


Novel Gammatone Filterbank Based Spectro-Temporal Features for Robust Phoneme Recognition

Ankit Nagpal^(✉)  and Hemant A. Patil 

Dhirubhai Ambani Institute of Information and Communication Technology,
Gandhinagar, Gujarat, India
{ankit_nagpal,hemant_patil}@daiict.ac.in

Abstract. Recently, Automatic Speech Recognition (ASR) technology is being used in practical scenarios and hence, robustness of ASR is becoming increasingly important. State-of-the-art Mel Frequency Cepstral Coefficients (MFCC) features are known to be affected by acoustic noise whereas physiologically motivated features such as spectro-temporal Gabor filterbank (GBFB) features intend to perform better in signal degradation conditions. The spectro-temporal GBFB feature extraction incorporates mel filterbank to mimic frequency mapping in the Basilar Membrane (BM) in the inner ear. In this paper, Gammatone filterbank is used and a comparison is done between GBFB with mel filterbank ($GBFB_{mel}$) features and GBFB with Gammatone filterbank ($GBFB_{Gamm}$) features. MFCC features and Gammatone Frequency Cepstral Coefficients (GFCC) features are concatenated with $GBFB_{mel}$ and $GBFB_{Gamm}$ features, respectively, to improve recognition performance. Experiments are carried out to calculate phoneme recognition accuracy (PRA), on TIMIT database (without ‘sa’ sentences), with additive white, volvo and high frequency noises at various SNR levels from -5 dB to 20 dB. Results show that, with acoustic modeling only, proposed feature set ($GBFB_{Gamm}+GFCC$) performs better (in terms of PRA %), than $GBFB_{mel}+MFCC$ features by an average of 1% , 0.2% and 0.8% for white, volvo and high frequency noises, respectively.

Keywords: Robust ASR · Gabor filterbank (GBFB) features · Gammatone filterbank · MFCC · Acoustic model · Language model

1 Introduction

Automatic Speech Recognition (ASR) is being used in practical scenarios which involves various noises and channel effects. Decades of research has brought several methods to improve performance of ASR system by increasing robustness against variability of speech signals. Methods include capturing of temporal cues from the speech signal (TempoRAI Patterns (TRAPS) [1]), spectral information from the speech signal (Mel Frequency Cepstral Coefficients (MFCC) [2] and Perceptual Linear Prediction (PLP) [3]). MFCC features are concatenated with their

first and second order temporal derivatives (i.e., delta and double-delta features), to capture temporal dynamics in the speech signal. It resulted in improvement in ASR performance and hence became a big motivation to use joint spectro-temporal features for ASR task. Another motivation to use spectro-temporal features in ASR is the fact that our brain responds to joint spectro-temporal patterns in the speech signal rather than temporal-only or spectral-only patterns [10]. Biological studies indicate that neurons in the primary auditory cortex (*A1*) of mammals are explicitly tuned to spectro-temporal patterns [4] and different neurons are excited by different spectro-temporal patterns depending upon their Spectro-Temporal Receptive Fields (STRFs). Hence, it would be worthwhile to explore and analyze spectro-temporal features of speech signal since these features are physiologically motivated and it is already known that human speech recognition system is better than any ASR system. The shape of STRF of a neuron looks like a $2-D$ Gabor filter as shown in Fig. 1(b) [8]. Arrow indicates highly varying $2-D$ impulse response region, where red and blue colors indicate region of strongly excitatory and suppressed responses. The stages of speech processing, from the signal entering the ear, till brain, is shown in Fig. 1(a) [7]. The final output of the speech processing is the response of the neuron in the *A1*, known as cortical representation, which the brain understands. The neural response is the convolution of the input time-frequency representation of speech signal with the STRF of the neuron (called as the cortical stage).

