A Comparative Study of Filter Bank Spacing for Speech Recognition

Ben J. Shannon, Kuldip K. Paliwal

***Abstract*— Although Mel scale filter bank spacing is used extensively in Automatic Speech Recognition (ASR), it will be shown in this paper that it provides little benefit over other perceptually motivated frequency warping scales. An MFCC like feature based on the Bark scale is shown to yield similar performance in speech recognition experiments as MFCC. The performance of MFCC and BFCC features are also compared to Uniform Frequency Cepstral Coefficients (UFCC) where it is shown that neither the Mel or Bark scale provide significant advantage over a Uniform scale if training and test conditions are matched.**

I. Introduction

MEL Frequency Cepstral Coefficients (MFCC) are used extensively in Automatic Speech Recognition (ASR). MFCC features are derived from the FFT magnitude spectrum by applying a filter bank which has filters evenly spaced on a warped frequency scale. The logarithm of the energy in each filter is calculated and accumulated before a Discrete Cosine Transform (DCT) is applied to produce the MFCC feature vector. The frequency warping scale used for filter spacing in MFCC is the Mel (*Mel*ody) scale. The Mel scale is a perceptually motivated scale that was first suggested by Stevens and Volkman in 1937 [4]. The scale was devised through human perception experiments where subjects were asked to adjust a stimulus tone to perceptually half the pitch of a reference tone. The resulting scale was one in which 1 Mel represents one-thousandth of the pitch of 1 kHz [3] and a doubling of Mels produces a perceptual doubling of pitch.

The Bark scale provides an alternative perceptually motivated scale to the Mel scale. Speech intelligibility perception in humans begins with spectral analysis performed by the basilar membrane (BM). Each point on the BM can be considered as a bandpass filter having a bandwidth equal to one *critical bandwidth* or one Bark [5]. The bandwidth of several auditory filters were empirically observed and used to formulate the Bark scale.

It will be shown in this paper that an MFCC like feature, based on the Bark scale and referred to as BFCC, yields similar performance in speech recognition experiments as MFCC. The performance of MFCC and BFCC features are also compared to Uniform Frequency Cepstral Coefficients (UFCC). It will be shown that the scale used to space the filter bank provides little advantage, especially when the training and testing conditions match.

The authors are with the Signal Processing Laboratory, School of Microelectronic Engineering, Faculty of Engineering and Information Technology, Griffi th University, Brisbane, QLD 4111 Australia (E-mail: ben.shannon@student.gu.edu.au, k.paliwal@me.gu.edu.au)

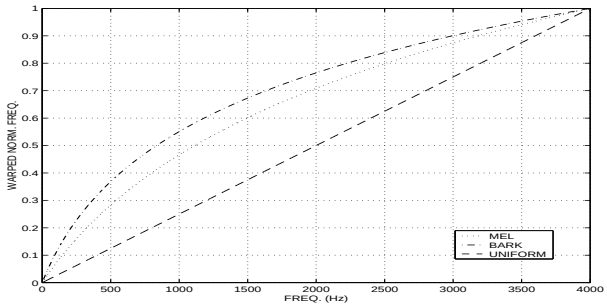

Fig. 1. Warped Freq. Comparison

II. Aurora Task

The Aurora database provides speech samples and scripts to perform speaker independent speech recognition experiments in clean and noisy conditions. The speech on Aurora consists of digit sequences (eg. sil-one-four-seven-three-five-three-three-sil) derived from the TI-digit database down sampled to 8kHz. These speech samples are then filtered with a G.712 characteristic [1].

Two training situations are presented in the scripts. The two training sets consist of 8440 utterances each. The clean set is not modified past the G.712 characteristic, but the multi set utterances are divided into 20 groups. These 20 groups consist of 5 different SNR levels (clean, 20dB, 15dB, 10dB, 5dB) with 4 different noise types (subway, babble, car, exhibition) artificially added.

The testing set consists of 28028 utterances and is the same for both training situations. The test set is divided evenly among 7 test SNRs (clean, 20dB, 15dB, 10dB, 5dB, 0dB, -5dB) and 4 noise types (subway, babble, car, exhibition) which gives 1001 utterances in each category.

III. Experiments

A. Recognition Framework

The HTK tool kit [6] is used for the speech recognition experiments. The ten digits (0-9) in the recognition dictionary are each modelled using a single continuous density Hidden Markov Model (HMM). Each HMM contains 16 emitting states with three Gaussian mixtures per state.

B. Warping Scales

When comparing the two warped scales, Mel and Bark, against the Uniform scale (Fig. 1), it is apparent that the Bark and Mel filters (Fig. 2.a, b) have narrow bandwidth at low

frequency and get wider as the frequency is increased unlike the Uniform filters (Fig. 2.c) which have constant spacing and bandwidth.

