Proceedings Paper:

Romero, H., Ma, N., Brown, G. orcid.org/0000-0001-8565-5476 et al. (2 more authors) (2019) Deep learning features for robust detection of acoustic events in sleep-disordered breathing. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-2019). IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 12-17 May 2019, Brighton, UK. IEEE . ISBN 9781479981311

https://doi.org/10.1109/ICASSP.2019.8683099

DEEP LEARNING FEATURES FOR ROBUST DETECTION OF ACOUSTIC EVENTS IN SLEEP-DISORDERED BREATHING

Hector E. Romero, Ning Ma, Guy J. Brown, Amy V. Beeston and Madina Hasan

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK
{heromeroramirez1, n.ma, g.j.brown, a.beeston, m.hasan}@sheffield.ac.uk

ABSTRACT

Sleep-disordered breathing (SDB) is a serious and prevalent condition, and acoustic analysis via consumer devices (e.g. smartphones) offers a low-cost solution to screening for it. We present a novel approach for the acoustic identification of SDB sounds, such as snoring, using bottleneck features learned from a corpus of whole-night sound recordings. Two types of bottleneck features are described, obtained by applying a deep autoencoder to the output of an auditory model or a short-term autocorrelation analysis. We investigate two architectures for snore sound detection: a tandem system and a hybrid system. In both cases, a 'language model' (LM) was incorporated to exploit information about the sequence of different SDB events. Our results show that the proposed bottleneck features give better performance than conventional mel-frequency cepstral coefficients, and that the tandem system outperforms the hybrid system given the limited amount of labelled training data available. The LM made a small improvement to the performance of both classifiers.

Index Terms— Sleep-disordered breathing, deep learning, hidden Markov model, bottleneck features, corpus

1. INTRODUCTION

Sleep-disordered breathing (SDB) is caused by the partial or complete collapse of the upper airway during sleep, whose forms include snoring, upper airway resistance syndrome (UARS), and obstructive sleep apnea (OSA) [1]. Snoring is produced by the vibration of structures such as the soft palate, epiglottis, pharyngeal walls, and tongue, due to turbulent airflow caused by partial collapse of the upper airway [2]. UARS is indicated by abnormal respiratory effort originating from the limited airflow [3]. A complete collapse of the upper airway results in OSA, characterised by an absence of airflow that interrupts sleep. OSA has a strong relationship with cardiovascular [4], metabolic, and neurocognitive diseases, affecting approximately 24-26% of men and 9-28% of women in Europe and the United States [5], and its prevalence is increasing [6].

The current gold standard for diagnosing SDB is polysomnography (PSG) [1, 7, 5, 8], which is expensive, time consuming, and uncomfortable for the patient. It involves sleeping for a complete night in a laboratory while physiological parameters are measured via at least 22 wired attachments to the body. A further problem is that data obtained from PSG in a hospital laboratory may not be representative of that which would be recorded in the home [7, 5]. For this reason, alternatives to the diagnosis of SDB have been explored including at-home PSG [9, 8] and smartphone-based solutions [10, 11]. In particular, smartphone apps offer the potential for a convenient, non-invasive, and low-cost method to diagnose SDB through acoustic analysis [12].

The focus of this paper is the detection of SDB sounds in whole-night audio recordings. Previous studies have typically used high-quality recordings made close to the patient's head while sleeping in clinic [1]; however a smartphone-based solution must work with higher levels of background noise, microphones that are designed for close-talking rather than ambient sound recording, and haphazard placement of the device. Recently, deep neural networks (DNNs) have proven to be very effective in such sound classification tasks. However, DNNs have not been fully exploited in SDB sound classification because it is a low-resource task; there are few (if any) large corpora of labelled SDB acoustic data.

Previous studies have used a variety of acoustic features and classifier architectures. Nonaka et al. [13] described a method to classify sleep audio recordings using a logistic regression classifier and features derived from the auditory image model (AIM). In another study, Dafna et al. [1] used an AdaBoost classifier with 34 features, such as periodicity, total energy, duration, and higher-order spectral statistics. Two studies are notable for their use of deep learning. Firstly, Emoto et al. [7] described a method for detecting low intensity snore and breathing events using a DNN. Good accuracy was obtained on a subject-specific task (91.8%) but performance was poorer on a subject-independent task (75.7%), underlining the problem of generalising a DNN when training data is limited. Secondly, Amiriparian et al. [14] proposed a deep learning approach for snore sound classification. They overcame the low-resource problem by using a convolutional neural network (CNN) that was pre-trained on an image classification task to generate deep spectrum features, which were then classified by a support vector machine (SVM).