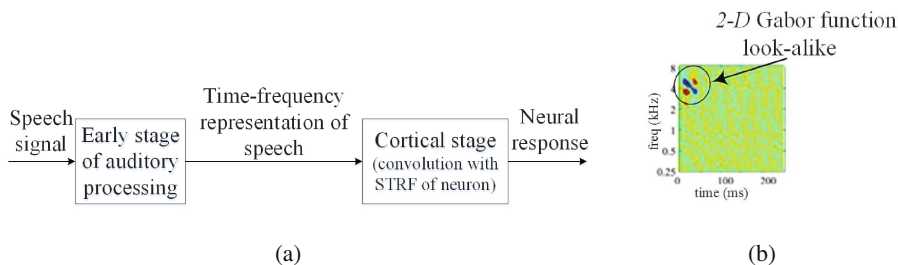


Fig. 1. (a) Speech processing stages in humans. (b) STRF of a neuron in *A1*. Adapted from [8]. All figures in the paper are best viewed in color.

Schadler *et al.* [4] tried to mimic speech recognition of mammals, in ASR task. Algorithm takes log-mel spectrogram (spectro-temporal patterns as input to neurons in *A1*) and passes it through a bank of $2-D$ Gabor filters (real part of Gabor filters, as $2-D$ impulse response of neurons known as STRFs) to generate corresponding time-frequency representations, known as the cortical representations. In this paper, Gammatone filters [5] are used instead of mel filters to generate spectro-temporal Gabor filterbank (GBFB) features with Gammatone filterbank ($GBFB_{Gamm}$), in contrast to spectro-temporal Gabor filterbank (GBFB) features with mel filterbank ($GBFB_{mel}$). Gammatone Frequency Cepstral Coefficients (GFCC) are seen to perform better than MFCC [6]

and hence, are concatenated with $GBFB_{Gamm}$ features to improve recognition performance. We have analyzed the performance of the proposed features (i.e., $GBFB_{Gamm}+GFCC$) on TIMIT database [9] with different additive noises such as white, volvo and high frequency noises, at various SNR levels. Performance of features is compared with $GBFB_{mel}+MFCC$ features and MFCC features alone. Experiments are carried out taking into consideration the effectiveness of Language Model (LM), with HTK as back end [12].

The rest of the paper is organized as follows. Section 2 describes the spectro-temporal feature extraction algorithm in detail. Section 3 contains the experimental results and finally, Sect. 4 concludes the paper along with future research directions.

2 Spectro-Temporal Feature Extraction

Figure 2 shows the architecture for spectro-temporal feature extraction from the speech signal. Log-Gammatone spectrogram is passed through 2-D Gabor filterbank to generate time-frequency representations corresponding to the Gabor filters. These time-frequency representations are combined and dimensionality is reduced to form GBFB with Gammatone filterbank ($GBFB_{Gamm}$) features.

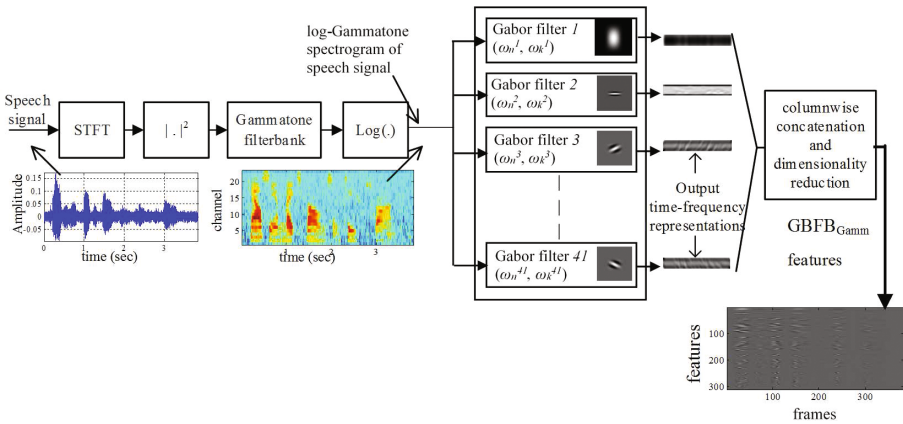


Fig. 2. Architecture for spectro-temporal feature extraction.

