PAPER Special Section on Recent Advances in Multimedia Signal Processing Techniques and Applications

Selective Gammatone Envelope Feature for Robust Sound Event Recognition

Yi Ren LENG[†a)], **Huy Dat TRAN**[†b)], ***Nonmembers***, **Norihide KITAOKA**[††c)], ***and*** **Haizhou LI**[†d)], ***Members***

SUMMARY Conventional features for Automatic Speech Recognition and Sound Event Recognition such as Mel-Frequency Cepstral Coefficients (MFCCs) have been shown to perform poorly in noisy conditions. We introduce an auditory feature based on the gammatone filterbank, the Selective Gammatone Envelope Feature (SGEF), for Robust Sound Event Recognition where channel selection and the filterbank envelope is used to reduce the effect of noise for specific noise environments. In the experiments with Hidden Markov Model (HMM) recognizers, we shall show that our feature outperforms MFCCs significantly in four different noisy environments at various signal-to-noise ratios.

key words: *gammatone filterbank, HMM, robust recognition, sound event recognition*

1. Introduction

Sound Event Recognition (SER) is to classify of generic sound events ([1]–[3]). The sounds to be classified can range from common sounds like door bells and foot steps to specific sounds such as gunshots and explosions. This is analogous to Automatic Speech Recognition (ASR) with the difference being the more diverse range of events to be recognized. This similarity motivates the use of conventional ASR features such as Mel-Frequency Cepstral Coefficients (MFCCs) [4] and auditory features ([5]–[8]) to be used for SER. Possible applications for SER include home automation [9], where door knocks and telephone rings trigger the appropriate response, and security [3], where gunshots and explosions trigger alarms to the relevant authorities.

MFCC-based classifiers have been shown perform poorly in noisy and noise-mismatched conditions [10] despite good results in clean conditions. To address the robustness issues, many noise filtering/compensation methods for ASR have been proposed in literature ([11]–[13]). While these methods have shown improvements of performance under moderate noise conditions, they are less effective under low SNR conditions.

The auditory-motivated gammatone feature [14] is an alternative feature extraction method which has shown better robustness in ASR under noisy conditions ([5]–[8]). The gammatone filterbank is an auditory model that is designed to resemble the basilar membrane in the human ear. The gammatone feature can be used directly as filterbank envelops [5] or as cepstral coefficients ([6]–[8]).

This paper is an extension of [15] where we proposed a novel feature, the Selective Gammatone Envelope Feature (SGEF), which uses some training to select the most robust gammatone filterbank channels to derive the robust feature for each noise condition. The motivation of this method is that the noise spectral shape varies with noise conditions (Fig. 1). The channels of the gammatone filterbank output are the results of applying gammatone filters with different filter center frequencies to the input signal. Each channel can thus be associated with a different center frequency of the filterbank center frequency distribution. Common examples of noisy environments such as a busy shopping mall and a train station do not have evenly distributed frequency spectra. Some frequencies will be less affected by noise due to this uneven distribution. By selecting the filterbank channels that are closest to these frequencies, the overall feature will be more robust to noise.

The principle of performing recognition on a subset of the filterbank channels for speech recognition was proposed in ([16], [17]) where the classification accuracy for each channel is merged with different weights to arrive at a final decision. Our method is different in that we identify the

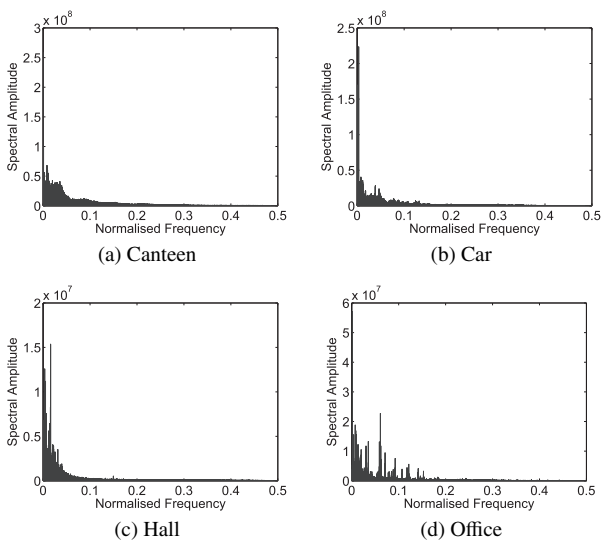


Fig. 1 Frequency spectra for four different noise environments.

Manuscript received May 2, 2011.
Manuscript revised August 8, 2011.

[†]The authors are with the Human Language Technology Department, Institute for Infocomm Research, A*STAR, Singapore.
[††]The author is with Nagoya University, Nagoya-shi, 464–8603 Japan.
a) E-mail: yrleng@i2r.a-star.edu.sg
b) E-mail: hdtran@i2r.a-star.edu.sg
c) E-mail: kitaoka@nagoya-u.jp
d) E-mail: hli@i2r.a-star.edu.sg

DOI: 10.1587/transinf.E95.D.1229

robust channels from a small set of noisy data and combine them into a robust feature. This idea is somewhat similar to Missing Feature Theory (MFT) ([18], [19]) where the reliable elements of a filterbank representation are identified and used for classification. The MFT concept of marginalization, where the unreliable elements are marginalized in the classification process, is close to our approach where the noisy channels are discarded in order to derive a feature comprising of the noise robust channels. Our method does not require modifications of the classifier thus it can be used with conventional recognizers, unlike the MFT approach where the classifier has to account for missing data in the feature.

Although the four noise environments presented in Fig. 1 have different spectra, they are similar in that the high frequency regions are relatively clean of noise. Theoretically, it is necessary to perform channel selection in each new noise environment, followed by feature extraction and recognizer model training. In practice however, if there are similarities between the different noise environments such as the ones we have shown in Fig. 1, the same features and recognizer models can be used. Similarly, the same features and models can be used if the events are not significantly different when changing the number and type of sound events to be classified. This is unlike model adaptation techniques which modify model parameters based on the statistical differences between different noise environments.

