Comparison of Auditory-Inspired Models Using Machine-Learning for Noise Classification

Salinna Abdullah [1], Andreas Demosthenous [1], and Ifat Yasin [2]

[1] *Department of Electronic and Electrical Engineering,*
[2] *Department of Computer Science*,
University College London, London, United Kingdom.
Email: salinna.abdullah.13@ucl.ac.uk; a.demosthenous@ucl.ac.uk; i.yasin@ucl.ac.uk

***Abstract* - Two auditory-inspired feature-extraction models, the Multi-Resolution CochleaGram (MRCG) and the Auditory Image Model (AIM) are compared on their acoustic noise classification performance, when combined with two supervised machine-learning algorithms, the ensemble bagged of decision trees or Support Vector Machine (SVM). Noise classification accuracies are then assessed in nine different sound environments with or without added speech and at different SNR ratios. The results demonstrate that classification scores using feature extraction with the MRCG model are significantly higher than when using the AIM model ($p< 0.05$), irrespective of machine-learning classifier. Using the SVM as a classifier also resulted in significantly better ($p<0.05$) classification performance over bagged trees, irrespective of feature-extraction model. Overall, the MRCG model combined with SVM provides a more accurate classification for most of the sound stimuli tested. From the comparison study, suggestions on how auditory model-plus-machine-learning can be improved for the purpose of sound classification are offered.**

Keywords - Acoustic noise classification, auditory model, cochlear implant, machine-learning.

** This work was supported by an EPSRC Industrial Strategy Studentship to S. Abdullah.*

I. INTRODUCTION

A lot of challenges persist in current research that impede Cochlear Implant (CI) listeners from achieving the same level speech intelligibility as normal hearing listeners in noisy conditions. One challenge is developing noise suppression algorithms for pre-processing noisy speech in CIs that will not degrade in challenging listening situations, such as in nonstationary noise and reverberant environments [1]. This is usually hard to achieve because many noise suppression algorithms rely on accurate noise estimation, but these challenging listening conditions make it hard to track and predict accurately their statistical properties [2]. Another reason is because the spectro-temporal characteristics of the various types of real-life background noise can vary widely, rendering a single speech processing algorithm often applied to all listening conditions too ambitious. A way to improve the effectiveness of CIs in a range of acoustic environments is for an audiologist to program the CI for the user to incorporate multiple maps (map: a set of parameters unique to the user, controlling speech processing via the CI). This allows a CI user to switch between different speech processing programs, each of which has been optimised for different listening environments [3].

However, the process of changing between maps may be tedious for many CI users and they may feel more comfortable with using solely the default setting without exploring the benefits of using other maps in different listening environments. Therefore, an automated identification of the listening environment and a real-time adaptation of speech processing by incorporating machine-learning approaches into the CI design is desirable. The front-end noise classification by the CI would enable automatic selection of an optimised set of parameters to process the incoming audio for a particular listening environment and this would improve speech intelligibility for the CI user. Furthermore, there is ongoing research on how to effectively restore and increase spectrotemporal cues to CI users since findings suggest that CI users' greater susceptibility to noise are caused by many factors such as reduced spectral resolution, high degree of spectral smearing that is associated with cochlear electrode channel interaction [4] and lack of spectrotemporal adaptation [5], a vital ability possessed by the human auditory system beneficial for discriminating sounds. Processing audio signals in a way that more closely represents how human ears perceive sound has led to an improved listening experience for CI users. For example, [6] reported that utilising a gammatone filterbank, which was designed to model the filtering process employed by the human auditory system, as a front-end was found to result in a significant improvement in melodic contour identification for both normal hearing and CI listeners. From this, it is hypothesised that feature extraction models inspired by the human auditory system, combined with machine-learning which can carefully consider non-obvious but important spectrotemporal patterns of different real-world listening environments, would lead to a robust noise event classification system beneficial for improving speech intelligibility for CI users.

In this study, two auditory-inspired feature-extraction models, the Multi-Resolution CochlearGram (MRCG) model [7] and the Auditory Image Model (AIM) [8], are combined with machine-learning approaches, an ensembled bagged (from the present continuous tense 'bagging', which stands for Bootstrap Aggregating) trees and a Support Vector Machine (SVM), and compared on their acoustic noise classification efficacy when exposed to various test stimuli and training methodologies. Fig. 1 depicts a high-level block diagram representation of the noise classification process.

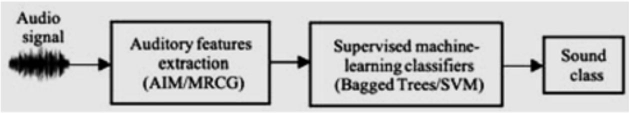

Fig. 1. Basic block diagram representation of the auditory-inspired noise classification process in this study.

The output measure of correct classification, measured by the classification accuracy in percentages, of the acoustic environment is used as the evaluation metric and the significance of the classification results are further assessed using the ANalysis Of VAriance (ANOVA) [9].

The rest of the paper is organised as follows. In the next section, operations of the auditory models, AIM and MRCG, and the machine-learning classifiers employed are explained. Section III and IV describe the datasets used and the experimental setup for the different experiments executed, respectively. Section V discusses the observations and results from the comparison. Finally, the concluding remarks are given.

II. METHODOLOGIES

In this section, the operations of the auditory models, AIM and MRCG, and the machine-learning classifiers, ensemble bagged trees and SVM, are explained, and example visualizations of the mentioned auditory models are provided.

A. Auditory Image Model (AIM)

The AIM feature extraction used in this study is the one described by [8]. However, the model was first described by [10]. Since the model has been comprehensively described in these papers, only the vital components of the model will be mentioned here.

