International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

Performance Evaluation of Mel and Bark Scale based Features for Text-Independent Speaker Identification

S. B. Dhonde, Amol A. Chaudhari, M. P. Gajare

Abstract*: *The performance of Mel scale and Bark scale is evaluated for text-independent speaker identification system. Mel scale and Bark scale are designed according to human auditory system. The filter bank structure is defined using Mel and Bark scales for speech and speaker recognition systems to extract speaker specific speech features. In this work, performance of Mel scale and Bark scale is evaluated for text-independent speaker identification system. It is found that Bark scale centre frequencies are more effective than Mel scale centre frequencies in case of Indian dialect speaker databases. Mel scale is defined as per interpretation of pitch by human ear and Bark scale is based on critical band selectivity at which loudness becomes significantly different. The recognition rate achieved using Bark scale filter bank is 96% for AISSMSIOIT database and 95% for Marathi database.

Keywords: Formants, MFCC, Text-independence, VQ

I. INTRODUCTION

A filter bank model and LPC model are widely used as signal processing front end tools for speech processing. The advantage of using filter banks is that they can be designed to cover an analysis band from 150 Hz to 7000 Hz. In a given frequency band, energy of speech signal is measured using filter bank analyser. Filter banks based on nonuniform spacing have been used in many practical systems [1]. Nonuniform filter banks are used to reduce overall computation and to characterize the speech spectrum in a manner considered more consistent with human perception [1]. The filters can be spaced along a logarithmic frequency scale which is often justified from a human auditory perception. Alternatively, the critical band scale can be used directly for designing of a nonuniform filter bank. The filters are placed according to perceptual studies for the intension of choosing bands that give equal contribution to speech articulation. Mel scale is used to define Mel filter bank in the computation of MFCCs for speaker recognition system. MFCC feature extraction is widely used speaker recognition systems [2, 3]. In recent years, there have been efforts for the robust speaker identification system.

Manuscript published on 30 September 2019.
*Correspondence Author(s)
Dr. S. B. Dhonde, Associate Professor, Department of Electronics & Telecommunication Engineering, A.I.S.S.M.S, Institute of Information Technology, Pune, India
Amol A. Chaudhari, Assistant Professor, Department of Electronics & Telecommunication Engineering, A.I.S.S.M.S, Institute of Information Technology, Pune, India
M. P. Gajare, Assistant Professor, Department of Electronics & Telecommunication Engineering, A.I.S.S.M.S, Institute of Information Technology, Pune, India

Robustness of speaker identification system mainly depends upon the features extracted from speech signal. MFCC features are modified in some ways either by concatenating them with complementary information or modifying the pre-processing steps of speech signal. Multitaper MFCC features [4, 5], Inverted MFCC features [6], Combining MFCC with phase information [7] and even replacing the Mel filter bank by Radon based features [3] or wavelet based features [8 - 12] have been extensively studied.

However, the frequency warping scales such as Bark scale and ERB scale are less studied in speaker recognition systems. The critical band i.e. bandwidth at which subjective response such as loudness perceived becomes significant can be approximated by Bark scale [13]. The use of Bark scale for audio signal processing have been studied in [14] [15]. Also, performance of Bark frequency cepstral coefficients (BFCC) has been studied for speech recognition in [16].

In this paper, the frequency scales Mel and Bark are studied and their performance is evaluated in speaker identification system.

II. FREQUENCY WARPING SCALES

A. Mel scale

According to psychoacoustics studies, the content of frequency in pure tones or speech signal perceived by human ear follows a nonlinear scale. The understandings from human auditory system are used to define subjective pitch of pure tones or speech signal [13]. Mel scale is used to measure a subjective pitch. When stimulus frequency is increased linearly, the subjective pitch in mels increases less rapidly [13]. Mel-filter bank represents different perceptual effects at different frequency bands. Mel scale is linear below 1 kHz and logarithmic above 1 kHz [13]. As shown in figure 1, the edges of filters are placed such that they coincide with the centre frequencies in adjacent filters. The following equation represents the formula for conversion of Linear to Mel frequencies.