The second measure we use to make our feature noise robust is by using the filterbank envelope (raw magnitude) instead of any compression such as logarithm. For matching conditions where the training and testing data are similar, this helps to reduce the differences between the training and testing data, improving the recognition accuracy. In the presence of noise, feature compression tends to confuse the signal with the background noise as the difference between them are reduced. The raw magnitude yields a sparse representation that clearly distinguishes the signal and noise, allowing us to select the cleaner regions with little noise for classification. Figure 2, where the vertical axis represents the 24 channel filterbank spread over the 8 kHz bandwidth, shows the difference between the raw and log filterbank outputs. The top two figures clearly show where the noise is added while in the bottom two figures, the noise appears to have been added everywhere.

In the experiments, we generate random sequences of sound events from 20 sound classes for training and testing, unlike [15] where sequences with a fixed number of events and a smaller set of 14 sound classes are used. Noisy conditions are simulated by adding random segments of four noise clips to the sound event clips. For the recognizer, we use Hidden Markov Models (HMM) ([20], [21]) as they allow for the recognition of sequences of events without the need for a separate event detection step. HMMs are trained in clean conditions and tested on a variety of signal-to-noise ratios (SNRs). From our experiments, we shall demonstrate the noise-robustness of our SGEF.

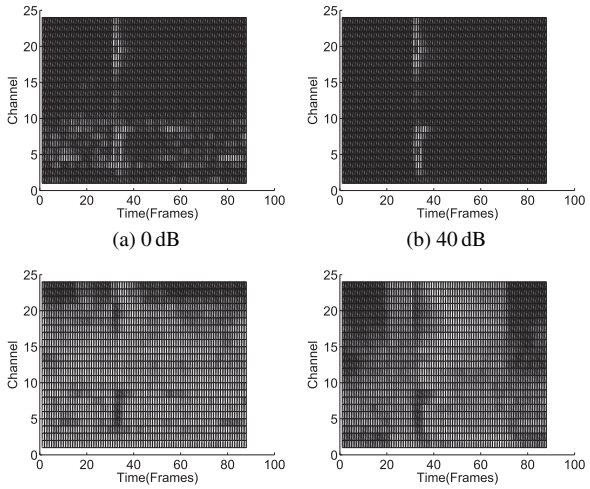


Fig. 2 Raw and log gammatone filterbank outputs in clean (40 dB) and noisy (0 dB) conditions.

2. Selective Gammatone Envelope Feature (SGEF)

Our proposed SGEF is derived from the envelope (raw magnitude) of the gammatone filterbank. We perform channel selection on the filterbank envelope features extracted from a small subset of the noisy test database to determine the set of channels that are least affected by noise. A selective gammatone filterbank, comprising of these noise-robust channels only, is then used for the actual feature. Finally, we perform additional processing on the envelope of the selective gammatone filterbank to generate our SGEF.

2.1 Gammatone Filterbank

The zero-phase n-th order gammatone filter impulse response (Fig. 3) is defined as follows:

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

where t and ω are time and angular frequency respectively and b is proportional to the bandwidth of the filter. We use the 4th order ($n = 4$) gammatone filterbank for our features as described in [22]. This involves taking the Laplace Transform of Eq. (1) to give the following:

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

where s is the Laplace variable. Equation (2) is then used to derive a real 8th order digital filter.

To construct our gammatone filterbank, we generate a distribution of filterbank center frequencies f_i for the i-th filter using Eq. (3) (Fig. 4).

$$f_i = -\alpha + (f_{max} + \alpha)\left[\frac{f_{min} + \alpha}{f_{max} + \alpha}\right]^{\left(1-\frac{i-1}{n}\right)} \tag{3}$$

where $\alpha = 9.26449 \times 24.7$, n is the total number of filters

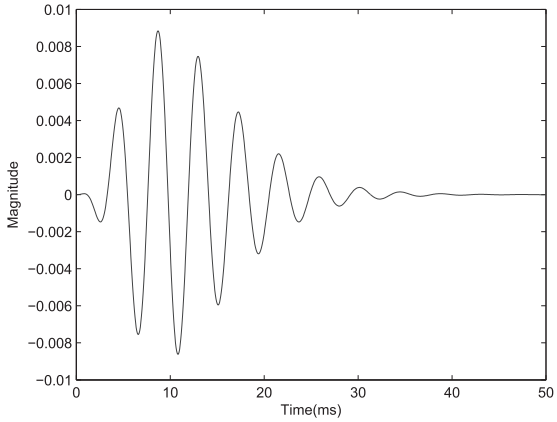


Fig. 3 Impulse response for a 4th order gammatone filter.

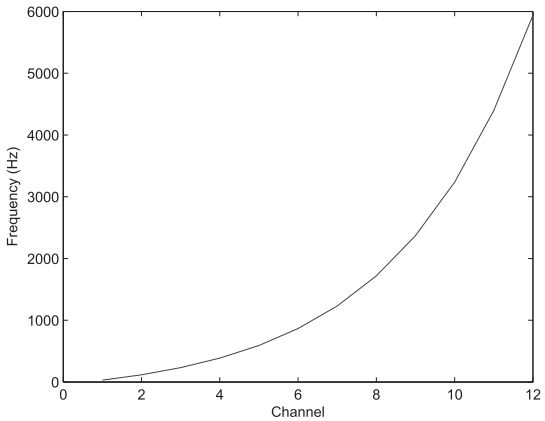


Fig. 4 Center frequencies for a 12 channel filterbank.