The AIM is a time-domain functional model of sound representations in the hearing system which can be associated with the cascade of processing stages in the auditory pathway. The principle functions of the AIM are to describe and simulate: (1) peripheral auditory processing, such as pre-cochlear processing, Basilar Membrane Motion (BMM) and the transduction process in the cochlea, (2) central auditory processing, such as neural activity patterns in the auditory nerve and cochlear nucleus and (3) higher auditory processing, such as strobed temporal integration and source size normalisation which will eventually yield the Size-Shape transformed auditory Image (SSI). The AIM feature set used for the noise classification only consists of the outputs from the BMM and SSI stages. The BMM features are obtained by calculating the logarithmic envelope power of a linear gammatone filter output. The gammatone function is defined in the time domain by its impulse response as shown in Equation 1 [11]:

$$g(t) = at^{n-1}\cos(2\pi ft + \phi)e^{-2\pi bt} \tag{1}$$

where n is the filter order, b is the filter bandwidth in Hz, t represents time in seconds, f is the filter centre frequency, a is the amplitude of the signal, ϕ is the phase of the signal. A gammatone filter bank with 64 frequency channels is used to match the number of channels used in the construction of the MRCG (described later). The SSI is a vocal tract length covariant representation of the input signal and it is obtained from the Stabilized Auditory Image (SAI) output at the higher auditory processing of the AIM, through processes described in [12]. The SAI is an interpretable stabilized representation of sound that preserves temporal fine structure through the strobed temporal integration process that converts the time dimension of the neural activity pattern into the time-interval dimension of the SAI [10]. After the processes described in [12], discrete cosine transform is applied to the 12 columns of the output with the greatest root mean square and then only the 2nd to 22nd coefficients are retained for the SSI features to be used for the classification. Figure 2 depicts the processes required for extracting the SSI features from the SAI. For a 20 ms long time frame (with 10 ms overlap), the resulting dimensionality was 316 for each feature vector, i.e., 64 BMM features and 252 SSI features.

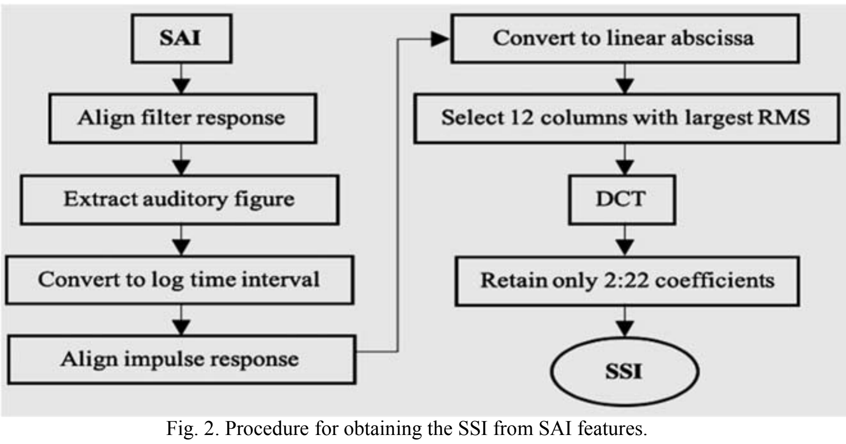

Fig. 2. Procedure for obtaining the SSI from SAI features.

Figure 3 provides example visualizations of the BMM and SAI features (of which the SSI features are obtained from) extracted from a 20 ms frame of a clean IEEE sentence – "the birch canoe slid on the smooth planks" – spoken by a male talker, and from a 20 ms frame of the same sentence contaminated with babble noise at 0 dB SNR.