The current paper introduces a number of innovations. First, we describe a corpus of SDB acoustic data recorded via smartphones in domestic environments, and an annotation scheme for labelling that data. Second, in order to leverage a large volume of unlabelled data, we employ bottleneck features for SDB classification learned by a deep autoencoder. Noting that conventional audio features such as mel-frequency cepstral coefficients (MFCCs) do not give a good representation of pitch, which might be important in SDB sound classification, we propose bottleneck features learned from the short-term autocorrelation function (ACF) in addition to those learned from an auditory representation of the sound spectrum. To further address the low-resource nature of this task, we investigate two snore detection architectures. The first is a *tandem* approach in which a hidden Markov model (HMM)-Gaussian mixture model (GMM) system is used to model the acoustics of the bottleneck features. The second is a *hybrid* approach in which a second DNN is used to classify the bottleneck features. Finally, we investigate the utility of a 'language model', obtained from our labelled data set, which captures information about the temporal sequence of different SDB events.

H.E. Romero was funded by a PhD studentship from Passion for Life Healthcare (PFLH) and the Department of Computer Science, University of Sheffield. A.V. Beeston was funded by KTP award 9905 with PFLH.

The remainder of the paper is organised as follows. Section 2 describes the SDB corpus and proposes an annotation scheme. Following this, the bottleneck features and classifier architectures are described in Section 3. An evaluation is presented in Section 4, using MFCCs as a baseline for comparison. The final two sections of the paper discuss our results and make some concluding remarks.

2. SLEEP BREATHING SOUND CORPUS

Given the lack of a suitable SDB corpus, large-scale data collection was undertaken. Whole-night recordings were made of 31 male and 13 female participants in their own homes, using a custom app for iOS devices (e.g., iPhone, iPad or iPod Touch). The device was placed at head level within arm's reach, while the participant was normally sleeping. Audio recordings were single channel, sampled at 16 kHz with 16 bit depth. Data collection and storage protocols were subjected to the ethical review procedures of the University of Sheffield.

2.1. Data selection

A subset of the data was selected for annotation. Since SDB sounds such as snoring occur sporadically throughout a whole night, a simple GMM-based classifier was implemented to identify signals of interest. Specifically, whole-night recordings were divided into 2-minute segments and those which contained at least 20% of snore sound were identified, according to the automatic classification. From those, 25 segments were selected for each of six male participants, amounting to a total of 5 hours data. To balance the amount of data for other sound classes, such as background noise, a further 54 minutes of audio recordings were included which were made in a home environment when the participants were not sleeping.

2.2. Annotation

Table 1 shows the scheme used for annotating the SDB sound recordings. In contrast to previous studies such as [1, 13, 7], which only considered 'snore' and 'non-snore', here we adopt a more detailed annotation scheme. Specifically, six event types were defined for annotation: *snore*, *breath*, *silence*, *wheezing*, *noisy in-breath*, and *other*. This approach provides more flexibility, since similar event types can be merged if necessary. For the purposes of the snore detection experiments described in this paper, the *snore*, *wheezing*, and *noisy in-breath* classes were merged into a single *snore* class.

Each 2-minute signal was labelled by one of three annotators. To confirm the validity of this approach (i.e., to ensure that all annotators shared a common understanding of the annotation scheme), two audio recordings were first segmented and labelled by all three annotators, and compared at the frame level. A Cohen's kappa [15] of 0.61±0.17 was obtained. Given the greater level of detail in our annotation scheme (i.e., six labels rather than two), this value suggests an acceptable level of agreement between annotators.

3. SYSTEM DESCRIPTION

Robust detection of acoustic events in SDB is challenging because breathing sounds display great variability between individuals. Like speech, SDB sounds vary according to the characteristics of an individual's vocal tract, and other factors can influence breathing sounds during sleep (e.g., medical conditions such as asthma and emphysema [16]). Audio recordings made via smartphones in a home environment present further variabilities, due to the microphone used in different phones, placement of the device, and domestic noises.