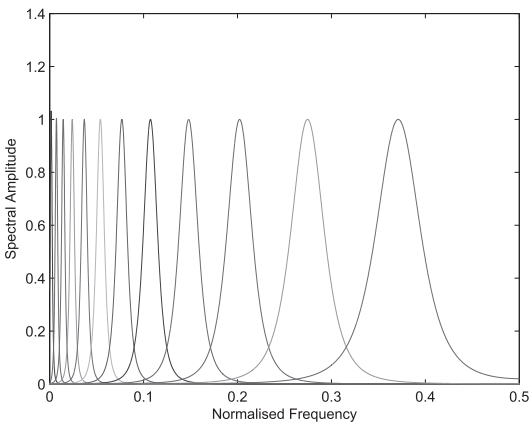


Fig. 5 Frequency response for a 4th order gammatone filterbank.

and f_{min} and f_{max} are the minimum and maximum frequencies in Hertz. The detailed description for our implementation of the gammatone filterbank can be found in [22]. The frequency response for a 12 channel 4th order gammatone filterbank is shown in Fig. 5.

The raw filterbank output $y_i(t)$ is full-wave rectified $E_i(t) = |y_i(t)|$ so that the subsequent time-averaging does not lose any information due to the sinusoidal behaviour of the gammatone filter. To reduce the overall size of the computed feature, the feature is time-averaged into frames $E_i(j)$ using rectangular windows $R_j(t)$. The windowing procedure is similar to the overlapping Hamming windows used for conventional Mel-Frequency Cepstral Coefficients (MFCCs) except the choice of rectangular windows for the SGEF.

The previous steps can be summarized as follows:

1. Calculate filterbank center frequencies f_i using Eq. (3)
2. Derive gammatone filterbank G_i from Eq. (2) with the above center frequencies
3. Compute filterbank output y_i from the signal $x(t)$: $y_i(t) = G_i[x(t)]$
4. Take envelope of the filterbank output: $E_i(t) = |y_i(t)|$
5. Frame the envelope using rectangular windows $R_j(t)$ of width T:

$$E_i(j) = \frac{1}{T}\sum_{t=1}^{T} E_i(t)R_j(t)$$

2.2 Channel Selection

The gammatone filterbank returns n channels of filtered waveforms $E_i(j)$ from a single input waveform $x(t)$. Based on the frequency characteristics of common noisy environments, some channels are less affected by noise than others. With a suitable selection criteria, it is possible to isolate these noise robust channels to construct a feature that is robust to such additive noise.

We choose the t-test distance as the selection criteria as it closely approximates the subband SNR which is empirically related to robust Automatic Speech Recognition performance.

$$d_{ic} = \frac{|\mu_i - \mu_c|}{\sqrt{\frac{\sigma_i^2}{n_i} + \frac{\sigma_c^2}{n_c}}} \tag{4}$$

μ, σ^2 and n are the mean, variance and length of the channel output while the index i refers to the various noise levels and c to the clean condition. Equation (4) makes use of the difference of the means of the filterbank outputs thus it is important to note that these outputs must not be mean normalized. For noise robust channels, there is less difference between the means and variances of the filtered waveforms at different noise levels thus the distance d_{ic} should be minimized.

Starting with $n = 36$, we select the 12 channels with the lowest t-test distance d_{ij} to use for our selective gammatone filterbank. The choice for the number of initial channels is entirely arbitrary while the choice of 12 channels is made to match the number of base cepstral coefficients that are commonly used for Mel-Frequency Cepstral Coefficients (MFCCs).

With reference to the four noise environments in Fig. 1, the computed t-test distances are shown in Fig. 6. The values shown are summed from the first 50 files in the testing

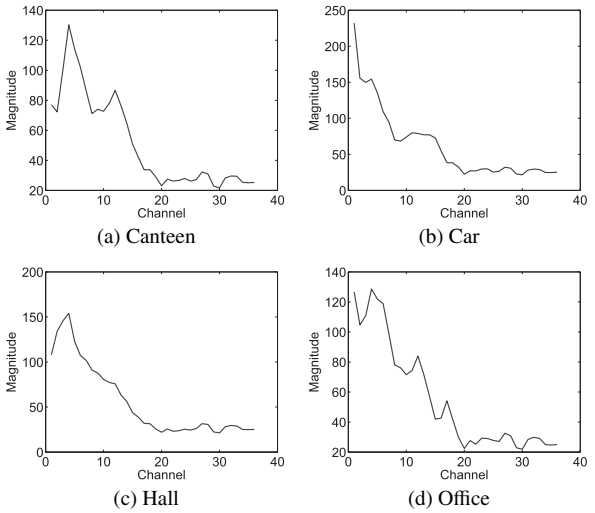


Fig. 6 t-test distance for four different noise environments.

database with i ranging over 5 signal-to-noise ratios (SNRs): 20, 15, 10, 5 and 0 dB. The 12 channels with the lowest t-test distance always occur in the right-hand side of the graphs, implying that the high frequency channels are less corrupted by noise. This can be seen in Fig. 1 where the noise spectra is only significant in the left-hand side of the graphs. The t-test distances also reflect that the characteristics of the noise environment are more important than those of the sound events. If more exotic noise conditions are used, such as ones where high frequency noise is dominant or multiple noise peaks are present, the t-test distance should be able to reflect these characteristics and select the less affected channels.

2.3 Additional Processing

The delta and double delta components are appended to the output of the selective gammatone filterbank to provide more information about the dynamics of the feature [23]. The addition of additive noise has a larger effect on the SGEF compared to log-compressed features like the MFCC as there is no compression factor to reduce the mismatch between noisy conditions. Although channel selection is able to remove the channels most heavily affected by noise, it is impossible to completely eliminate noise from the final output. We apply mean normalization on our features to partially offset this effect. The final feature contains 12 + 12 (*delta*) + 12 (*double delta*) = 36 dimensions.