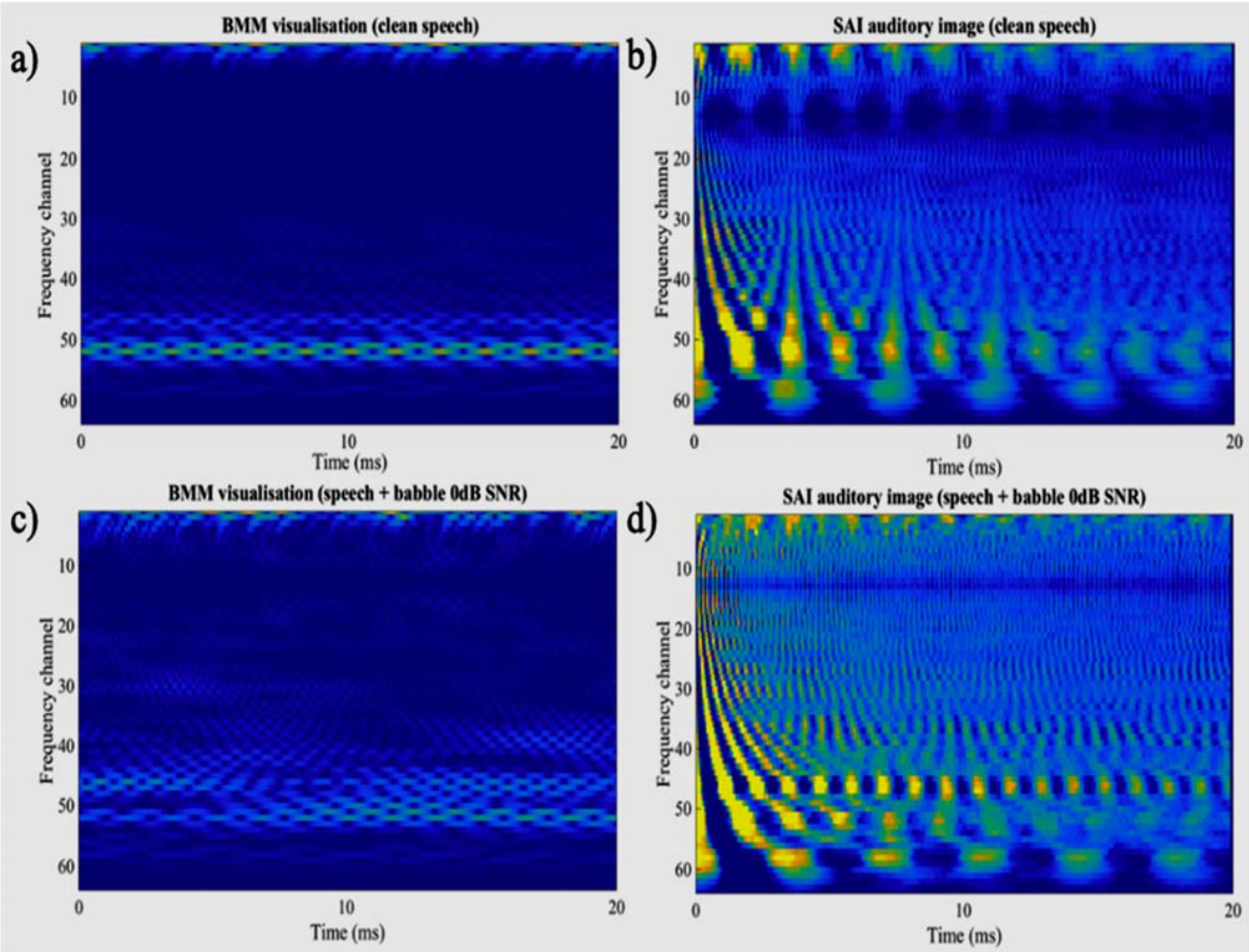

Fig. 3. (a) BMM of the clean speech. (b) SAI of the clean speech. (c) BMM of the speech contaminated with babble noise at 0 dB SNR. (d) SAI of the speech contaminated with babble noise at 0 dB SNR.

B. Multi-Resolution CochleaGram (MRCG)

The MRCG proposed by [7] is a combination of four cochleagrams (CG1, CG2, CG3 and CG4) that encode power distributions of an audio signal in the time-frequency representation at different resolutions. The high-resolution cochleagram captures the local information while the three low-resolution cochleagrams capture the spectrotemporal contexts at different scales. Fig. 4 is provided to summarise the construction process. In addition to the MRCG, [7] suggested adding delta (Δ) and double-delta (ΔΔ), the first and second-order derivatives of the MRCG feature vector respectively, features to yield the MRCG+Δ+ΔΔ feature set.

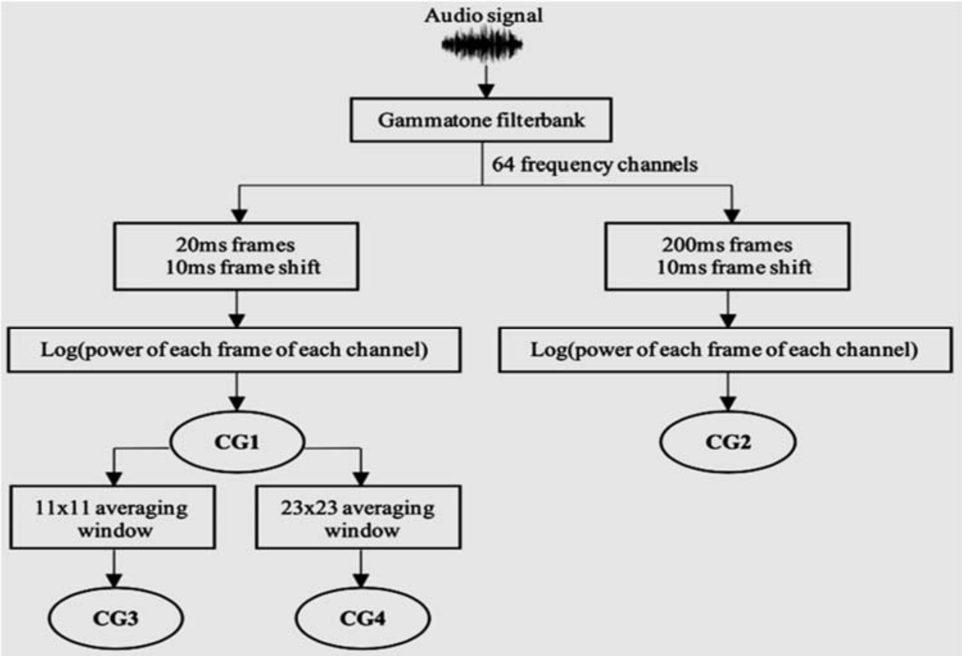

Fig. 4. Procedure for obtaining the MRCG features. 20 ms frames are used to obtain CG1, 200 ms frames are used to obtain CG2 and averaging windows are applied on CG1 to obtain CG3 and CG4.