$$F_{mel} = 2595 * \log_{10}\left(1 + \frac{f_{linear}}{700}\right) \quad (1)$$

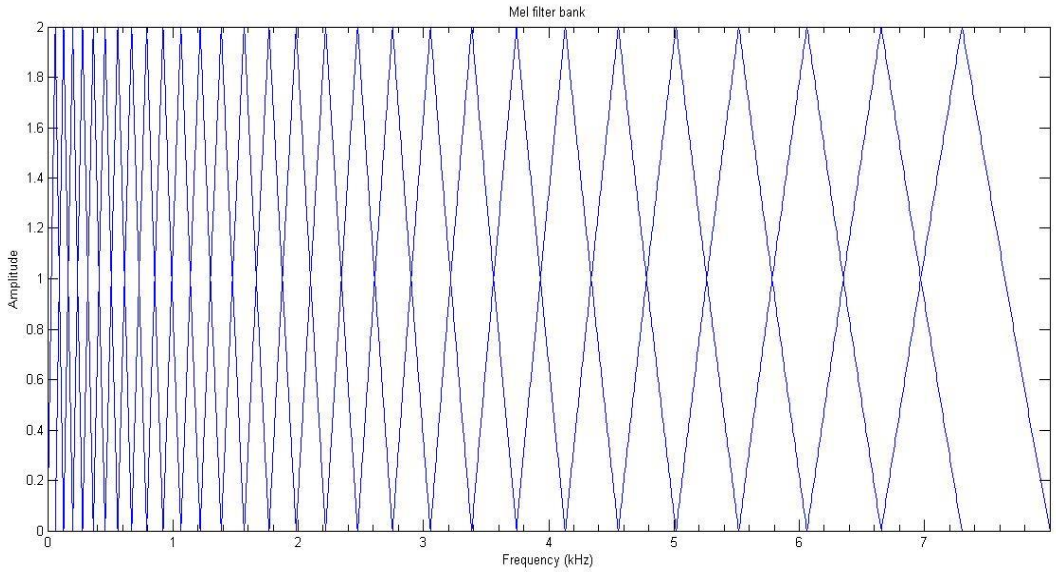

Fig. 1. Mel filter bank

B. Bark scale

Bark scale is psychoacoustical scale which is based upon loudness measurement. The scale has values ranging from 1 to 24. These 24 values relates to 24 critical bands of hearing. In Bark scale, the distance from the centre to left edge is different from that from the centre to the right edge. The band edges and band centres are given in [14]. The interpretation of the centre-frequencies and bandwidths are discussed in [14]. Following equation shows relationship between frequencies in Hz and Bark scale frequencies

$$(\omega) = 6\, ln\left(\frac{\omega}{1200\pi} + \left[\left(\frac{\omega}{1200\pi}\right)^2 + 1\right]^{0.5}\right) \quad (2)$$

where, Ω represents the angular frequency in Bark scale, and ω represents angular linear frequency = 2πf

Apart from equation 2, many analytical formulae have been proposed for Bark scale [17] [18].

$$(bark) = \ 13\tan^{-1}\left(0.76\frac{F(Hz)}{1000}\right) + 3.5\ \tan^{-1}\left(\frac{F(Hz)}{7500}\right)^2 \quad (3)$$

$$Critical\ Bark\ Rate = \ \left[\frac{(26.81f)}{(1960+f)}\right] - 0.53 \quad (4)$$

$$F(bark) = 6\sinh^{-1}\left(\frac{F(Hz)}{600}\right) \quad (5)$$

The important constants required to specify to define Bark filter bank are the number of filters, the minimum frequency, and maximum frequency [17]. The frequency range for these filters will be specified by minimum and maximum frequencies. As it can be seen from figure 1 and 2, Bark scale has high bandwidth as compared to mel scale particularly at higher frequency.