3. Comparison with Other Features

For all of the features described below, the deltas and mean normalization are applied so that they share the same dimensionality as our proposed feature (36). The frame length and period for each feature are 25 and 10 ms respectively, the same as that used for automatic speech recognition. In order to justify the steps used to derive the Selective Gammatone Envelope Feature (SGEF), we compare our feature with the following features:

3.1 Mel-Frequency Cepstral Coefficients (MFCC)

We choose MFCCs as the baseline for comparison as it is well-established for Automatic Speech Recognition (ASR) [24]. The following steps describe the MFCC extraction procedure:

1. Short-Time Fourier Transform of the Hamming-windowed waveform into a frequency-time representation
2. Using a triangular Mel-filterbank to bin the spectral power and taking the natural logarithm
3. Converting the Log Mel-power vector into cepstral cofficients with the Discrete Cosine Transform (DCT)

The first 12 cepstral coefficients obtained after the DCT, excluding the 0-th coefficient, form the MFCC.

3.2 Full Gammatone Filterbank

Instead of selecting the best 12 channels out of 36, the Full Gammatone Filterbank only has 12 channels ($n = 12$ in Eq. (3)). This comparison studies the effect of channel selection on the gammatone filterbank-based features.

3.3 Selective Log-Gammatone Filterbank

The extraction process for this feature is the same as that for the SGEF, with the sole addition of taking the natural logarithm of the filterbank output. The log is taken both at the channel selection step and for the final selective filterbank. This comparison studies the effect of log-compression on the overall feature.

3.4 Gammatone Cepstral Coefficients

The gammatone filterbank output replaces the Mel filterbank output used in MFCCs. As with most cepstral coefficients, the natural logarithm of the filterbank output is taken. 24 filters for the gammatone filterbank and 12 cepstral coefficients are taken based on the common setting for Mel filters and cepstral coefficients for MFCCs. This feature is used to compare the effectiveness of gammatone filterbank-based cepstral coefficients against using the filterbank output directly.

3.5 Selective Mel Filterbank

Finally, we adopt the procedure for generating the SGEF to the Mel filterbank used for MFCCs to generate the Selective Mel Filterbank Feature. We follow the MFCC extraction process to generate the Mel-power, skipping the subsequent log compression and DCT steps. This Mel-power vector shares the same dimensions as the gammatone filterbank envelope if we choose 36 filters for the Mel filterbank thus we can apply the channel selection procedure to create a Selective Mel Filterbank with 12 channels.

4. Experimental Setup

4.1 Sound Event Database

Our sound event database is made up of sequences of one to five concatenated sound events. A random amount of silence is inserted between each sound event to simulate short pauses. At the start and the end of each sequence, 0.2 seconds of silence is appended.

The sound events are randomly selected from a pool of 20 sound classes sampled at 16 kHz: 19 sound events from "RWCP Sound Scene Database in Real Acoustic Environment" [25] and one speech class from CENSREC-4 [26]. The 19 sound events are:

1. bank: Beating a handheld moneybox with a metal stick
2. bells: Ringing suspended bells
3. bowl: Beating a handheld bowl with a metal stick
4. buzzer: Sound of an electronic sound toy
5. castanet: Clicking castanets
6. china: Beating of china placed on a sound absorbing board with a wooden stick or a spoon
7. clock: Ringing of an electronic alarm clock
8. dice: Dropping dice on a wooden board
9. dryer: Sound of hair dryers
10. file: Filing a metal stick with a metal file
11. horn: Blowing a bugle
12. maracas: Shaking maracas
13. mechbell: Ringing of mechanical bell of a bicycle
14. phone: Beep of a cellular phone
15. ring: Ringing a bell by shaking
16. string: Twanging of a stringed musical instrument
17. tear: Tearing copy paper
18. whistle: Blowing a whistle
19. wood: Beating a wooden board with a wooden stick

The speech class comprises of the digits one to five spoken by different females in Japanese. Each class consists of 100 clips, 60 of which are allocated to the training pool and the remaining 40 are allocated to the testing pool.

For the training database, a total of 600 sequences are generated from the sound clips in the training pool. 40 dB of additive noise is added to the sequences to simulate a more realistic recording condition. The noise clip used for the training database is not used in the testing database.

For the testing database, a total of 400 sequences are generated from the sound clips in the testing pool. Four types of additive noise are added to the sequences at 40, 20, 15, 10, 5 and 0 dB to create a total of 2400 testing clips. The 40 dB condition is defined as the "clean" testing condition.

The clips for additive noise are taken from the CENSREC-4 and NOISEX-92[†] databases. The "Car", "Hall" and "Office" clips from CENSREC-4 are used directly while the "Speech Babble" clip from NOISEX-92 is downsampled to 16 kHz to give the "Canteen" noise types. The training database utilizes the "Japanese Room" noise clip from CENSREC-4.

4.2 Hidden Markov Model (HMM) Recognizer

The statistical models used are trained and tested using the Hidden Markov Model Toolkit (HTK)[††]. The recognizer setup is based on AURORA-2J [27] 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. We note that this HMM configuration is chosen due to the existence of speech class in the database. Although some impulsive sound classes may require less states than speech, to have a fair comparison of the classification, the same configuration is applied for all the sound classes. We note that it is possible to optimize the number of states for each class separately which might improve the performance and is worth studying for future works.