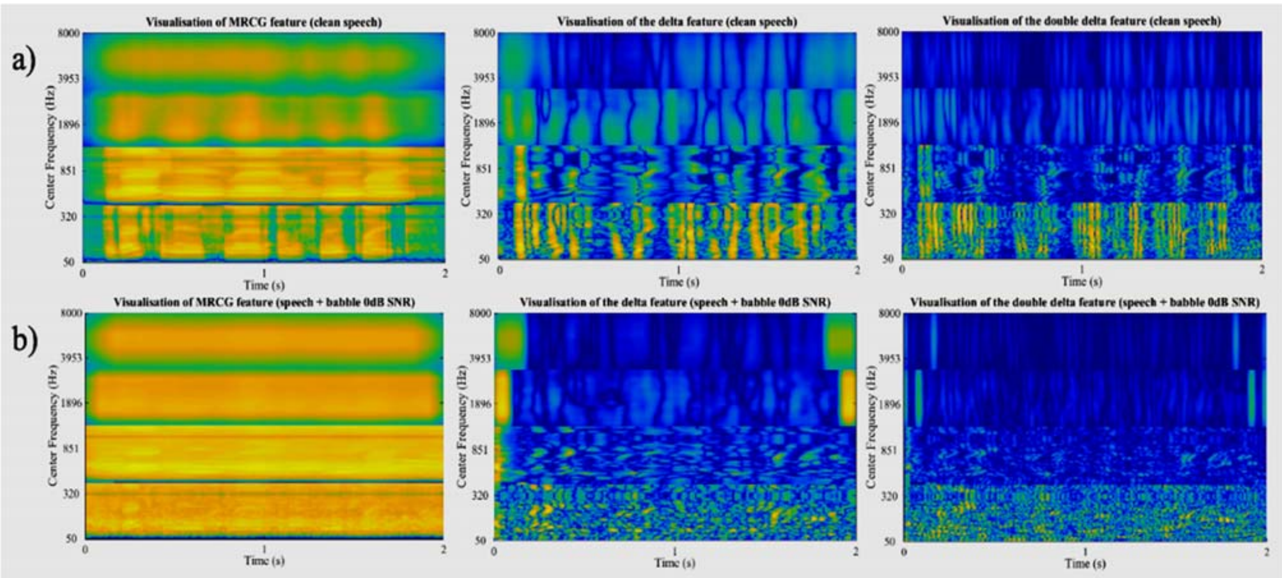

Fig. 5. (a) From left to right: MRCG, Δ and ΔΔ of the clean speech. (b) From left to right: MRCG, Δ and ΔΔ of the speech contaminated with babble noise at 0 dB SNR.

Adding Δ and ΔΔ is a popular post-processing technique in speech processing, widely used to capture temporal dynamics. In [13] the addition of Δ and ΔΔ features was found to improve speech separation results. The MRCG+Δ+ΔΔ feature set resulted in a dimensionality of 768 for each 20 ms frame (64×4 for MRCG + 256×Δ + 256×ΔΔ). Hereinafter the term 'MRCG' includes the Δ and ΔΔ features. Example visualisations of MRCG features extracted from a clean and contaminated speech are shown in Fig. 5. The same clean IEEE sentence and IEEE sentence contaminated with babble noise at 0 dB SNR mentioned in the previous section were used for attaining the examples. The figure shows that the energy of a clean speech signal tends to have a sparse distribution in time and frequency. In other words, a clean speech signal usually posesses significant energy in small, isolated regions of a time-frequency representation. On the contrary, a signal is likely to be dominated by noise when relatively high energy is

DOI 10.5013/IJSSST.a.21.02.20

ISSN: 1473-804x online, 1473-8031 print

observed across a large spread of time-frequency units. Sparser energy distribution in time and frequency leads to larger energy differences between many adjacent time-frequency units as observed in the Δ and ΔΔ features shown in same figure.

C. Machine-Learning Classifiers

The extracted MRCG and AIM features are used to train two supervised machine-learning classifiers: an ensemble bagged of decision trees (hereinafter referred to as bagged trees) and a Support Vector Machine (SVM). An SVM is a discriminative classifier, formally defined by a separating hyperplane. Initially, SVMs with several variations of kernel functions – linear, quadratic and cubic – were used for testing. However, the cubic SVM resulted in deterioration in the validation accuracy score (a prediction of the performance of the trained model on test sets), potentially due to its higher sensitivity to outliers, and the quadratic SVM gave the same results as the linear SVM. Therefore, only the results obtained from the linear SVM are reported here. The linear kernel is given by the inner product of the feature space <x,y> plus an optional constant c as shown in Equation 2:

$$k(x, y) = x^T \mathrm{y} + \mathrm{c} \qquad (2)$$

where T represents the transpose operation.

The bagged trees method combines several decision trees to produce better predictive performance than when a single decision tree is utilised. The decision to use a bagged trees classifier stemmed from [14]'s finding that using a random forest tree classifier in a hierarchical sound classification algorithm leads to higher robustness to non-trained sound signals.

All algorithms in this study were implemented in MATLAB R2019b. The classifiers, bagged trees and linear SVM, were constructed and the trained prediction models were obtained using the Classification Learner application from MATLAB (provided in the Statistics and Machine Learning Toolbox). 5-folds cross-validation (the default validation scheme) was chosen for all training to prevent overfitting during the training process. The bagged trees classifier was implemented using the Breiman's random forest algorithm [15]. The number of learners (or number of trees) and the maximum number of splits (or branch points) were kept at the Classification Learner app's default setting of 30 and 28,727 respectively.

III. DATASETS

The dataset used in the study is a noisy speech corpus, NOIZEUS, obtained from [16]. This corpus was developed with the intention of providing a common noisy speech (speech tokens combined with noise) database for the evaluation of speech-enhancement algorithms. The database consists of 30 IEEE sentences (produced by three male and three female speakers) corrupted by eight different real-world noises (airport noise, babble noise, car noise, exhibition noise, restaurant noise, train-station noise, street noise and suburban train noise) at different SNRs (0 dB, 5 dB, 10 dB and 15 dB SNRs). The noise files used for the noisy speech mixture were obtained from the AURORA database and different portions of a single noise file were used for different mixtures to ensure that the noisy segments used in testing and training were different. Clean speech files are also provided in NOIZEUS. The NOIZEUS files are sampled at 8kHz and each file is 2s long.