Table 1. Annotation scheme

Acoustic event	Description
Snore	Pitched breathing sound.
Wheezing	Whistling sound produced in the respiratory airways during breathing.
Noisy in-breath	Like a snore, but without a strong pitch.
Breath	Unpitched breathing sound, high frequency noise.
Other	Speech, street noise, alarm clock, etc.
Silence	Hiss, no other structure.

For these reasons, various features and classifiers were studied with the objective of developing a robust SDB classifier. The following two sections describe the proposed bottleneck features and classifier architectures.

3.1. Bottleneck features

Annotating the SDB data is time-consuming: it takes on average 30 minutes to annotate a 2-minute section. The annotated corpus described in Section 2 represents only a small portion of the data collected. To leverage the large amount of unlabelled data, we adopted an unsupervised approach by using a deep autoencoder. The autoencoder was trained on 20 hours of unlabelled data and extracts bottleneck features specialised for SDB signals. These were derived from an auditory spectral representation and a pitch-based representation arising from the short-term ACF.

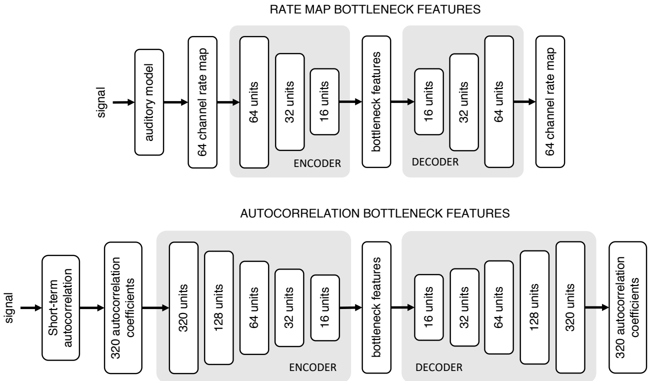

Fig. 1. Rate map and autocorrelation bottleneck features.

3.1.1. Bottleneck features from auditory nerve firing rate maps

Auditory-motivated representations have been successful in various sound understanding applications [17]. Here, we derived bottleneck features from a model of peripheral auditory processing based on a gammatone filterbank. 64 auditory filter channels were used, with centre frequencies spaced on the ERB-rate scale [18] between 80 Hz and 7500 Hz. The envelope was then extracted from each channel, forming a so-called rate map (RM). The envelope in each frequency channel can be interpreted as the instantaneous firing rate of an auditory nerve fibre [17].

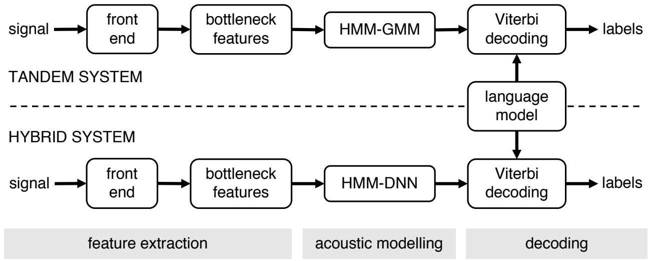

Fig. 2. System diagram showing both the tandem and the hybrid snore detection systems.

An autoencoder DNN was implemented in TensorFlow [19] to learn RM bottleneck features. It consists of three fully connected layers that encode RMs, followed by a further three fully connected layers that decode them (Fig. 1, upper panel). During training, the aim is to reconstruct the input at the output layer, via a compressed intermediate representation. The input to the autoencoder is the 64-channel RM, which is transformed to a compressed 16-channel representation by three layers of 64, 32 and 16 sigmoid units. After the RM is encoded, it is decoded (i.e., reconstructed) back to its original 64-channel form by three layers of 16, 32 and 64 sigmoid units. The bottleneck features are obtained from the output of the 16-unit encoder layer; the second half of the network is discarded after training. Bottleneck features were appended with first-order and second-order difference (deltas and accelerations), resulting in a 48-element feature vector.

3.1.2. Bottleneck features from the autocorrelation function

Snoring is a *pitched* acoustic event, since it is produced by the vibration of structures such as the soft palate, epiglottis and pharyngeal walls [2]. The RM bottleneck features, which can be seen as a compressed time-frequency representation, do not give a good representation of pitch, however. To provide pitch-related information to the system, we propose another kind of bottleneck features learned from the ACF. The short-term ACF is a popular means of pitch estimation [20] and is defined as:

$$A(\tau) = \sum_{n=0}^{N-1} y(n)y(n-\tau) \tag{1}$$

where τ is the time lag and y is a windowed frame with N samples. In this study $N = 400$ for a 25 ms window sampled at 16 kHz. The first 320 lags were selected as input to the deep autoencoder, giving a lower pitch limit of 50 Hz.