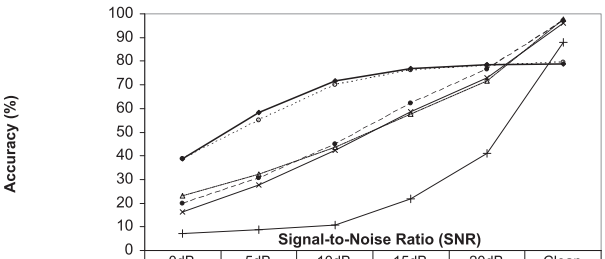
The word network used for the recognizer is allowed to select any combination of the 20 sound events with optional silence at the beginning and end of each event and optional short pauses between events. The detection and classification of these optional silence and short pauses are not considered in our reported results. The HTK Viterbi word recognizer *HVite* is allowed to make any number of insertions and deletions in the reported word sequences. To control this, we vary the log-insertion penalty for each word from 0 to -1000 at -100 intervals and report the highest accuracy from these variations. For the recognition results, we use the "Word Accuracy Rate" given by the HTK recognizer. This value reflects the number of events per sound clip reported correctly and accounts for insertions and deletions.

5. Results and Discussion

The results for our experiments in four noise environments and over six signal-to-noise ratios (SNRs) are presented in Fig. 7. In clean (40 dB) or matching conditions, the features with log-compression (Cepstral Coefficients and Log-Gammatone) show similar good results with accuracies greater than 95%. We are more concerned with robust recognition where noise is present thus the clean condition scores are not relevant. The results for the SGEF and Selective Mel Filterbank Feature are almost identical due to their similar extraction procedures thus we can treat them to be the same for the purpose of comparisons with the other four features.

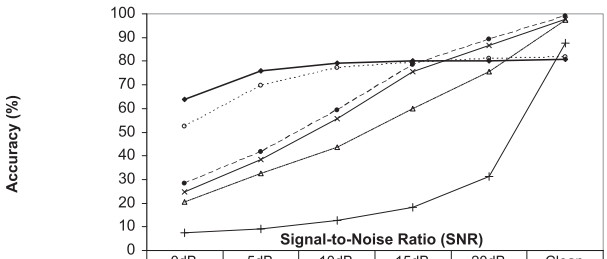
Figure 8 shows the average accuracy results for our experiments. When the clean condition results are removed for the second graph, the accuracies decreased for all the features except the Selective Gammatone Envelope Feature (SGEF). Even with the inclusion of the clean condition results where the conventional Mel-Frequency Cepstral Coefficients (MFCC) is known to be superior, the SGEF is still able to perform better on average.

[†]http://spib.rice.edu/spib/select_noise.html
[††]http://htk.eng.cam.ac.uk/



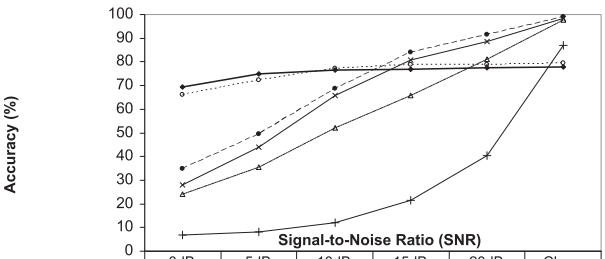
	0dB	5dB	10dB	15dB	20dB	Clean
SGEF	38.85	58.15	71.78	76.72	78.43	78.99
MFCC	19.95	30.66	44.85	62.37	76.56	97.00
Full Gamma	7.22	8.92	10.71	21.74	40.96	87.83
Sel. Log-G	23.11	32.12	43.80	57.50	71.53	97.57
Gamma Cep.	16.14	27.74	42.25	58.64	72.83	96.19
Sel. Mel	38.69	55.15	70.07	76.24	78.18	79.48

(a) Canteen



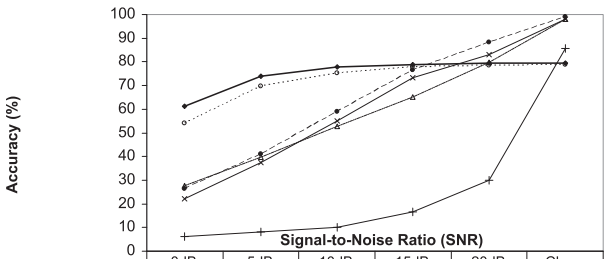
	0dB	5dB	10dB	15dB	20dB	Clean
SGEF	63.91	75.83	79.24	80.05	80.13	80.62
MFCC	28.22	41.61	59.21	78.51	89.38	99.11
Full Gamma	7.46	9.16	12.57	18.09	31.14	87.75
Sel. Log-G	20.68	32.52	43.55	60.02	75.67	97.32
Gamma Cep.	24.74	38.52	55.64	75.51	86.7	97.73
Sel. Mel	52.55	69.67	77.05	79.48	81.02	81.67

(b) Car



	0dB	5dB	10dB	15dB	20dB	Clean
SGEF	69.42	74.94	76.56	76.89	77.45	77.94
MFCC	34.79	49.39	68.86	83.94	91.65	99.03
Full Gamma	6.73	8.19	12.17	21.41	40.55	86.86
Sel. Log-G	24.25	35.44	52.07	65.69	81.02	97.65
Gamma Cep.	28.06	43.88	65.94	80.86	88.65	98.30
Sel. Mel	66.18	72.42	77.29	78.67	78.75	79.48

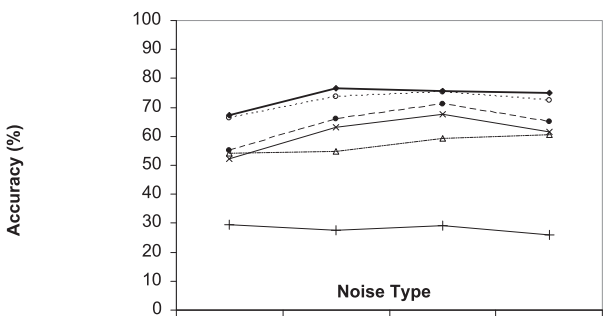
(c) Hall



	0dB	5dB	10dB	15dB	20dB	Clean
SGEF	61.23	73.8	77.94	78.75	79.32	79.48
MFCC	26.36	41.04	58.88	76.64	88.40	99.03
Full Gamma	6.16	8.03	10.14	16.63	29.93	85.81
Sel. Log-G	27.66	39.82	52.72	65.13	79.72	98.05
Gamma Cep.	22.14	37.31	54.91	73.24	83.21	98.05
Sel. Mel	54.18	69.59	75.26	77.94	78.51	78.91