IV. EXPERIMENTAL SETUP

The experimental setups, such as how the training and testing sets are generated, for the different experiments directed are described here.

A. Experiment 1: AIM and MRCG's classification performance

12 IEEE sentences from NOIZEUS (recorded by male speakers only) from each category (i.e., clean speech and speech corrupted by the eight different real-world noise types – airport, babble, car, exhibition, restaurant, station, street and train noise) are used for training. The corrupted sentences are all presented at 0 dB SNR. 3 different IEEE sentences (obtained from the same database, also recorded by male speakers and corrupted by eight different real-world noise types at 0dB SNR) are used for the testing.

B. Experiment 2: AIM and MRCG's ability to distinguish between SNRs

12 recordings of babble noise at different SNRs (clean speech, 0 dB, 5 dB, 10 dB and 15 dB SNR) from NOIZEUS are used for training and 3 other different recordings of babble noise at the same SNRs (clean speech, 0 dB, 5 dB, 10 dB and 15 dB SNR) are used for testing. Each recording used for training/testing consists of a different IEEE sentence.

C. Supplementary Experiment: AIM's classification performance with a longer frame length

The methodology is exactly as described for Experiment 2. However, the frame length is increased to 200ms (the overlap is kept at 10ms) and the supplementary experiment is conducted only with AIM as the MRCG model has already been designed to extract cochleagrams at different frame lengths (20 ms and 200 ms).

V. RESULTS AND DISCUSSION

Fig. 6 presents the outcome from Experiment 1.

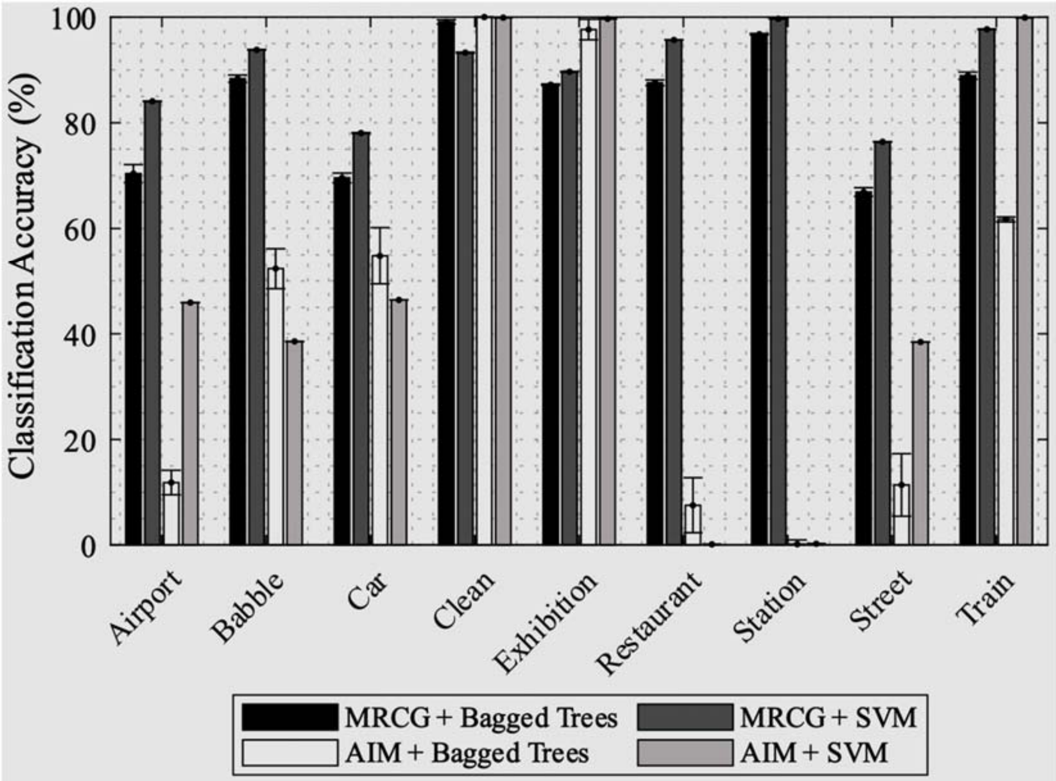

Fig. 6. AIM's and MRCG's classification performance, when combined with bagged trees and SVM, for Experiment 1.