2.1 Log-Gammatone Spectrogram as Input to Gabor Filterbank

Gammatone filterbank is commonly used filterbank to simulate the motion of the basilar membrane in the cochlea. Slaney’s Auditory toolbox [5] is used to generate the Gammatone filterbank. Spectrogram is expressed as:

$$S(k, \tau) = \sum_{n=0}^{N-1} x[n]w[n, \tau] \exp^{-j2\pi kn/N}, \quad (1a)$$

$$X(k, \tau) = |S(k, \tau)|^2, \tag{1b}$$

where $x[n]$ is the speech signal, $w[n]$ is the window function, τ is the time frame, N is the window length in samples and $S(k, \tau)$ is the short-time Fourier transform (STFT). Spectrogram is represented by $X(k, \tau)$. Figure 3 shows log-Gammatone spectrogram for the segment of clean speech signal, dr1_fdac1_sx394_te from TIMIT database (with sampling frequency 16 kHz) and for additive white noise with 5 dB SNR. Parameters used for calculating log-Gammatone spectrogram are window (Hanning) length = 25 ms, window shift = 10 ms, number of channels/subband filters in Gammatone filterbank = 23, with center frequencies ranging from 100 Hz to 8000 Hz. Figure 3 clearly indicates that joint spectro-temporal intensity pattern in the noisy signal has varied significantly from that of the clean version and thus recognizing speech from a noisy speech signal is indeed a challenging task.

2.2 Gabor Filterbank

The localized complex Gabor filters are defined in (2a), b and c, with the channel and time-frame variables k and n , respectively; ω_k and ω_n the spectral and the temporal modulation frequencies respectively; v_k and v_n the number of semi-cycles under the envelope in spectral and temporal dimension. A Gabor filter is the product of a complex sinusoid carrier (2b) with the corresponding modulation frequencies ω_k and ω_n , and an envelope function defined in (2a).

$$h_b(x) = \begin{cases} 0.5 - 0.5 \cos(2\pi x/b), & 0 < x < b \\ 0, & \text{else} \end{cases}, \tag{2a}$$

$$s_\omega(x) = e^{j\omega x}, \tag{2b}$$

$$g(n, k) = s_{\omega_k}(k) s_{\omega_n}(n) h_{\frac{v_k}{2\omega_k}}(k - p) h_{\frac{v_n}{2\omega_n}}(n - q), \tag{2c}$$

where p and q represent the shift in the envelope of the Gabor filter to align the filter at the origin. The above definition would lead to infinite support for purely temporal or purely spectral modulation ($\omega_k = 0$ or $\omega_n = 0$) filters. Thus, filter size is limited to 69 channels and 40 time frames.

There is a linear relationship between the modulation frequency and the extension of the envelope (Eq. (2a), b and c) and hence all the filters with same values for v_k and v_n are constant Q (i.e., quality factor) filters. DC bias of each filter is removed since relative energy fluctuations are important for speech classification. Mean removal on a logarithmic scale is same as dividing on a linear scale and thus this corresponds to a normalization. While cepstral coefficients normalize spectrally, and RASTA (Relative Spectra) [11] processing and discrete derivatives normalize temporally, DC-free Gabor filters naturally normalize in both directions.

Temporal modulation frequencies up to 16 Hz and spectral modulation frequencies up to 0.5 cycle/channel are most sensitive to humans [10] and therefore, best performance is attained if maximum modulation frequencies of the filters are

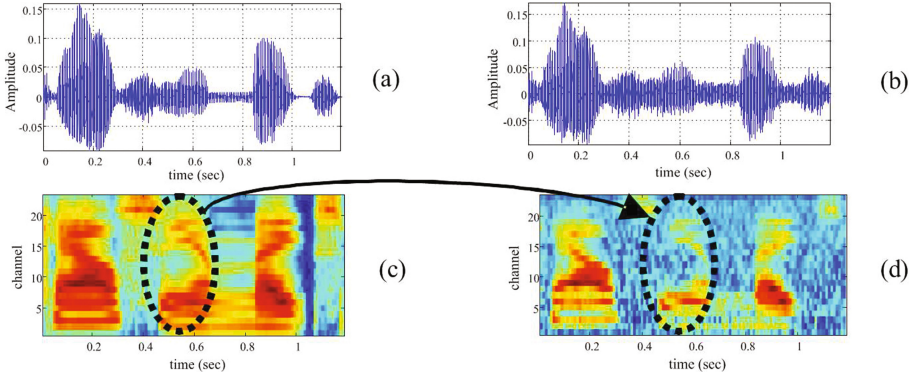


Fig. 3. (a) Segment of the clean speech signal dr1_fdac1_sx394_te from TIMIT database ($F_s = 16$ kHz), (b) signal in (a) with additive white noise at 5 dB SNR-level, (c) log-Gammatone spectrogram of clean signal, (d) log-Gammatone spectrogram of the noisy speech signal.