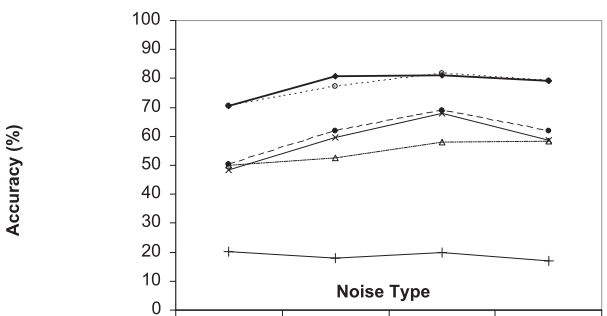
(d) Office

Fig. 7 Recognition accuracy (%).



	Canteen	Car	Hall	Office
SGEF	67.15	76.63	75.53	75.09
MFCC	55.23	66.01	71.28	65.06
Full Gamma	29.56	27.70	29.32	26.12
Sel. Log-G	54.27	54.96	59.35	60.52
Gamma Cep.	52.30	63.14	67.62	61.48
Sel. Mel	66.30	73.57	75.47	72.40

(a) All conditions



	Canteen	Car	Hall	Office
SGEF	70.42	80.66	81.20	79.03
MFCC	50.28	61.78	69.03	61.88
Full Gamma	20.05	17.81	19.82	17.11
Sel. Log-G	49.89	52.43	57.97	58.38
Gamma Cep.	48.55	59.59	67.86	58.75
Sel. Mel	70.62	77.11	81.70	79.08

(b) Noise mismatched conditions

Fig. 8 Average accuracy (%) for all (0–40 dB) and noise mismatched conditions (0–20 dB).

5.1 Selective Gammatone Envelope Feature (SGEF)

Our proposed feature is shown to have the highest overall accuracy in these experiments. Compared to the other four features, this feature shows the least variation over signal-to-noise ratios (SNRs) with an average decrease of about 20%. The notable exception is the "Canteen" noise condition where the decrease is about 50%. This particular condition appears to give the worst results for all the other features thus we can attribute the poor performance to be specific to the noise condition.

The other main characteristic of the SGEF is the poor clean condition accuracy of about 79%. This can be explained by the loss of information due to the channel selection process and the lack of feature compression which exacerbates the slight differences between the training and testing conditions. The need to eliminate noisy channels necessitates the loss of certain channels of the gammatone filterbank thus it no longer spans the entire frequency spectrum. This loss of information will affect the representation of the signal unless the signal contains no information in those missing channels. The second factor regarding the lack of compression is related to the differences between the training and testing data, even if both are in matching conditions. These differences tend to be small and easily overcome by compressing the feature thus the lack of compression makes them more readily apparent, resulting in a decrease in recognition accuracy.

5.2 Mel-Frequency Cepstral Coefficients (MFCC)

MFCCs have the second highest average accuracy among the five features compared and the highest accuracy at clean (40 dB) and nearly clean (20 dB) conditions with accuracies of around 99% and 85% respectively. At 15 dB, both the MFCC and the SGEF share similar results but at lower SNRs, the SGEF shows its superior performance. This relative improvement ranges from around 30% at 10 dB to nearly 100% at 0 dB.

5.3 Full Gammatone Filterbank

The results show clearly that the Full Gammatone Feature offers the worst recognition performance. This is contrasted by the SGEF which has the highest accuracy in noisy (<15 dB) conditions and the highest overall accuracy. The only difference between the two set of features is channel selection. Without removing the noisy channels from the feature, there is a large mismatch between the clean training condition and the noisy test conditions that mean normalization alone cannot compensate. Even in the clean test condition, the accuracy of the Full Gammatone Feature ($\approx$ 86%) is inferior to the log-compressed feature. This justifies the use of some form of dynamic range compression if the main concern is clean or matching condition accuracy.

For a fair comparison in terms of feature dimension, the 12 non-selective filter system is chosen as the reference. We also note that for the non-selective systems, increasing the number of filters to 36 did not significantly improve the performance accuracy while resulting in a much larger feature thus it not recommended.

5.4 Selective Log-Gammatone Filterbank

The Selective Log-Gammatone Feature produces results that are inferior to the two cepstral features by about 20%, especially in the middle SNRs (5–20 dB). The only difference between this feature and the SGEF is the inclusion of log-compression for both the channel selection and the final filterbank output. Comparing the results between the two, it is clear that log-compression improves the clean condition accuracy at the expense of the noisy conditions.

Assuming that the clean signal $X(\omega)$ and noise $N(\omega)$ are uncorrelated ($X(\omega)N(\omega) = 0$), the amplitude of the filterbank output $S(\omega)$ is given by Eq. (5):

$$S(\omega) = X(\omega) + N(\omega) \tag{5}$$

Taking the natural log entangles the signal and noise as shown in Eq. (6):

$$\log S(\omega) = \log [X(\omega) + N(\omega)] \tag{6}$$

For $X >> N$, $\log [X(\omega) + N(\omega)] \approx \log X(\omega)$ thus the noise is largely suppressed at high SNRs. As the noise increases, this effect is lost and taking the log only serves to confuse the signal and the noise since they cannot be separated easily from the non-linear log function.