The error bars presented in some of the result figures indicate the standard error of mean calculated from conducting five repetitions of the experiment. The MRCG feature set resulted in a high classification accuracy with majority of its classification accuracies at above 80%. The average classification accuracy is best at 89.8% when MRCG in combination with linear SVM is used. The AIM feature set performed worse than the MRCG feature set for identifying most noise types, except for the exhibition (with both bagged trees and SVM) and train noise (with SVM only). However, it was able to identify clean speech better than the MRCG method. The ANOVA results are reported as F-statistic and its associated p-value, and eta squared ($\eta 2$). The F-statistic and p-value are used in combination to determine the significance of the data whereas $\eta 2$ represents the proportion of variance. For the ANOVA, the three factors considered are the auditory model (two levels; MRCG or AIM), machine-learning approach (two levels: bagged trees and SVM) and sound stimulus (nine levels; airport, babble, car, clean speech, exhibition, restaurant, station, street and train). There is a significant effect of auditory model [$F(1,4) = 12001.2$, $p < 0.001$ (two-tailed) with effect size, $\eta 2 = 1.00$)] and machine-learning approach [$F(1,4) = 330.4$, $p < 0.001$ (two-tailed) with effect size, $\eta 2 = 0.98$]. There are also significant 2-way interactions between the auditory model and machine-learning approach [$F(1,4) = 8.14$, $p < 0.05$ (two-tailed) with effect size, $\eta 2 = 0.67$], auditory model and sound stimulus [$F(8,32) = 306.4$, $p < 0.001$ (two-tailed) with effect size, $\eta 2 = 0.98$)], machine-learning approach and sound stimulus [$F(8,32) = 34.1$, $p < 0.001$ (two-tailed) with effect size, $\eta 2 = 0.89$)], and a 3-way interaction between the auditory model, machine-learning approach and sound stimulus [$F(8,32) = 20.76$, $p < 0.001$ (two-tailed) with effect size, $\eta 2 = 0.84$)].

Post hoc pairwise comparisons were conducted with a Bonferroni correction to keep Type I error at 5%. Post hoc pairwise comparisons showed that when using the MRCG model, classification scores using SVM are significantly higher for sound samples of airport, babble, car, exhibition, restaurant, station, street and train than when using bagged trees ($p<0.05$). Only for sound stimulus clean speech, with the MRCG model, is the score significantly higher when using bagged trees than SVM ($p<0.05$). Post hoc pairwise comparisons also showed that when using AIM, classification scores using SVM are significantly higher for sound samples of airport, street and train than when using bagged trees ($p<0.05$). Only for sound stimulus clean speech, with the AIM feature set is the score significantly higher when using bagged trees than SVM ($p<0.05$). For the AIM model, there is no significant difference in classification accuracy when comparing between machine-learning type (SVM and bagged trees) for sound stimuli babble, clean speech, exhibition, restaurant and station.

The results for Experiment 2 shown in Fig. 7. It shows that AIM has a better distinguishing power between various SNRs compared to the MRCG model.

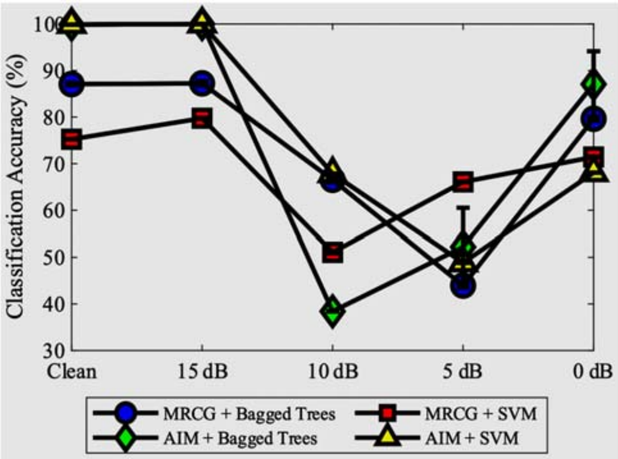

Fig. 7. AIM's and MRCG's classification performance, when combined with bagged trees and SVM, when tested with the same noise environment at different SNRs (Experiment 2).

MRCG's better performance in Experiment 1 indicates that temporal cues at different resolutions are beneficial for distinguishing between different noise types. However, more sophisticated auditory model such as AIM, which includes more complex processing stages (e.g., converting BMM to neural activity patterns to model central auditory processing and identification of neural peak times to preserve fine-structure of noise etc.) before finally yielding the SSI features, might lead to the ability to better distinguish between different SNR values. There were many cases where both models confused adjacent SNRs with the correct one (e.g., 10 dB SNR is confused for 15 dB or 5 dB SNR). This is especially true for identifying babble noises at 5 dB and 10 dB SNR, where the classification accuracies are observed to be poorest. This outcome suggests that better classification accuracy is achieved when the signal is either clean or dominated by noise (two extreme ends). In this test condition, the bagged trees performed better than the linear SVM for the MRCG model but linear SVM is preferred when used in combination with AIM.

The reason for AIM's poorer classification performance in Experiment 1 may be due to its extraction of solely fine temporal context. Unlike the MRCG model which extracts features from the audio signals using two different timescales, the AIM only extracts features using frames of 20 ms. This time scale may be too small to the extent that useful temporal information that describes a certain noise class may not be visible, and that many frames from the training and testing data are significantly different from each other, thus preventing the classifiers when combined with AIM, from generalising from the training data. [17] emphasised that natural sounds, music and vocal sounds have rich temporal structure over multiple timescales. Therefore, it is imperative that an audio classification system accounts for both: fine-grained information in short timescales (<50 ms) and global pattern in long timescales (~200ms).

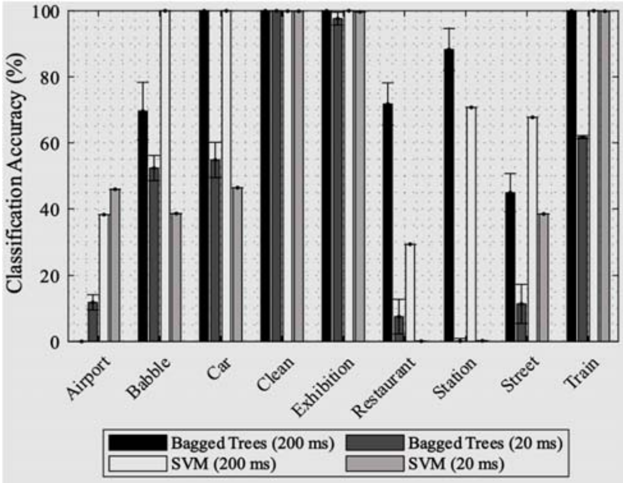

Fig. 8. AIM's classification performance, when combined with bagged trees and SVM, when a longer frame length of 200 ms is employed (Experiment 3).