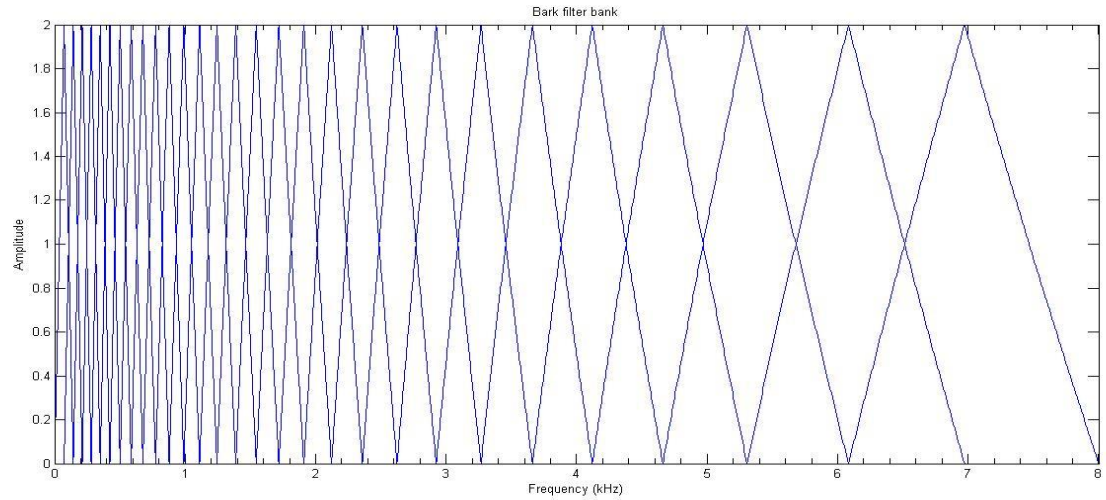

Fig. 2. Bark filter bank

III. EXPERIMENTAL SET-UP

The experiments have been evaluated on our recorded database AISSMSIOIT database [19] of 75 speakers. Another databases used for speaker identification experiments are TIMIT [20] consisting of 630 speakers and subset of 120 speakers and Marathi database [21] consisting of 120 speakers (60 male and 60 female). A speaker model is trained using five SX and three SI sentences (almost twenty-four seconds of duration) from TIMIT database and remaining two SA sentences (each of almost three seconds of duration) are used for testing phase. In Marathi database, a speaker model is trained using ten sentences approximately three seconds of duration and for testing two sentences roughly each of duration three seconds are used. The pre-processing steps of speech signal are carried out as mentioned in [19]. In framing step 256 samples per frame in case of TIMIT database, 512 samples per frame for Marathi database and 1024 samples per frame for AISSMSIOIT database are carried out along with 50% overlap and hamming windowing. Twelve MFCCs are computed by changing number of MFCC filters and speaker model is created using LBG algorithm (VQ). The testing procedure is carried out as mentioned in [19].The same procedure is carried out to calculate Bark scale based frequency cepstral coefficients (BFCC). In above mentioned procedure, Mel scale filter bank is substituted by Bark scale filter bank and twelve cepstral coefficients are computed by varying Bark filters.

IV. RESULTS AND DISCUSSION

The percentage recognition rate for AISSMSIOIT database and average recognition rate for Marathi and TIMIT databases is carried out.

$$\text{Recognition Rate} = \frac{\text{Number of accurately matched speakers}}{\text{Total number of test speakers sample}} \times 100\%$$

Table- I: Recognition rate (%) evaluated on AISSMSIOIT database

Sr. No.	No. of Filters	Cepstral Coefficients	Recognition Rate for Mel Scale	Recognition Rate for Bark Scale
1	13	12	90.66	94.66
2	20	12	92	96
3	32	12	94.66	96

Table- II: Average recognition rate (%) evaluated on Marathi database

Sr. No.	No. of Filters	Cepstral Coefficients	Average Recognition Rate for Mel Scale	Average Recognition Rate for Bark Scale
1	13	12	94.16	93.75
2	20	12	93.75	95
3	30	12	93.75	94.16

Table- III: Average recognition rate (%) evaluated on 120 speakers of TIMIT database

Sr. No.	No. of Filters	Cepstral Coefficients	Average Recognition Rate for Mel Scale	Average Recognition Rate for Bark Scale
1	13	12	98.75	93.75
2	20	12	99.16	97.50
3	29	12	99.58	99.16

Table- IV: Recognition rate evaluated (%) on AISSMSIOIT database

Sr. No.	No. of Filters	Cepstral Coefficients	Average Recognition Rate for Mel Scale	Average Recognition Rate for Bark Scale
1	13	12	94.44	88.89
2	20	12	95.95	92.61
3	29	12	95.80	93.96