Taking the log of the filterbank output and performing channel selection appears to have two effects: improving the clean condition accuracy and reducing the noise robustness of the feature. Despite the loss of information due to channel selection, the Selective Log-Gammatone Feature is able to produce clean condition accuracies that are similar to the cepstral features. This suggests that some of the channels used are redundant since there is little to no improvement to using all of the available channels. On the other hand, taking the log reduces the noise robustness that channel selection offers thus for robust recognition, we should not take the log of the filterbank output.

5.5 Gammatone Cepstral Coefficients

The extraction procedure for this feature is more similar to the MFCC than the other three gammatone-based features and the results verify this. The relative improvement of the MFCC over the Gammatone Cepstral Coefficient is around 10% at all SNRs thus the gammatone cepstral coefficients can be considered to be an inferior version of the MFCC. Compared to the Selective Log-Gammatone Feature, the Gammatone Cepstral Cofficients replaces channel selection with the DCT. The result is an overall improvement in accuracy over all SNRs thus we can conclude that channel selection is inferior to using cepstral coefficients for log-compressed filterbank features.

5.6 Selective Mel Filterbank Feature

The performance of the Selective Mel Filterbank Feature is very similar to that of the SGEF. The better accuracy of the SGEF in high noise conditions (<15 dB) can be attributed to the increased robustness of the gammatone filterbank over the Mel filterbank. The difference in accuracy between the Selective Mel Filterbank Feature and the MFCC show that our method of combining channel selection and using the uncompressed feature leads to increased noise robustness at the expense of matching condition performance.

6. Conclusion

We have presented a novel feature, the Selective Gammatone Filterbank Feature (SGEF), and studied the three main differences between it and Mel-Frequency Cepstral Coefficients (MFCCs): the use of the raw filterbank output, channel selection and the choice of not using cepstral coefficients. Our feature is designed to maximize the recognition accuracy in noisy conditions and the experimental results show that indeed, the SGEF is more effective in noisy conditions at signal-to-noise ratios (SNRs) of below 15 dB.

The major drawback of our proposed feature is the poor clean condition (>15 dB) accuracy when compared to dynamic range compressed features. It is impractical to use a feature designed for noisy recognition at low noise levels when conventional features already cater to such requirements. A possible solution is for the recognizer to obtain an estimate of the noise level before deciding which feature to use. Based on our findings, it would be best to use the conventional MFCC if the SNR is high, only switching to the SGEF in noisy conditions (<20 dB). The other disadvantage of the SGEF, the need for channel selection with different noise environments and sound events, can be eliminated if similar noise environments and sound events are used. By reducing the differences between the new settings with an existing one, the t-test distance will not vary significantly thus the channels selected will be the same.

It must be noted that there are many possible ways to employ the gammatone filterbank in a feature vector and that our results are not representative of all such methods. The simple implementation presented can serve as a baseline for further studies to improve upon gammatone-based features in the future. The methods we employed to develop the SGEF can also be used for similar time-frequency representations such as the Gabor filterbank. This should allow similar noise robust features to be implemented as we have shown with the Selective Mel Filterbank Feature.

References

[1] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini, "Automatic sound detection and recognition for noisy environment," European Signal Processing Conference (EUSIPCO), pp.1033–1036, 2000.

[2] W. Huang, T.K. Chiew, H. Li, T.S. Kok, and J. Biswas, "Scream detection for home applications," 5th IEEE Conference on Industrial Electronics and Applications (ICIEA), pp.2115–2120, 2010.

[3] S. Ntalampiras, I. Potamitis, and N. Fakotakis, "On acoustic surveillance of hazardous situations," Proc. ICASSP, pp.165–168, 2009.

[4] S.B. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process., vol.28, no.4, pp.357–366, 1980.

[5] J. Tchorz and B. Kollmeier, "A model of auditory perception as front end for automatic speech recognition," J. Acoustical Society of America, vol.106, no.4, pp.2040–2050, 1999.

[6] M.E. Munich and L. Qiguang, "Auditory image model features for automatic speech recognition," INTERSPEECH-2005, pp.3037–3040, 2005.

[7] 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.

[8] Y. Shao, Z. Jin, D. Wang, and S. Srinivasan, "An auditory-based feature for robust speech recognition," Proc. ICASSP, vol.00, pp.4625–4628, 2009.

[9] J.C. Wang, H.P. Lee, J.F. Wang, and C.B. Lin, "Robust environmental sound recognition for home automation," IEEE Trans. Autom. Sci. Eng., vol.5, no.1, pp.25–31, Jan. 2008.

[10] Y. Gong, "Speech recognition in noisy environments: A survey," Speech Commun., vol.16, no.3, pp.261–291, 1995.

[11] R. Martin, "Spectral subtraction based on minimum statistics," Proc. EUSIPCO, pp.1182–1185, 1994.

[12] M.D. Skowronski and J.G. Harris, "Increased mfcc filter bandwidth for noise-robust phoneme recognition," Proc. ICASSP, vol.1, pp.801–804, 2002.

[13] C.P. Chen and J.A. Bilmes, "Mva processing of speech features," IEEE Trans. Audio Speech Lang. Process., vol.15, no.1, pp.257–270, 2007.

[14] R.D. Patterson and J. Holdsworth, "A functional model of neural activity patterns and auditory images," Advances in Speech, Hearing and Language Processing, vol.3, pp.547–563, 1996.

[15] Y.R. Leng, T.H. Dat, N. Kitaoka, and H. Li, "Selective gammatone filterbank feature for robust sound event recognition," INTERSPEECH 2010, pp.2246–2249, 2010.

[16] H. Hermansky, S. Tibrewala, and M. Pavel, "Towards asr on partially corrupted speech," Proc. ICSLP, vol.1, pp.462–465, 1996.

