



# Selective Gammatone Filterbank Feature for Robust Sound Event Recognition

Yi Ren Leng<sup>1</sup>, Huy Dat Tran<sup>1</sup>, Norihide Kitaoka<sup>2</sup>, Haizhou Li<sup>1</sup>

<sup>1</sup>Human Language Technology Department, Institute for Infocomm Research, A\*STAR, Singapore 138632

<sup>2</sup>Nagoya University, Japan

urleng@i2r.a-star.edu.sg, hdtran@i2r.a-star.edu.sg, kitaoka@nagoya-u.jp, hli@i2r.a-star.edu.sg

## Abstract

This paper introduces a novel feature based on the raw output of the gammatone filterbank. Channel selection is used to enhance robustness over a range of signal-to-noise ratios (SNR) of additive noise. The recognition accuracy of the proposed feature is tested on a sound event database using a Hidden Markov Model (HMM) recogniser. A comparison with a series of similar features and the conventional Mel-Frequency Cepstral Coefficients (MFCC) shows that the proposed feature offers significant improvement in low SNR conditions.

**Index Terms:** gammatone filterbank, Hidden Markov Model, robust recognition, sound event recognition

## 1. Introduction

The gammatone filterbank is designed to be an auditory model for the human basilar membrane due to the similarity between its frequency response and the frequency selectivity of the human auditory system [1]. Most of the previous work involving the gammatone filterbank is conducted from the physiological perspective, relating to its similarity to the human or mammalian auditory system. In speech recognition, gammatone filterbanks are usually used in cepstral-like features involving the Discrete Cosine Transform (DCT) ([2],[3]) or as physiologically-inspired feature systems [4]. In contrast, little work has been done in applying the gammatone filterbank for sound event recognition.

Sound event recognition is the ability to detect and classify sound events such as doors closing, the breaking of glass and footsteps on a hard surface. Similar to speech recognition, sound event recognition is useful for a large range of tasks such as audio searching [5], hearing aids [6] and home automation [7]. Sound events can be modelled like isolated words in order to tap on the well-established speech recognition methods. A notable exception is the inability to separate the sound events into smaller components or analogs of sub-word units which prevents the use of a language model for general recognition tasks. We choose to use a Hidden Markov Model (HMM) recogniser as it can produce good results without extensive calibration and optimisation. A major limitation of automatic recognition system is the deterioration of recognition accuracy with increasing levels of environmental or background noise. In this paper, we shall only consider the effect of additive noise as it is the easiest to quantify. Most of the existing systems only perform well in matching conditions where the system is trained and tested in a similar setting. We propose a form of channel selection that retains most of the recognition accuracy over a range of mismatched conditions. This method only requires a sampling of the testing condition to determine its

noise characteristic. The optimal subset of channels of the filterbank output is then automatically selected and used to train the mismatched model.

Figure 1 shows the raw output of a 36 channel gammatone filterbank in noisy (0db) and clean (40db) conditions. The two images share essentially the same distinct peaks from channels 10 and onwards. The only difference is the addition of noise in the lower channels in the noisy condition. Based on this observation, selecting the higher channels that are less

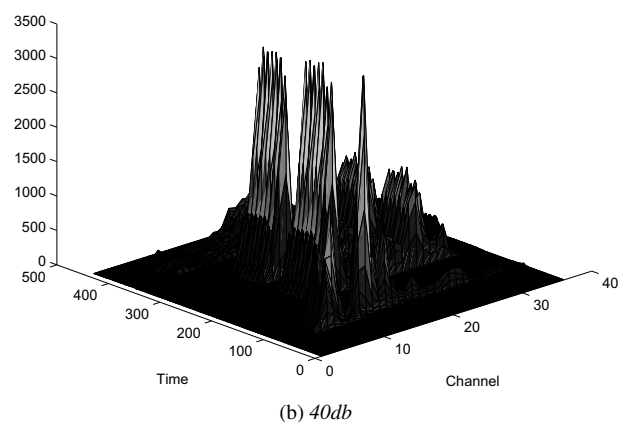
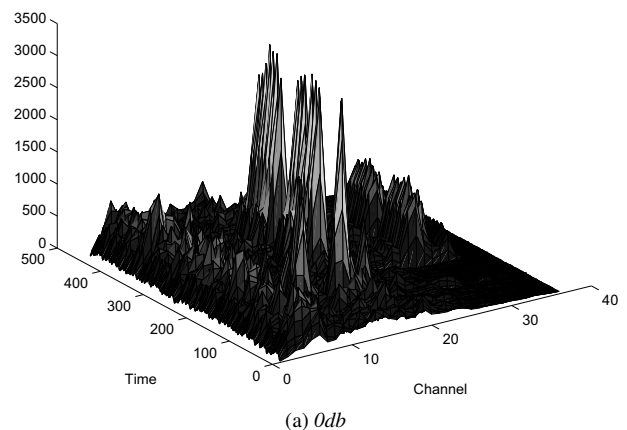