Fig. 8 depicts the results attained from a supplementary experiment conducted to determine whether the performance of the AIM model could be improved when a longer frame length of 200 ms is used for feature extraction. The test stimuli and methodology are exactly the same as that in Experiment 1 with exceptions that only the AIM model is tested, and the frame length is increased to 200 ms with 10 ms overlap. Indeed, remarkable improvement is observed for almost all noise types except for airport noise when both bagged trees and SVM were used. When the airport noise is tested, a considerable number of frames are identified as restaurant noise. To investigate a possible reason for this, a sample BMM output of the tested airport noise is compared against sample BMM outputs of the trained airport and restaurant noise.

The result, depicted in Fig. 9, suggests that the machine-learning model achieved low classification accuracy for the tested airport noise as the BMM output of the tested airport noise more closely resembles the BMM output of the trained restaurant noise.

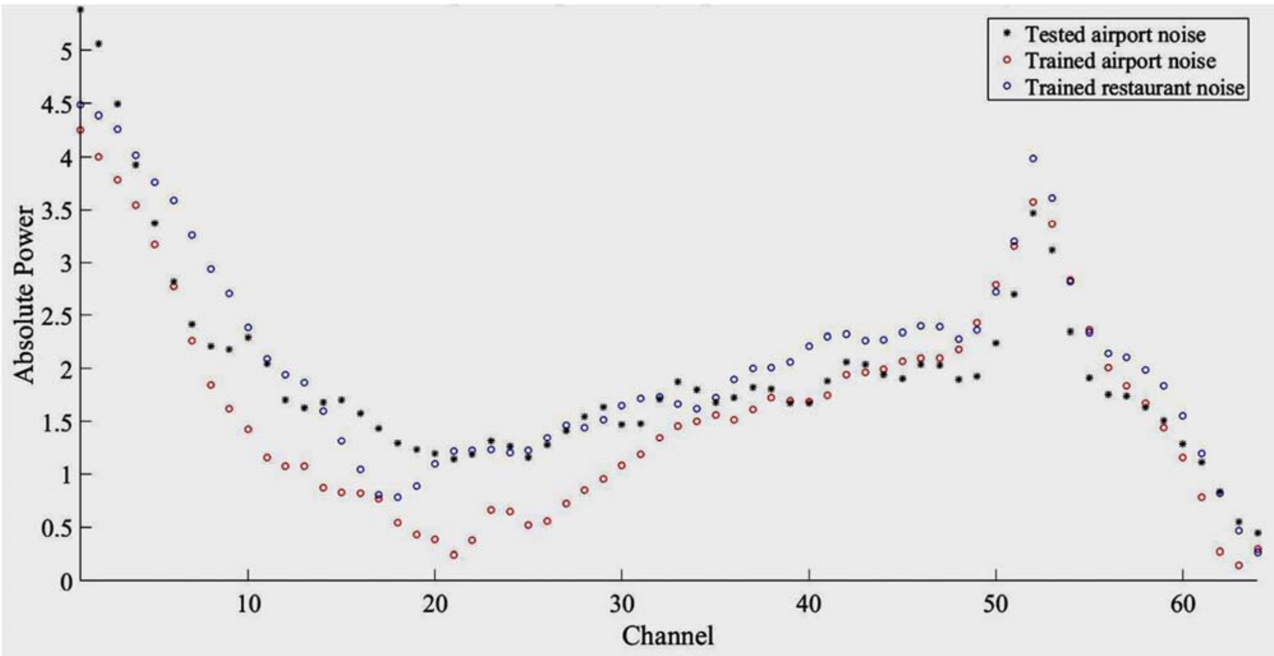

Fig. 9. The BMM output of the tested airport noise compared with the BMM output of the trained airport noise and trained restaurant noise.

It appears that confusions are inevitable as many noise types share similar spectrotemporal contexts that are not easily distinguishable. Therefore, to implement a sound classification system in CIs, similar noise types should be classed together (i.e., training class labels should be more carefully deliberated) to minimise classification errors, and they will share the same noise reduction parameters during the speech processing stage.

Although the bagging method used in the bagged trees classifier is widely used to reduce the variance of a decision tree, a greater variance is still observed for the bagged trees classifier than the linear SVM. This could be seen from the error bars included in Fig. 6-8 which show larger variability in the results for MRCG+bagged trees and AIM+bagged trees. Utilising a larger set of training and testing samples should further reduce this variance. Moreover, the optimal combination of number of learners and maximum number of splits could be explored to further reduce variance, prevent overfitting and produce high classification accuracy but in the expense of training and testing time, and memory usage. Post hoc pairwise comparisons from the ANOVA analyses conducted for Experiment 1 showed that the classification accuracy is significantly higher only for the clean speech stimulus when the bagged trees classifier is used ($p<0.05$). This possibly implies that the combination of number of learners and maximum number of splits was already sufficiently optimal for the clean speech stimulus which has more distinct spectrotemporal features than the rest of the stimuli.

