey, Korbinian Riedhammer, Jan Trmal and Sanjeev Khudanpu. A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 2014.

[8] Neville Ryant, Malcolm Slaney, Mark Liberman, Elizabeth Shriberg, and Jiahong Yuan. Highly Accurate Mandarin Tone Classification In The Absence of Pitch Information. *Proceedings of Speech Prosody #7*, Dublin, Ireland, 2014.

[9] Michael Seltzer, Dong Yu, and Yongqiang Wang. An Investigation Of Deep Neural Networks For Noise Robust Speech Recognition. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP),* 2013.

[10] Adrian P. Simpson. Phonetic Differences between Male and Female Speech. *Language and Linguistics Compass*, 3/2 pp. 621640, 2009.

[11] Malcolm Slaney. *An Efficient Implementation of the Patterson-Holdsworth Auditory Filter Bank,* Apple Computer Technical Report #35, 1993.

[12] Malcolm Slaney. *Auditory Toolbox (version 2).* Interval Research Corporation Technical Report #1998-010, Palo Alto, CA, 1998.

[13] S. Young, J Jansen, J. Odell, D. Ollason, P. Woodland. *HTK: Hidden Markov Toolkit.* 1995.