The ACF autoencoder is shown in the lower panel of Fig. 1, which follows the same procedure as that of the RM bottleneck features. A 320-lag ACF is encoded to a 16-channel representation with five fully connected layers of 320, 128, 64, 32 and 16 sigmoid activation units, and decoded back to its original form with another five fully connected layers of 16, 32, 64, 128 and 320 sigmoid activation units. Again, the aim during training is to reconstruct the input at the output layer. The 16-channel encoded ACF features were appended with first-order and second-order difference (deltas and accelerations), resulting in a 48-element feature vector as before.

3.2. Snore detection architectures

Given the limited amount of training data, two snore detection architectures were investigated using the bottleneck features, namely a *tandem system* and a *hybrid system*. A schematic overview of both systems is shown in Fig. 2.

3.2.1. Tandem system

In the tandem system, bottleneck features are extracted from the autoencoder DNN but they are modelled by a conventional HMM system with GMMs modelling the state distributions [21]. The number of HMM states used by each class is summarised in Table 2. Each state is represented by a GMM with 7 Gaussian components using diagonal covariance matrices. The system is implemented using HTK [22] and the parameters are selected heuristically according to the complexity of each class.

Table 2. Number of HMM states used in the tandem system

Class	Snore	Breath	Other	Silence
States	7	5	3	3

3.2.2. Hybrid system

In the hybrid system, the HMM state distributions are directly represented by class posterior probabilities from a DNN. The acoustic features are mapped by a DNN with 3 hidden layers to the class labels. For the system that takes separate bottleneck features (RM bottleneck or ACF bottleneck) as input, each hidden layer consists of 96 fully connected hidden units with sigmoid activation functions. For the system that combines both bottleneck features, each hidden layer uses 192 hidden units. Finally, the output layer has four softmax units. This system is implemented using TensorFlow [19].

3.3. Language model

Breathing sounds, like language, usually follows certain patterns, e.g., a snore sound typically begins with an inspiration and is followed by a brief period of silence before an expiration. A language model (LM), such as is normally used in speech recognition, might therefore be leveraged for more accurate snore detection during the decoding process.

HTK is used to compute event bigram statistics from the labelled training dataset. Given the limited amount of training data, the bigram model proves a suitable choice and is easy to incorporate. The LM is employed by both the tandem and the hybrid systems during the decoding process. In both systems, the Viterbi algorithm is employed to search for the most probable acoustic event sequence given the acoustic model and LM.

4. EVALUATION

4.1. Experiment setup

The annotated SDB corpus described in Section 2 was used for evaluation. The six-snorer data was divided into a training set which included all the annotated data from four snorers, and a test set consisting of data from the remaining two snorers. In this way, the systems we present here are ‘snorer-independent’.

For the tandem system, the HMM parameters were estimated from the training data using the Baum-Welch algorithm. For the hybrid system, the DNN was trained with a learning rate of 0.001, 60 epochs and a batch size of 256. During Viterbi decoding, an insertion penalty was used to balance the number of event deletion and insertion errors. Furthermore, a LM scaling factor was also introduced

Table 3. Snore event error rates for various systems

	Tandem		Hybrid	
	No LM	LM	No LM	LM
MFCCs	19.94%	17.59%	9.40%	9.52%
RM bottleneck	12.00%	12.13%	13.40%	13.40%
ACF bottleneck	15.24%	14.48%	14.92%	14.92%
RM+ACF bottleneck	10.86%	**8.89%**	10.22%	9.90%

Table 4. Frame-based snore F-measures for various systems

	Tandem		Hybrid	
	No LM	LM	No LM	LM
MFCCs	90.78%	91.67%	93.60%	93.45%
RM bottleneck	**95.29%**	95.23%	90.74%	90.74%
ACF bottleneck	88.34%	88.47%	86.96%	86.96%
RM+ACF bottleneck	94.43%	94.36%	94.73%	94.75%

for each system in order to adjust the impact of the LM. Given the limited amount of annotated data, a randomly selected subset of the training data was used to empirically optimise the two parameters.