Figure 1: Full 36 channel gammatone filterbank

corrupted by noise should lead to better recognition results.

Logarithm or root functions are usually used to reduce the dynamic range of the raw values in the feature domain. For HMM-based speech recognition systems, this reduction allows for a better representation by the statistical model, thereby leading to improved accuracy. However, we feel that a large dynamic range allows for better distinguishability between the signal and background noise which should lead to better results in noisy conditions.

The discriminative power of a feature utilising the full dynamic range is largely offset by the presence of channels that are highly polluted by noise in conventional features since the HMM makes use of all the present channels for the final output. We propose to use channel selection to reduce or eliminate these noisy channels to retain the channels where the signal is dominant. This should improve the accuracy of the full dynamic range feature by cutting out irrelevant information that only confuses the HMM recogniser.

## 2. Selective Gammatone Filterbank Feature

The proposed feature vector is based on the raw output of the gammatone filterbank applied to the signal in the time domain. Each filter in the filterbank generates a filtered waveform that is full-wave rectified and time-averaged into frames. Using this set of vectors, we apply the t-test distance measure to select the most noise robust vectors or channels. The selected channels are then combined with their delta and double delta components and mean normalised to form the final feature vector.

### 2.1. Gammatone Filterbank

The gammatone function is defined as

$$g(t) = t^{n-1} e^{-bt} \cos(\omega t + \phi) \quad (1)$$

$n$  represents the order of the function,  $b$  is related to the bandwidth of the gammatone filter and  $\phi$  is a phase constant. We shall use a fourth order ( $n = 4$ ) order gammatone filterbank and calculate  $b$  as documented in [8], setting  $\phi = 0$ . The filterbank can be implemented as an eighth order digital filter in the temporal domain after accounting for the complex data. The Laplace Transform of the gammatone function is used to obtain the filter coefficients:

$$G(s) = \frac{6(-b^4 - 4b^3s - 6b^2s^2 - 4bs^3 - s^4) + 6b^2\omega^2 + 12bs\omega^2 + 6s^2\omega^2 - \omega^4}{(b^2 + 2bs + s^2 + \omega^4)^4} \quad (2)$$

The above equation is used to derive four biquad filters for computational stability and implemented in MatLab.

### 2.2. Channel Selection

For robust recognition, the feature vector should appear similar under different noise conditions. Channel selection of the filterbank outputs is done to retain the channels with little variation over noise conditions while discarding the remainder. In order to perform such a selection, a measure is required to gauge the degree of variability over noise. Examples are the Euclidean distance, Mahalanobis distance, chi-square distance and Kullback-Leibler (KL) distance.

We choose the t-test distance as a measure as it closely approximates the subband SNR which is empirically related to robust automated speech recognition performance.

$$d_{ij} = \frac{|\mu_i - \mu_j|}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_j^2}{n_j}}} \quad (3)$$

$\mu$ ,  $\sigma^2$  and  $n$  are the mean, variance and length of the channel output while the index  $i$  refers to the various noise levels and  $j$  to the clean condition. This distance measure is taken over a small subset of the testing database to represent adaptation of the system to the testing condition.

### 2.3. Summary of Feature Extraction

The full gammatone filterbank is designed with a total of 36 filters. This filterbank is applied to 50 sequences from the test database for use in the t-test distance measure. The filterbank output is full-wave rectified and time-averaged into frames using 400ms rectangular windows to create a 36 dimension feature. Mean normalisation is not applied for this feature vector as the t-test measure involves a comparison of the means of the two vectors.

A selective filterbank containing the 12 best channels is created based on the t-test results. The filterbank output is full-wave rectified and time-averaged into frames using 400ms rectangular windows to create a 12 dimension feature. This feature is then augmented by the delta and double delta components to give a 36 dimension vector which is normalised to have zero mean in each dimension. The proposed method is less computation-intensive compared to the Mel-Frequency Cepstral Coefficients (MFCC). The only significant step involves the digital filtering of the input waveform with the remainder being simple arithmetic operations.