around these values. Empirically, we found that maximum modulation frequencies of 12.5 Hz and 0.25 cycle/channel produced the best performance. With the aim of evenly covering the modulation transfer space, modulation frequencies of the filterbank are decided as in (3a, b).

$$\omega_x^{i+1} = \omega_x^i \frac{1 + c/2}{1 - c/2}, \tag{3a}$$

$$c = d_x \frac{8}{v_x}, \tag{3b}$$

where d_x (in x -domain) is the distance factor between the two adjacent filters. Gabor filters with following frequencies are considered.

$$\begin{aligned} \omega_k &= -0.25, -0.12, -0.06, -0.03, 0, 0.03, 0.06, 0.12, 0.25, \\ \omega_k &= 0, 3.09, 4.92, 7.84, 12.5, \end{aligned}$$

in cycles/channel and Hz, respectively. Hence, 41 unique 2-D spectro-temporal Gabor filters are achieved whose real parts are used to process the log-Gammatone spectrogram of the speech signal. The parameters for Gabor filterbank used here are given in Table 1. These parameters, empirically, found to perform the best and are thus used for the speech recognition task considered in this paper.

2.3 Output of the Gabor Filter

2-D convolution of log-Gammatone spectrogram is done with the real part of the Gabor filter to get time-frequency representation that contains patterns matching the modulation frequencies associated with the filter (Fig. 4). The dimension

Table 1. Parameters used for Gabor filterbank

| Parameter | ω_n (max) | ω_k (max) | v_n | v_k | d_n | d_k |
|-----------|------------------|------------------|-------|-------|-------|-------|
| Value | 12.5 | 0.25 | 3.5 | 3.5 | 0.2 | 0.3 |

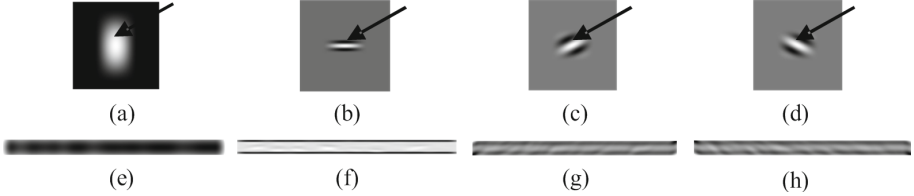


Fig. 4. Four Gabor filters with (ω_n, ω_k) as $(0,0)$, $(0,0.12)$, $(3.09,-0.06)$, $(3.09,0.06)$ in (a), (b), (c) and (d), respectively. Corresponding output using log-Gammatone spectrogram of speech signal dr1_fdac1_sx394_te from TIMIT database, with white noise added at 5 dB SNR, in (e), (f), (g) and (h). Gabor filters parameters used are $v_n = v_k = 3.5$, $d_n = 0.2$ and $d_k = 0.3$.

of single filter's output time-frequency representation is same as that of the log-Gammatone spectrogram of the speech signal, i.e., 23 (number of Gammatone channels) \times number of frames of the speech signal. Outputs of all the 41 Gabor subband filters are concatenated columnwise to form the features. Figure 4 shows some Gabor filters with different combinations of modulation frequencies (ω_n, ω_k) and corresponding outputs of noisy speech signal (generated by adding white noise at 5 dB SNR, to clean speech signal from TIMIT database). The orientation of the Gabor filters are depicted by arrows, indicating that different combination of modulation frequencies (ω_n, ω_k) leads to different orientation of the Gabor filter.

The resultant concatenated output would be quite high-dimensional ($23 \times 41 = 943$). To reduce computational complexity, dimensionality needs to be reduced. Dimensionality is reduced by exploiting the fact that the filter output between adjacent channels is highly correlated when the subband filter has a large spectral extent. Thus, channel selection scheme as discussed in [4] is applied to the complete feature matrix and dimensionality is reduced to 311 . Since, Gabor filter size is limited to 40 time frames, these features encode upto 400 ms (40×10 ms window duration) context while MFCC features encode upto 45 ms context. To improve recognition performance, GFCC features [6] are concatenated with GBFB_{Gamm} features to give GBFB_{Gamm}+GFCC features which results in the dimension of 350 (i.e., $311 + 39 = 350$).