4.2. Baseline systems

MFCCs have conventionally been used in automatic speech recognition (ASR) and non-speech-related tasks such as rare sound event detection [23] and snore detection [24]. Both a tandem system and a hybrid system using MFCCs, whose architectures are identical to the proposed systems, were introduced as the baseline systems. 12 MFCCs excluding the C0 coefficient were extracted for each 25 ms frame with a 10 ms step. They were further appended with deltas and accelerations, forming 36-dimensional feature vectors.

4.3. Evaluation metrics

All systems were evaluated both at event level and frame level. At event level, the snore event error rate (SER) was computed, which takes into account all the possible types of errors: insertions, deletions and substitutions, similar to the word error rate (WER) commonly used in ASR.

At frame level, the snore F-measure was computed to evaluate the segmentation quality of the snore detection systems. This metric takes into account both the precision (proportion of frames classified as ‘snore’ that are actually ‘snores’) and the recall (proportion of actual ‘snore’ frames that are classified as ‘snore’):

$$F\text{-measure} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (2)$$

5. RESULTS AND DISCUSSION

The event-level SERs and frame-level snore F-measures are shown in Tables 3 and 4, respectively. Considering the SER, the overall best performance was achieved by the tandem system using both RM and ACF bottleneck features with a LM (SER: 8.89%), outperforming the baseline MFCC system. The tandem systems using RM or ACF bottleneck features on their own produced significantly higher SERs, which suggests the DNN was able to exploit both the spectral and pitch information from the bottleneck features. A similar pattern of results can be seen for the hybrid systems. The use of the LM improved the performance in most systems, especially in the tandem systems, which showed an average relative improvement of 9%.

In most cases the tandem system outperformed the hybrid system. This is likely due to the amount of training data needed for the hybrid HMM-DNN system to perform well, while the HMM-GMM architecture adopted in the tandem system requires less data. This kind of behaviour has also been observed in other studies (e.g., [25]). An exception is the baseline MFCC system, which performed significantly better in the hybrid configuration (best SER: 9.40%) than in the tandem configuration (best SER: 17.59%). This pattern is also observed in the frame-level results discussed below; the reason for this discrepancy is currently under investigation.

Looking at the frame-based snore F-measure as a way to assess the segmentation quality, the best results were achieved by the tandem configuration using only RM bottleneck features (F-measure: 95.29%), although the best result using both RM+ACF bottleneck features is very close (F-measure: 94.75%). Comparing the F-measures and the SERs, it is reasonable to hypothesise that a F-measure above 93% demonstrates a reasonable quality in the segmentation, as a score below 93% corresponds to a significant drop in the SER.

It should be noted that in the hybrid bottleneck systems, feature extraction and acoustic modelling are currently done in two separate DNNs. This could potentially be a limitation. A better strategy might be to combine the two DNNs into a single network, therefore allowing the bottleneck DNN parameters to be adapted together with the acoustic DNN model.

6. CONCLUSIONS

Robust snore detection in a home environment, from recordings made using a smartphone, is a challenging task. There are two major problems: (i) collection and annotation of a large amount of SDB data; (ii) selection of appropriate acoustic features. We have described a solution to both problems by learning bottleneck features from a large corpus of unlabelled SDB data, and then employing a tandem architecture that makes the most of a limited amount of labelled data. We obtained the best performance using bottleneck features that encode both spectral shape and pitch information; these were shown to outperform conventional MFCC features when used in a tandem system.

Detailed annotation of sleep breathing sounds is a surprisingly difficult task due to the subjectivity involved in the process, the inherent variability of breathing sounds, and the significant amount of time that it demands. For example, there is no commonly accepted acoustic definition of a snore [2]. In our experience, annotators require careful training if they are to adhere to agreed standards such as the one shown in Table 1.

Although the contribution of a LM was small for some systems that we investigated, it does provide useful information in the tandem system. Preliminary work suggests that the LM helps to enforce realistic snore event durations. The LM is likely to have a larger effect when identifying conditions that involve a temporal sequence of events, such as OSA.

In the future we will focus on building systems to detect other forms of SDB, such as UARS and OSA. Such systems will build on the snore detection approaches described here, for example by using a bag-of-audio-words approach [26] to combine snore event detection with other acoustic features over a whole night recording.

7. REFERENCES

[1] E. Dafna, A. Tarasiuk, and Y. Zigel, "Automatic detection of whole night snoring events using non-contact microphone," *PLoS ONE*, vol. 8, no. 12, December 2013.