## 3. Experimental Setup

### 3.1. Sound Event Database

The testing of the features are carried out on a sound event database created by the concatenation of three distinct sound events to form a sequence. A short period of silence is inserted between events and at the start and end of the sequence. To simulate a realistic recording environment, additive noise at 40db is added to the sequences to represent the clean condition. For the testing database, noise is added at 0, 5, 10, 15, 20 and 40db to the sequences. The 40db events are used to represent the clean condition while the others represent noisy data.

A total of thirteen sound events and one speech event make up the pool of sound events. The speech event comprises of the Japanese digits one to five spoken by different female speakers. The events used in the sound event database are constructed from sound files extracted from "Real World Computing Partnership Sound Scene Database in Real Acoustical Environments"<sup>1</sup>[9] and voice files from "Reverberant Speech Recognition Evaluation Environment CENSREC-4"<sup>2</sup> [10]. The additive noise used is obtained from CENSREC-4 and the NOISEX-92<sup>3</sup> database.

The 40db training noise used is the "Japanese style room" from CENSREC-4. The four testing conditions are labelled as "car", "hall", "canteen" and "factory". The additive noise used are "in-car" and "elevator hall" from CENSREC-4 and "speech babble" and "factory floor noise 1" from NOISEX-92 respectively. The 19.98khz noise samples from NOISEX-92 are down-sampled to 16khz to correspond with the rest of the data.

Each sound event contains 100 separate samples, 70 of which are used for the training database and the remaining 30 are used for testing. A total of 980 training sequences and

<sup>1</sup><http://tosa.mri.co.jp/sounddb/indexe.htm>

<sup>2</sup><http://research.nii.ac.jp/src/eng/list/detail.html>

<sup>3</sup>[http://spib.rice.edu/spib/select\\_noise.html](http://spib.rice.edu/spib/select_noise.html)

420 testing sequences are generated such that each individual sample is expected to appear thrice in the database.

### 3.2. Other Features for Comparison

The other features to be compared are described as follows:

1. Full Gammatone:

Instead of selecting the 12 best channels out of 36, an evenly distributed 12 channel gammatone filterbank is applied to all the training and testing samples. The filterbank output is time-averaged into frames.

2. Selective log-Gammatone:

Before the 36 channel filterbank is applied to the 50 test sequences, the natural logarithm of the filterbank output is taken. To avoid errors with values close to or equal to zero when taking the log, a minimum bound is set for the filterbank output.

3. Mel-Frequency Cepstral Coefficients (MFCC):

12 MFCC are created using the Matlab function *melcepst* from “VOICEBOX: Speech Processing Toolbox for MATLAB”<sup>4</sup> with 24 Mel filters

4. Selective Mel Power:

The Matlab function *melcepst* is modified to give the Mel power using 36 Mel filters by skipping the Discrete Cosine Transform and natural logarithm. From these 36 channels, the t-test is used to find the 12 best channels.

Like the selective gammatone filterbank feature, each of the above four features has 12 dimensions. Following the same processing, they are augmented with the delta and double delta components and normalised to have zero mean so that all five features are mean normalised and have 36 dimensions.

### 3.3. Hidden Markov Model (HMM) Recogniser

The statistical models used are trained and tested using the Hidden Markov Model Toolkit<sup>5</sup> (HTK). The recogniser setup is based on AURORA-2J<sup>6</sup>[11] with 18 HMM states. Of the 18 states, only 16 are emitting states with 20 Gaussians for the sound events and 36 Gaussians for the silence model. This configuration is chosen as the modelling of sound events is similar to that for isolated words or digits. The effect of varying the number of states or Gaussians is not studied in this paper.

The word network used for the recogniser is allowed to select any combination of the fourteen sound events with optional silence at the beginning and end of each event and optional short pauses between events. Although both the training and test databases are created with exactly three events per sequence, the recogniser is capable of operating on sequences with an arbitrary number of events. A drawback of this flexibility is the possibility of reporting the wrong number of events due to erroneous insertions or deletions. To compensate for this, the log-insertion penalty of the HTK recogniser program *HVite* is allowed to vary from 0 to -1000 at intervals of -100. The final reported accuracy is the highest reported accuracy over this range of values.