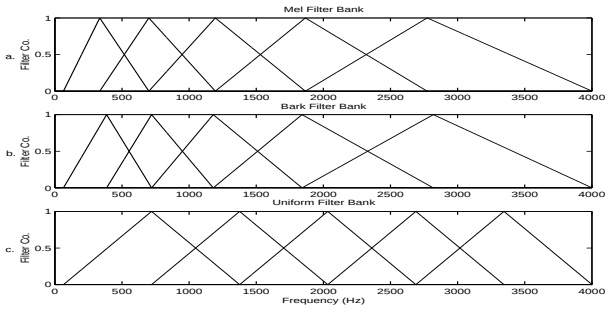

Fig. 2. Example Filter Banks

C. Feature Extraction

Feature frames with 36 dimensions are extracted every 10ms. Each feature frame consists of 12 cepstral coefficients (excluding c0), concatenated with the delta and acceleration coefficients of the cepstral features. In these experiments an energy coefficient was not used.

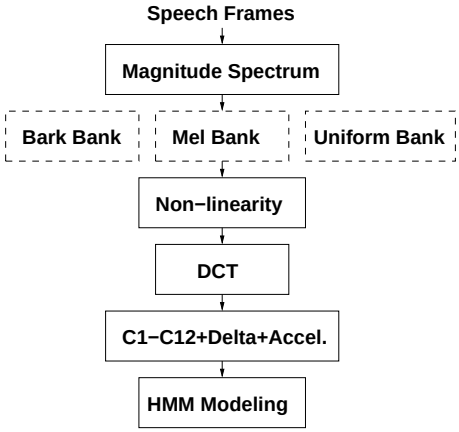

Fig. 3. Block Diagram

From the magnitude spectrum and filter banks, 23 Filter Bank Energies (FBE) are calculated before being converted to cepstral coefficient (Fig. 3). The spacing and bandwidth of the 50% overlapped filters is determined on the warped scale. When the filters are returned to the Hz scale they are given a triangular shape (Fig. 2) where each filter starts and ends at the centre of the adjacent filter. The 23 channel filter bank is positioned between 64 and 4000 Hz.

D. Mel Scale Expression

Equation (1) is the Hz to Mel warping used in the experiments [6].

$$Mel(f) = 2595 log_{10}(1 + (f/700)) \quad (1)$$

E. Bark Scale Expression

An approximate expression (Eq. 2) for the Bark scale frequency warping, due to Schroeder [2], is used in these experiments.

$$Bark(f) = 6 log_e \left((f/600) + \sqrt{(f/600)^2 + 1} \right) \quad (2)$$

F. Scoring

The recognition results are scored using the HTK accuracy expression [6] (Eq.3). In (Eq.3) the number of inserted labels are subtracted from the number of correct labels, then divided by the total number of labels in the true transcription to give the accuracy score.

$$Accuracy = \frac{Correct - Insertions}{Total} \times 100\% \quad (3)$$

IV. Recognition Results

A. Clean Training

In clean training conditions and noise free testing conditions, all three feature sets performed well (96.85%-97.97%), but degraded rapidly when noise was introduced (62.11%-78.42% @ 10dB). In this experiment the Mel and Bark scale filter spacing performance difference was insignificant. The accuracy results for both spacing schemes appears to overlay each other (Fig.4) in all four noise conditions. The uniform spacing though can be as much as 8% lower on accuracy than the nearest warped scale (MFCC Sub. @ 10dB)(Tbl. I).

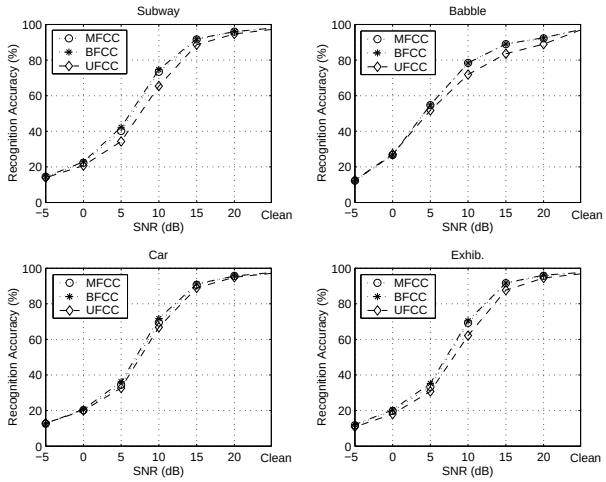

Fig. 4. Test Set A Clean Training

B. Multi Training

In this experiment the Mel and Bark scale filter spacing performance is insignificantly different (Fig.5). The recognition accuracy for all three feature sets is maintained longer with SNR, as expected, since the noise condition is trained in. Unlike in the clean training experiments, the UFCCs are much closer to the warped frequency features other than in the babble noise case where they are 5.7% lower at 10dB (Tbl. II).