3 Experimental Results

Recognition experiments are conducted on TIMIT database with additive white, volvo and high frequency noises at various SNR levels ranging from 20 dB to -5 dB. Core training sentences (3696) and core testing sentences (192) of TIMIT

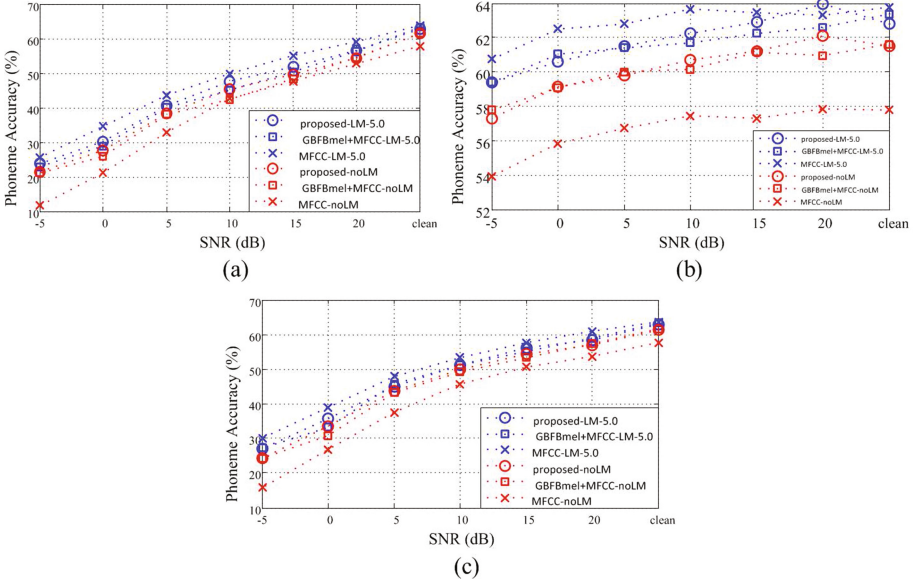


Fig. 5. Comparison of phoneme-level accuracy (in %) between the proposed features, GBFB_{mel}+MFCC features and MFCC features for LM 5.0 and with no LM, for additive white, volvo and high frequency noises at various SNR levels in (a), (b) and (c), respectively.

database are used in the experiments. For our experiments, training and testing environments are kept same. Hidden Markov Model (HMM) is used as the back end and phoneme-level accuracy, as given in (4), is used as the performance measure with one phoneme modeled by 5 states and each state modeled by mixture of 8 Gaussians. HTK is used to carry out the experiments. The % phoneme recognition accuracy (*PRA*) is defined as [12]:

$$\%PRA = \frac{N - D - S - I}{N} \times 100, \tag{4}$$

where N is the total number of labels (phonemes) in the reference transcriptions, S is the substitution errors, D is the deletion errors and I is the insertion errors. A comparison between proposed features, i.e., GBFB_{Gamm} concatenated GFCC (GBFB_{Gamm}+GFCC, dimension = 350) features, GBFB_{mel} concatenated MFCC (GBFB_{mel}+MFCC, dimension = 350) features and MFCC features (dimension = 39) is shown in Fig. 5, for additive white, volvo and high frequency noises, for various SNR levels. Experiments are conducted with 5.0 weighted LM and for without LM. When experimented with 5.0 weighted language model (LM), it is found that MFCC features perform better than the other two features for clean and noisy environments with SNR ranging from 20 dB to -5 dB. For SNR = ∞ (clean conditions), 20 dB, 15 dB, 10 dB, 5 dB,