## 4. Results and Discussion

The recognition accuracy is given by the “Word Accuracy Rate” reported by HTK shown in Table 1. The performance of

<sup>4</sup><http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html>

<sup>5</sup><http://htk.eng.cam.ac.uk/>

<sup>6</sup><http://sp.shinshu-u.ac.jp/CENSREC/en/CENSREC/AURORA-2J/>

Table 1: Summary of Recognition Accuracy

#### Car

SNR	Sel. G	Full G	Sel. log G	MFCC	Sel. Mel
0db	81.03	11.35	32.06	40.08	80.71
5db	83.81	19.05	47.86	60.08	83.33
10db	84.60	38.57	65.87	79.29	84.13
15db	84.44	63.57	84.37	88.73	84.05
20db	84.44	78.57	90.48	91.27	84.05
clean	84.44	83.57	90.32	92.06	84.05
avg	83.79	49.11	68.49	75.25	83.39

#### Canteen

SNR	Sel. G	Full G	Sel. log G	MFCC	Sel. Mel
0db	73.73	9.84	26.67	21.75	69.52
5db	82.30	18.33	41.59	44.60	78.65
10db	84.13	42.54	58.57	64.44	81.83
15db	83.97	67.30	75.24	79.44	82.22
20db	83.97	80.00	87.62	86.51	82.22
clean	83.89	83.65	90.40	92.06	82.22
avg	82.00	50.28	63.35	64.80	79.44

#### Hall

SNR	Sel. G	Full G	Sel. log G	MFCC	Sel. Mel
0db	38.57	6.98	8.17	12.38	30.71
5db	66.51	7.38	19.13	26.59	60.95
10db	80.32	10.16	33.33	47.94	79.76
15db	82.86	18.73	50.87	66.03	82.38
20db	83.89	38.41	68.73	84.76	83.97
clean	84.21	83.57	92.30	92.54	84.21
avg	72.73	27.54	45.42	55.04	70.33

#### Factory

SNR	Sel. G	Full G	Sel. log G	MFCC	Sel. Mel
0db	69.52	8.89	12.46	19.92	38.25
5db	78.65	14.68	22.78	40.24	62.70
10db	81.83	30.79	39.92	64.21	79.29
15db	82.22	51.83	57.46	82.62	64.21
20db	82.22	72.86	75.79	89.21	92.06
clean	82.22	83.57	91.03	84.52	84.84
avg	79.44	43.77	49.91	64.71	72.30

the selective gammatone feature shows little variation over the range of SNR. This is a good indication of its noise robustness as there is little deterioration in performance despite the drastic drop in SNR from 40db to 0db.

#### 4.1. Selective Gammatone vs Full Gammatone

The recognition accuracy of the full gammatone feature improves with SNR but is inferior to that of the selective feature at all SNR. This implies that the raw gammatone filterbank output consists largely of redundant information as the selective feature only utilises a third of the available channels to produce superior results. Although a comparison of the full 36 channel filterbank might be more appropriate in this situation, tripling the number of dimensions introduces its own problems in addition to the increased computational load.

The inclusion of redundant or poor channels only serves to reduce recognition accuracy as the performance of the full gammatone feature is only comparable to that of the selective feature under clean conditions. This reinforces the choice of applying channel selection to the raw gammatone filterbank output as it packs more useful information in the given feature space.

#### 4.2. Selective Gammatone vs Selective log-Gammatone

The recognition accuracy of the selective log-gammatone feature increases with SNR like the full gammatone, albeit with significantly higher accuracy. The log feature only surpasses the base selective feature near clean conditions.

The log function is known to reduce the dynamic range of the raw feature such that dissimilar values are brought closer to each other. This serves to enhance the performance of matching conditions as minor differences between the training and test databases are attenuated. However, it also reduces the difference in noisy regions where distinct peaks in the raw feature are greatly smoothened. The net effect of the log function is to improve the matching condition performance while reducing the mismatched condition performance.

#### 4.3. Selective Gammatone vs MFCC

Among the features with an increasing accuracy with SNR, the MFCC feature offers the best average performance. Like the selective log-gammatone feature, the MFCC feature is inferior to the selective gammatone feature at low SNR.