[17] H. Bourlard and S. Dupont, "A new asr approach based on independent processing and recombination of partial frequency bands," Proc. ICSLP, vol.1, pp.426–429, 1996.

[18] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," Speech Commun., vol.34, no.3, pp.267–285, 2001.

[19] B. Raj and R.M. Stern, "Missing-feature approaches in speech recognition," IEEE Signal Process. Mag., vol.22, no.5, pp.101–116, 2005.

[20] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," Proc. IEEE, vol.77, no.2, pp.257–286, 1989.

[21] J. Bilmes, "What hmms can do," IEICE Trans. Inf. & Syst., vol.E89-D, no.3, pp.869–891, March 2006.

[22] M. Slaney, "An efficient implementation of the patterson-holdsworth auditory filter bank," Apple Computer Technical Report #35, 1993.

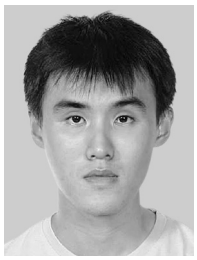
[23] J. Wilpon, C.H. Lee, and L. Rabiner, "Improvements in connected digit recognition using higher order spectral and energy features," Proc. ICASSP, vol.1, pp.349–352, 1991.

[24] C.J. Jr., H.D.H. Vi, and R.P. Lippmann, "A comparison of signal processing front ends for automatic word recognition," IEEE Trans. Speech Audio Process., vol.3, no.4, pp.286–293, 1995.

[25] S. Nakamura, K. Hiyane, F. Asano, and T. Endo, "Sound scene data collection in real acoustical environments," J. Acoust. Soc. Jpn, vol.20, no.3, pp.225–231, 1999.

[26] 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.

[27] S. Nakamura, K. Takeda, K. Yamamoto, T. Yamada, S. Kuroiwa, N. Kitaoka, T. Nishiura, A. Sasou, M. Mizumachi, C. Miyajima, M. Fujimoto, and T. Endo, "Aurora-2j, an evaluation framework for japanese noisy speech recognition," IEICE Trans. Inf. & Syst., vol.E88-D, no.3, pp.535–544, March 2005.



Yi Ren Leng received his B.Sc. in Physics from the National University of Singapore (NUS) in 2009. He has been a Research Enginner with the Institute for Infocomm Research (I2R), Singapore, since 2009. His research interests includes speech and sound recognition.

Huy Dat Tran received his M.ESc. from the National Technical University of Ukraine (NTUU) in 1995 and Ph.D. from the National Academy of Sciences (NAS) of Ukraine in 2000, with both degrees specialized in acoustics. From 2000 to 2002 he did his postdoc research at the Institute of Hydromechanics, NAS of Ukraine. From 2002 to 2005 he had been a postdoc fellow at F. Itakura and K. Takeda Labs, Nagoya University, Japan. Since 2005 he has been a Senior Research Fellow at the Institute for Infocomm Research, Singapore. His research interest includes acoustic and speech signal processing, sound recognition, and statistical methods for pattern recognition. He serves as an Editorial member of the Open Acoustic Journal and a regular reviewer for many journals including IEEE Transaction on Audio, Speech and Language Processing, Speech Communication, Signal Processing Letters, and Pattern Recognition Letters.

Norihide Kitaoka received the B.E. and M.E. degrees from Kyoto University, Kyoto, Japan, in 1992 and 1994, respectively, and the Dr. Eng. degree from Toyohashi University of Technology, Toyohashi, Japan, in 2000. He joined DENSO Corporation, Kariya, Japan, in 1994. He joined the Department of Information and Computer Sciences, Toyohashi University of Technology, as a Research Associate in 2001 and was a Lecturer from 2003 to 2006. Since 2006, he has been an Associate Professor with the Department of Media Science, Graduate School of Information Science, Nagoya University, Nagoya, Japan. His research interests include speech processing, speech recognition, and spoken dialog. He is a member of International Speech Communication Association (ISCA), Acoustical Society of Japan (ASJ), Information Processing Society of Japan (IPSJ), and Japan Society for Artificial Intelligence (JSAI).

Haizhou Li (M'91 - SM'01) received the B.Sc., M.Sc., and Ph.D. degrees in electrical & electronic engineering from the South China University of Technology (SCUT), Guangzhou, in 1984, 1987, and 1990, respectively. Dr. Li was a Research Assistant from 1988 to 1990 at the University of Hong Kong, Hong Kong, China. In 1990, he joined SCUT as an Associate Professor where he became a Full Professor in 1994. From 1994 to 1995, he was a Visiting Professor at CRIN, Nancy, France. In 1995, he became the Manager of the ASR group at the Apple-ISS Research Centre in Singapore where he led the research of Apple's Chinese Dictation Kit for Macintosh. In 1999, he was appointed Research Director of Lernout & Hauspie Asia Pacific, where he oversaw the creation of the first multimodal speech, pen and keyboard input solution for Chinese computing. From 2001 to 2003, he was the Vice President of InfoTalk Corp. Ltd. Since 2003, he has been with the Institute for Infocomm Research (I2R), Singapore, where he is now the Principal Scientist and Head of Human Language Technology Department. Dr. Li was a Visiting Professor at the University of New South Wales, Australia in 2008 and the University of Eastern Finland in 2009 and 2010. His current research interests include automatic speech recognition, speaker and language recognition and natural language processing. Dr. Li was named one of the two Nokia Visiting Professors 2009 by the Nokia Foundation to recognize his contributions to Speaker and Language Recognition research. He was a recipient of the National Infocomm Award 2001 and the TEC Innovator's Award 2004 in Singapore. He is now an Associate Editor for the Springer International Journal of Social Robotics, and the IEEE Transactions on Audio, Speech and Language Processing. He is also an elected Board Member of International Speech Communication Association (ISCA) and the Asian Federation of Natural Language Processing (AFNLP).