0 dB, -5 dB, MFCC features perform better (in terms of *PRA* %) than the proposed features by an average (computed over various SNR levels from -5 dB to 20 dB) of 2.6% , 1% and 2.3% and perform better than GBFB_{mel}+MFCC by an average of 3.5% , 1.2% and 2.5% , for additive white, volvo and high frequency noise, respectively. Thus, with 5.0 LM, the proposed features perform better than GBFB_{mel}+MFCC by an average of 0.9% for white noise and 0.2% for volvo and high frequency noises. When experimented without incorporating LM, it is seen that the proposed features outperform both MFCC and GBFB_{mel}+MFCC under signal degradation conditions. For signal degradation conditions, the proposed features perform better than MFCC by an average of 4.6% , 3.5% and 5.4% and perform better than GBFB_{mel}+MFCC by an average of 1% , 0.2% and 0.8% for white, volvo and high frequency noise, respectively. Under clean conditions, without LM, the proposed features perform almost similar to GBFB_{mel}+MFCC features but perform better than MFCC features by 3.7% . It can be observed that, with acoustic modeling only, spectro-temporal Gabor filterbank (GBFB) features (whether incorporating Gammatone filterbank or mel filterbank) when concatenated with cepstral coefficients perform better than the state-of-the-art MFCC features in clean conditions as well as in the presence of various additive noises. This is because GBFB features are able to capture more local joint spectro-temporal information in the speech signal. In addition, when Gammatone filterbank is used instead of mel filterbank, to extract GBFB features, the recognition performance under signal degradation conditions (SNR ranging from 20 dB to -5 dB), is improved.

4 Summary and Conclusions

With acoustic modeling only, the spectro-temporal GBFB features when concatenated with cepstral coefficients perform better than the state-of-the-art MFCC features because of the fact that GBFB features are able to capture more local joint spectro-temporal information in the speech signal (by passing spectrogram of speech through various 2 -D Gabor subband filters aligned at modulation frequencies important for speech intelligibility). Thus, spectro-temporal features are preferred for the languages/ databases which do not have enough accurate language models (due to scarcity of training data). When Gammatone filterbank is used instead of the standard mel filterbank, the recognition performance of the spectro-temporal features is improved. Future work will be to reduce the dimension of such high-dimensional spectro-temporal features and to see the effect of context window of the features (defined by temporal dimension of the Gabor filter) on the recognition performance.

References

1. Hermansky, H., Sharma, S.: Temporal patterns (TRAPS) in ASR of noisy speech. In: Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP), Phoenix, Arizona, USA, vol. 1, pp. 289–292 (1999)

2. Davis, S.B., Mermelstein, P.: Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoust. Speech Signal Process.* **28**(4), 357–366 (1980)
3. Hermansky, H.: Perceptual linear predictive (PLP) analysis of speech. *J. Acoust. Soc. Am.* **87**(4), 1738–1752 (1990)
4. Schadler, M., Meyer, B., Kollmeier, B.: Spectro-temporal modulation subspace spanning filter bank features for robust automatic speech recognition. *J. Acoust. Soc. Am.* **131**(5), 4134–4151 (2012)
5. Slaney, M.: Auditory Toolbox, version 2. <http://engineering.purdue.edu/malcolm/interval/1998-010/>. Accessed 7 Apr 2015
6. Shao, Y., Jin, Z., Wang, D., Srinivasan, S.: An auditory-based feature for robust speech recognition. In: *Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Taipei, Taiwan, pp. 4625–4628 (2009)
7. Chi, T., Ru, P., Shamma, S.A.: Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **118**(2), 887–906 (2005)
8. Depireux, D.A., Ru, P., Shamma, S.A., Simon, J.Z.: Response-field dynamics in the auditory pathway. In: *Computational Neuroscience*, pp. 1–6 (1998)
9. Lee, K., Hon, H.: Speaker-independent phone recognition using hidden Markov Models. *IEEE Trans. Acoust. Speech Signal Process.* **37**(11), 1642–1648 (1989)
10. Chi, T., Gao, Y., Gutyon, M.C., Ru, P., Shamma, S.A.: Spectro-temporal modulation transfer functions and speech intelligibility. *J. Acoust. Soc. Am.* **106**(5), 2719–2732 (1999)
11. Hermansky, H., Morgan, N.: RASTA processing of speech. *IEEE Trans. Speech Audio Process.* **2**(4), 578–589 (1994)
12. Young, S.J., Evermann, G., Gales, M.J.F., et al.: *The HTK book for HTK version 3.4*. Microsoft Corporation (2006)