Assuming that the Mel power of the signal  $X(\omega)$  and noise  $N(\omega)$  are uncorrelated ( $X(\omega)N(\omega) = 0$ ), the Mel power of the total signal  $S(\omega)$  is given by:

$$S(\omega) = X(\omega) + N(\omega) \quad (4)$$

Taking the natural log entangles the signal and noise

$$\log S(\omega) = \log [X(\omega) + N(\omega)] \quad (5)$$

The Discrete Cosine Transform (DCT) is unable to separate the two components thus the noise is spread over the entire cepstral domain. However, this spreading appears to improve on the robustness of the feature as compared to the selective log-gammatone feature which is similar in principle, minus the additional DCT step.

#### 4.4. Selective Gammatone vs Selective Mel Power

The recognition accuracy of the selective Mel power feature appears to be slightly worse than the selective gammatone feature but is otherwise exactly the same.

This result suggests that using the raw filterbank output in conjunction with channel selection, regardless of the filterbank itself, leads to noise robust features.

### 5. Conclusion

Our experiments have shown that the raw gammatone filterbank output with channel selection gives a feature vector that is highly noise robust. With a generic HMM recogniser, this feature showed a significant improvement of up to 100% over the baseline MFCC feature at low SNR conditions. The comparisons with the full filterbank feature and log feature justified the use of channel selection and the raw filterbank output respectively.

The main drawback of the proposed selective gammatone filterbank feature is the reduced accuracy in matching condition as compared to features with a reduced dynamic range. Supplementing the robust features with more conventional features like the MFCC should be considered for a flexible feature that is effective in all situations.

The results with the selective Mel power feature showed that the choice of filterbank affects the recognition result

while maintaining the robustness of the proposed feature. By optimising the filterbank to the recognition task to be performed, it should be possible to improve on the robust recognition accuracy. The HMM parameters used in our experiment were originally designed for isolated word recognition experiments. It is worthwhile to study the optimal configuration of the system for sound recognition to account for the differences between words and sound events.

### 6. References

- [1] Roy D. Patterson and John Holdsworth, "A functional model of neural activity patterns and auditory images," *Advances in Speech, Hearing and Language Processing*, vol. 3, pp. 547–563, 1996.
- [2] R. Schlüter, I. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," *Proc. IEEE ICASSP*, vol. 4, pp. 649–652, 2007.
- [3] Yang Shao, Zhaozhang Jin, DeLiang Wang, and Soundararajan Srinivasan, "An auditory-based feature for robust speech recognition," *ICASSP Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 00, pp. 4625–4628, 2009.
- [4] Mario E. Munich and Lin Qiguang, "Auditory image model features for automatic speech recognition," *INTERSPEECH-2005*, pp. 3037–3040, 2005.
- [5] Erling Wold, Thom Blum, Douglas Keislar, and James Wheaton, "Content-based classification, search, and retrieval of audio," *IEEE MultiMedia*, vol. 3, no. 3, pp. 27–36, september 1996.
- [6] In-Chul Yoo and Dongsuk Yook, "Automatic sound recognition for the hearing impaired," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 4, pp. 2029–2036, November 2008.
- [7] J. C. Wang, H. P. Lee, J. F. Wang, and C. B. Lin, "Robust environmental sound recognition for home automation," *IEEE Transactions on Automation Science and Engineering*, vol. 5, no. 1, pp. 25–31, January 2008.
- [8] Malcom Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," *Apple Computer Technical Report #35*, 1993.
- [9] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene data collection in real acoustical environments," *J. Acoust. Soc. Japan*, vol. 20, no. 3, pp. 225–231, 1999.
- [10] M. Nakayama, T. Nishiura, Y. Denda, N. Kitaoka, K. Yamamoto, T. Yamada, S. Tsuge, C. Miyajima, M. Fujimoto, T. Takiguchi, S. Tamura, T. Ogawa, S. Matsuda, S. Kuroiwa, K. Takeda, and S. Nakamura, "Censrec-4: Development of evaluation framework for distant-talking speech recognition under reverberant environments," *INTERSPEECH 2008*, pp. 968–971, Sept 2008.
- [11] Satoshi Nakamura, Kazuya Takeda, Kazumasa Yamamoto, Takeshi Yamada, Shingo Kuroiwa, Norihide Kitaoka, Takanobu Nishiura, Akira Sasou, Mitsunori Mizumachi, Chiyomi Miyajima, Masakiyo Fujimoto, and Toshiki Endo, "Aurora-2j, an evaluation framework for japanese noisy speech recognition," *IEICE Transactions on Information and Systems*, vol. E88-D, no. 3, pp. 535–544, March 2005.