The dimensionality is larger for the MRCG+Δ+$\Delta\Delta$ (total dimensionality of 768 for each 20 ms frame) but AIM (total dimensionality of 316 for each 20 ms frame) took a longer time to extract the required features due to the various processing stages it must go through before finally yielding the SSI features. With a desktop platform equipped with a 3.6GHz clock processor, it took an average of 8.62 ms and 23.8 ms to extract the MRCG and AIM features, respectively, for a 20 ms frame of a clean IEEE sentence sampled at 8 kHz. The processing time of the MRCG features was below the frame overlap time of 10 ms (50% of 20 ms). The time required to extract the AIM features, on the other hand, exceeds the overlap time. This renders AIM unsuitable for real-time operation as this factor might lead to a snowball effect of creating a larger latency. Investigating the significance of the SSI features for accurate sound classification is vital as the same classification accuracy may be achieved with less complex features obtained with less computations.

From this study, auditory-inspired feature extraction models are observed to provide promising results for the task of sound classification even when trained with a small dataset. However, a good compromise must be achieved between classification accuracy and computation complexity in order for them to be suitable to be implemented in hearing devices such as cochlear implants, where low power consumption and compacity are sought after. Although the models have displayed a fair capability in discriminating between different SNR values, the classification system should be made insensitive to changes in SNR when it comes to classifying noise types. Further testing should be carried out to investigate the extent of the impact varying SNR values have on the sound classification performance of the models. More could also be done to optimize the classifier employed for sound classification applications when used in conjunction with auditory-inspired feature extraction.

VI. CONCLUSION

Overall, the MRCG model gave more consistent results across experiments but AIM can better distinguish between different SNRs when the same noise type is used for testing. The linear SVM gave more favourable classifications in most of the tests conducted but it is clear that the classifiers' performance varies when combined with different features and different test materials. There is a significant interaction between the feature extraction model and machine-learning approach, such that when using the MRCG model classification, scores using SVM are significantly higher for the majority of sound samples (i.e., all except for clean speech) than when using bagged trees ($p<0.05$). Overall, the best classification scores are obtained with MRCG when combined with SVM.

Limitations and challenges found in the comparison study have been discussed. Using variable frame sizes can account for temporal structure over multiple different timescales to a certain extent but this is still far from modelling the superior spectrotemporal adaptation exhibited by the human auditory system. Future work could look into identifying the combination of suitable frame lengths and filter channels to use to account for various important spectrotemporal contexts of different sound classes; characteristics of auditory model that when combined with the appropriate machine-learning approach best represents human performance in speech recognition in a variety of noise environments without requiring high processing demands; and incorporating behavioural and sensory context into auditory-inspired models to increase spectrotemporal adaptation.

REFERENCES

[1] Y. Lai et al., "Deep Learning–Based Noise Reduction Approach to Improve Speech Intelligibility for Cochlear Implant Recipients", Ear and Hearing, vol. 39, no. 4, pp. 795-809, Jul. 2018.

[2] P. Loizou and G. Kim, "Reasons why Current Speech-Enhancement Algorithms do not Improve Speech Intelligibility and Suggested Solutions", IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 1, pp. 47-56, Jan. 2011.

[3] Y. Hu, and P. Loizou, "Environment-specific noise suppression for improved speech intelligibility by cochlear implant users", The Journal of the Acoustical Society of America, vol. 127, pp. 3689-3695, Jun. 2010.

[4] Q. Fu and G. Nogaki, "Noise Susceptibility of Cochlear Implant Users: The Role of Spectral Resolution and Smearing", Journal of the Association for Research in Otolaryngology, vol. 6, no. 1, pp. 19-27, Apr. 2005.

[5] S. David, "Incorporating behavioral and sensory context into spectrotemporal models of auditory encoding", Hearing Research, vol. 360, pp. 107-123, Mar. 2018.

[6] S. Tabibi, A. Kegel, W. Lai and N. Dillier, "Investigating the use of a Gammatone filterbank for a cochlear implant coding strategy", Journal of Neuroscience Methods, vol. 277, pp. 63-74, Feb. 2017.

[7] J. Chen, Y. Wang and D. Wang, "A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no.12, pp. 1993-2002, Dec. 2014.

[8] J. Monaghan et al., "Auditory inspired machine learning techniques can improve speech intelligibility and quality for hearing-impaired listeners", The Journal of the Acoustical Society of America, vol. 141, no. 3, pp. 1985-1998, Mar. 2017.

[9] E. Girden, "ANOVA", Newbury Park, Calif.: Sage Publ., 2003.

[10] R. Patterson, M. Allerhand and C. Giguère, "Time□domain modelling of peripheral auditory processing: A modular architecture and a software platform", The Journal of the Acoustical Society of America, vol. 98, no. 4, pp. 1890-1894, Oct. 1995.

[11] R. Patterson et al., "An efficient auditory filterbank based on the gammatone function", Institute of Acoustics on Auditory Modelling, pp. 1-33, Dec. 1987

[12] T. Irino and R. Patterson, "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform", Speech Communication, vol.36, no. 3-4, pp. 181-203, Mar. 2002.

[13] Y. Wang, K. Han and D. Wang, "Exploring Monaural Features for Classification-Based Speech Segregation", IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 2, pp. 270-279, Feb. 2013.

[14] F. Saki and N. Kehtarnavaz, "Real-time hierarchical classification of sound signals for hearing improvement devices", Applied Acoustics, vol. 132, pp. 26-32, Mar. 2018.

[15] L. Breiman, "Random Forests", Machine Learning, vol. 45, no. 1, pp.5-32, Oct. 2001.

[16] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms", Speech Communication, vol. 49, no. 7-8, pp. 588-601, Jul. 2007.

[17] X. Teng, X. Tian and D. Poeppel, "Testing multi-scale processing in the auditory system", Scientific Reports, vol. 6, no. 1, Oct. 2016.