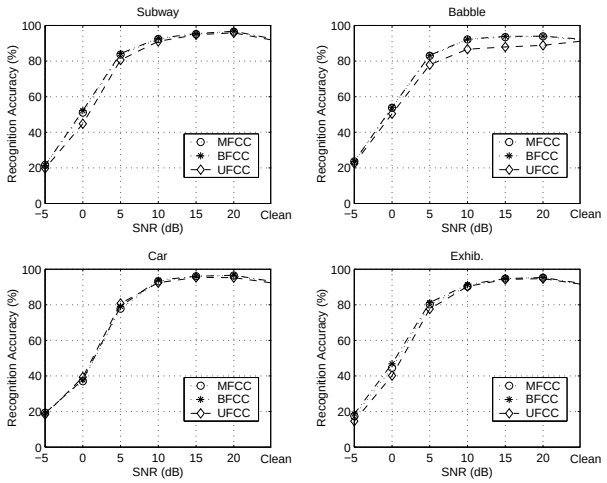

Fig. 5. Test Set A Multi Training

V. Conclusions

In this paper it has been shown that Bark scale and Mel scale filter banks have equivalent performance in speech recognition tasks. It has also been shown that these features provide an advantage over uniformly space filter banks only when the training and testing condition are not matched.

	MFCC				BFCC				UFCC			
	Sub.	Bab.	Car.	Exhib.	Sub.	Bab.	Car.	Exhib.	Sub.	Bab.	Car.	Exhib.
Clean	97.79	97.52	97.61	97.50	97.97	97.25	97.58	97.59	97.27	97.13	97.17	96.85
20 dB	96.01	92.23	95.71	95.71	96.01	92.59	95.62	96.11	94.69	89.00	94.93	94.51
15 dB	91.59	88.97	90.67	91.64	92.02	89.09	91.14	91.70	88.64	83.62	88.99	87.63
10 dB	73.44	78.42	69.49	69.27	74.67	78.36	71.73	70.41	65.40	71.89	66.84	62.11
5 dB	40.19	54.72	34.54	33.35	42.25	54.87	36.18	35.24	34.30	51.72	32.69	30.76
0 dB	22.29	26.66	20.22	19.75	22.84	26.66	20.73	20.39	20.72	27.15	20.01	18.02
-5 dB	14.52	12.18	12.59	11.76	14.40	12.45	12.59	12.06	13.88	12.55	12.94	10.80
Mean	62.26	64.39	60.05	59.85	62.88	64.47	60.80	60.50	59.27	61.87	59.08	57.24

TABLE I
Test Set A Clean Training Results (Accuracy%)

	MFCC				BFCC				UFCC			
	Sub.	Bab.	Car.	Exhib.	Sub.	Bab.	Car.	Exhib.	Sub.	Bab.	Car.	Exhib.
Clean	92.57	91.81	92.78	91.30	92.69	92.35	93.17	92.04	91.86	91.23	92.28	91.61
20 dB	96.53	93.89	96.45	95.28	96.75	94.01	96.63	95.34	95.89	88.88	95.26	94.72
15 dB	95.18	93.47	96.06	94.75	95.46	93.89	96.03	94.75	94.81	87.97	95.35	94.23
10 dB	92.42	92.17	93.23	90.28	92.78	92.32	93.86	91.33	91.10	86.70	92.42	90.31
5 dB	83.64	82.98	77.81	80.25	84.28	83.28	79.12	81.24	80.63	77.96	80.73	77.63
0 dB	51.06	53.75	37.07	44.55	52.20	54.05	37.85	46.84	44.77	50.42	39.37	40.20
-5 dB	21.61	23.64	19.36	17.46	21.37	23.58	19.18	18.64	19.71	22.46	18.31	14.63
Mean	76.14	75.96	73.25	73.41	76.50	76.21	73.69	74.31	74.11	72.23	73.39	71.90

TABLE II
Test Set A Multi Training Results (Accuracy%)

References

[1] Ericsson, "Aurora database readme," *ETSI/AURORA PROJECT*, January, 25th 2000.
[2] H. Hermansky, "Perceptual linear prediction (plp) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.
[3] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A guide to theory, algorithm, and system development.* Prentice Hall, Inc., 2001, iSBN 0-13-022616-5.
[4] B. Moore, *Hearing*. Academic Press, Inc., 1995, iSBN 0-12-505626-5.
[5] ——, *An introduction to the psychology of hearing*. Academic Press, 1997, iSBN 0-12-505627-3.
[6] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book.* Cambridge University Engineering Department, 2002.