It is observed that Bark filters are effective than Mel filters in case of Indian dialect speaker database. This is because centre frequencies are effectively captured in case of bark scale than mel scale. As shown in table 5, the location of formants in voiced part has more closeness with bark centre frequencies than mel centre frequencies. This is observed specially in case of formant F1. For speaker recognition, the formant F1 is important as it brings the maximum energy along with it which reveals the language structure and speaking style [22]. It is observed that formants in case of Indian dialect speakers are effectively mapped using bark centre frequencies. It is observed that for twenty Bark filters, the recognition rate of the system is improved as compared to 13, 30 and 32 Mel filters in case of Indian dialect speaker database. This is because twenty Bark filters have optimally covered the critical band of human hearing system. This signifies that number of filters and selection of critical band is an important aspect to define Bark filter bank for speaker identification. The critical band is the bandwidth at which loudness perceived becomes significantly different. Psychoacoustic critical band can be measured using bark scale. However, as shown in table 3 and 4, for TIMIT database the average recognition rate of Mel scale is better than Bark scale. This is because Mel scale is more suitable for pitch perception and phonetic approaches whereas, Bark scale is suitable for measurement of loudness. The bark scale is physiological scale in which bandwidth is constant. Therefore, number of filters and critical band selection are primary for Bark scale filter bank. Mel scale is psychological scale in which as increment in the filter number increases the bandwidth thereby increasing the recognition rate.

Table- V: Location of formants with respect to Mel centre frequency and Bark centre frequency for AISSMSIOIT Database

Sr. No.	Speaker Number	Formant F1	Mel centre Frequency	Bark centre Frequency
1	1	**558**	77.88	77.52
2	2	**521**	164.43	152.74
3	3	**479**	**260.60**	226.76
4	4	**575**	367.47	**300.31**
5	5	**407**	**486.23**	376.49
6	6	**463**	**618.21**	**457.99**
7	7	**501**	764.88	**545.38**
8	8	**523**	927.85	639.32
9	9	**544**	1109	740.58
10	10	**618**	1310	850.04
11	21	**439**	1534	968.76
12	32	**420**	1782	1098
13	36	**463**	2059	1239
14	40	**320**	2366	1394
15	56	**289**	2707	1564
16	69	**457**	3086	1753
17	70	**591**	3507	1963

In this table, bold faces are for Bark centre frequencies and bold with underlined are for Mel centre frequencies.

V. CONCLUSION

The performance of Mel scale and Bark scale is evaluated. Mel scale and Bark scale are filter bank structure which is used in speaker recognition system. This filter banks are mainly based human auditory system. It is found that centre frequencies of Bark scale are closely matched with the formant F1, than Mel scale centre frequencies in case of Indian dialect speaker database. Formant F1 is mainly useful for inter-variability between speakers. Also, Bark scale is good approximation for measurement of critical band at which loudness becomes significantly different. Whereas, Mel scale is suitable for pitch perception and phonetical information. Therefore, Mel scale has shown better performance as compared to Bark scale on TIMIT database. The recognition rate achieved is 96% on AISSMSIOIT database and average recognition rate achieved is 95% on Marathi database using 20 Bark filters.

REFERENCES

1. L. Rabiner, B. Juang, B. Yegnanarayana, "Fundamentals of Speech Recognition", published by Pearson Education.
2. Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors", *Journal on Speech Communication*, Elsevier, vol. 52, no. 1, pp. 12–40, 2010.
3. Pawan K. Ajmera, Dattatray V. Jadhav, Ragunath S. Holambe, "Text-independent speaker identification using Radon and discrete cosine transforms based features from speech spectrogram", *Journal on Pattern Recognition*, Elsevier, vol. 44, no. 10-11, pp. 2749-2759, 2011.
4. Tomi Kinnunen, Rahim Saeidi, Filip Sedlák, Kong Aik Lee, Johan Sandberg, Maria Hansson-Sandsten, Haizhou Li, "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification", *IEEE Transactions Audio, Speech and Language Processing*, vol.20, no.7, pp. 1990-2001, 2012.
5. Md Jahangir Alam , Tomi Kinnunen , Patrick Kenny , Pierre Ouellet, Douglas O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors", *Journal on Speech Communication*, Elsevier, vol. 55, no. 2, pp. 237-251, 2013.
6. R.Shantha Selva Kumari, S. Selva Nidhyananthan, Anand.G, "Fused Mel Feature sets based Text-Independent Speaker Identification using Gaussian Mixture Model", *International Conference on Communication Technology and System Design 2011, Journal on Procedia Engineering*, Elsevier, vol. 30, pp. 319–326, 2012.
7. Seiichi Nakagawa, Longbiao Wang, and Shinji Ohtsuka, "Speaker Identification and Verification by Combining MFCC and Phase Information", *IEEE Transactions Audio, Speech and Language Processing*, vol.20, no.4, pp. 1085-1095, 2012.
8. Jian-Da Wu, Bing-Fu Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition", *Journal on Expert Systems with Applications*, Elsevier, vol. 36, no. 2, pp. 3136–3143, 2009.
9. Deshpande, Mangesh S., and Raghunath S. Holambe. "Speaker identification using admissible wavelet packet based decomposition", *International Journal of Signal Processing* 6.1 (2010): 20-23.
10. Sumithra Manimegalai Govindan, Prakash Duraisamy, Xiaohui Yuan, "Adaptive wavelet shrinkage for noise robust speaker recognition", *Journal on Digital Signal Processing*, Elsevier, vol. 33, pp. 180-190, 2014.
11. Noor Almaadeed, Amar Aggoun, Abbes Amira, "Speaker identification using multimodal neural networks and wavelet analysis", *IET Journals and Magazines*, vol. 4, no. 1, pp. 18-28, 2015
12. Khaled Daqrouq, Tarek A. Tutunji, "Speaker identification using vowels features through a combined method of formants, wavelets, and neural network classifiers", *Journal on Applied Soft Computing*, Elsevier, vol. 27, pp. 231-239, 2015.
13. Holambe, Raghunath S., Deshpande, Mangesh S., "Advances in Non-Linear Modeling for Speech Processing", *Springer Briefs in Speech Technology*, Section 2, Section 6, pp. 11-15, 77-82, ISBN 978-1-4614-1505-3, 2012.
14. Julius O. Smith III, Jonathan S. Abel, "Bark and ERB Bilinear Transforms", *IEEE Transactions on Speech and Audio Processing*, vol. 7, no. 6, pp. 697-708, 1999.
15. Hyung-Min Park, Sang-Hoon Oh, Soo-Young Lee, A Bark-scale filter bank approach to independent component analysis for acoustic mixtures, *Journal on Neuro computing*, Elsevier, vol. 73, issues 1–3, pp. 304-314, December 2009.
16. J. Rajnoha, P. Pollak, "Modified feature extraction methods in robust speech recognition", in Proceedings of the 17th IEEE International Conference on Radio elektronika, pp.1–4, 2007.
17. Dr. Shaila D. Apte, "Speech Processing Applications", in Speech and Audio Processing, Section 1, Section 2 and Section 3, pp. 1-6, 67, 91-92, 105-107, 129-132, Wiley India Edition.
18. VoiceboxToolbox, http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html
19. Amol A Chaudhari and S B Dhonde. "Effect of Varying MFCC Filters for Speaker Recognition", *International Journal of Computer Applications* 128(14):7-9, October 2015. Published by Foundation of Computer Science (FCS), NY, USA.
20. J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgren, V. Zue, TIMIT acoustic-phonetic continuous speech corpus, http://catalog.ldc.upenn.edu/ldc93s1, 1993.

21. Department of Computer Engineering, Dr. Babasaheb Ambedkar Marathwada University.
22. M.A. Yusnita, M.P. Paulraj, Sazali Yaacob, M. Nor Fadzilah, A.B. Shahriman, "Acoustic Analysis of Formants Across Genders and Ethnical Accents in Malaysian English Using ANOVA", *Procedia Engineering*, Elsevier, vol. 64, pp. 385-394, ISSN 1877-7058, 2013.

AUTHORS PROFILE

Dr. S. B. Dhonde has completed his Ph.D. in Electronics Engineering from Dr. Babasaheb Ambedkar Marathwada University, Aurangabad. He has 19 Years of experience in teaching, research and industry. His area of interest includes speech signal processing, computer networking, wireless sensor network. Currently, He is working as an Associate Professor in A.I.S.S.M.S. Institute of Information Technology, Pune. He has published several papers in the area of speech signal processing, networking, embedded system at international conference and journals.

Mr. Amol A. Chaudhari has completed his M.E. in E & TC Engineering from Savitribai Phule Pune University. His area of interest includes speech signal processing and Embedded Systems. Currently, He is working as an Assistant Professor in A.I.S.S.M.S. IOIT, Pune. He has published several papers in the area of speech signal processing including speaker identification.

Mr. M. P. Gajare has completed M.E. and pursuing Ph.D. from Savitribai Phule Pune University. His area of interest includes signal processing and CMOS VLSI. Currently, He is working as an Assistant Professor in A.I.S.S.M.S. IOIT, Pune. He has published several papers in the area of CMOS.