[2] D. Pevernagie, R. Aarts, and M. De Meyer, "The acoustics of snoring," *Sleep Medicine Reviews*, vol. 14, pp. 131–144, 2010.

[3] T. Masri and C. Guilleminault, "Upper airway resistance syndrome," in *Encyclopedia of Sleep*. Academic Press, 2013.

[4] G. Lee, L. Lee, C. Wang, N. Chen, T. Fang, C. Huang, W. Cheng, and H. Li, "The frequency and energy of snoring sounds are associated with common carotid artery intima-media thickness in obstructive sleep apnea patients," *Nature Scientific Reports*, 2016.

[5] M. Shokoueinejad, C. Fernandez, E. Carroll, F. Wang, J. Levin, S. Rusk, N. Glattard, A. Mulchrone, X. Zhang, A. Xie, M. Teodorescu, J. Dempsey, and J. Webster, "Sleep apnea: a review of diagnostic sensors, algorithms, and therapies," *Physiological Measurement*, vol. 38, pp. 204–252, 2017.

[6] BBC, "Sleep disorder testing carried out by NHS doubles," http://www.bbc.co.uk/news/uk-england-40122979, June 2017.

[7] T. Emoto, U. R. Abeyratne, K. Kawano, T. Okada, O. Jinnouchi, and I. Kawata, "Detection of sleep breathing sound based on artificial neural network analysis," *Biomedical Signal Processing and Control*, vol. 41, pp. 81–89, 2018.

[8] F. Mendonça, S. S. Mostafa, A. G. Ravelo-Garcia, F. Morgado-Dias, and T. Penzel, "Devices for home detection of obstructive sleep apnea: A review," *Sleep Medicine Reviews*, 2018.

[9] J. Miller, P. Schulz, B. Pozehl, D. Fiedler, A. Fial, and A. M. Berger, "Methodological strategies in using home sleep apnea testing in research and practice," *Sleep and Breathing*, November 2017.

[10] R. Nandakumar, S. Gollakota, and N. Watson, "Contactless sleep apnea detection on smartphones," in *MobiSys 2015*, 2015, pp. 45–57.

[11] SnoreLab, "SnoreLab," https://snorelab.com, 2018.

[12] S. K. Koo, S. B. Kwon, Y. J. Kim, J. S. Moon, Y. J. Kim, and S. H. Jung, "Acoustic analysis of snoring sounds recorded with a smartphone according to obstruction site in OSAS patients," *European Archives of Oto-Rhino-Laryngology*, vol. 274, pp. 1735–1740, 2017.

[13] R. Nonaka, T. Emoto, U. R. Abeyratne, O. Jinnouchi, I. Kawata, H. Ohnishi, M. Akutagawa, S. Konaka, and Y. Kinouchi, "Automatic snore sound extraction from sleep sound recordings via auditory image modeling," *Biomedical Signal Processing and Control*, vol. 27, pp. 7–14, 2016.

[14] S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, and B. Schuller, "Snore sound classification using image-based deep spectrum features," in *Proceedings of Interspeech*, Stockholm, Sweden, 2017, pp. 3512–3516.

[15] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[16] P. Forgacs, A. R. Nathoo, and H. D. Richardson, "Breath sounds," *Thorax*, vol. 26, no. 3, pp. 288–295, 1971.

[17] G. J. Brown and D. L. Wang, "Fundamentals of computational auditory scene analysis," in *Computational Auditory Scene Analysis*, G. J. Brown and D. L. Wang, Eds., pp. 1–44. IEEE, 2006.

[18] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, pp. 103–138, 1990.

[19] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation*. USENIX, 2016, pp. 265–283.

[20] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Am.*, vol. 11, no. 4, pp. 1917–1930, 2002.

[21] S. P. Rath, K. M. Knill, A. Ragni, and M. J. F. Gales, "Combining tandem and hybrid systems for improved speech recognition and keyword spotting on low resource languages," in *Interspeech Proceedings 2014*. Interspeech, 2014.

[22] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book, HTK version 3.4 edition*, Cambridge University Engineering Department, 2006.

[23] J. Gutierrez, R. Fraile, A. Camacho, T. Durand, J. Jarrin, and S. Mendoza, "Synthetic sound event detection based on MFCC," in *DCASE proceedings*. DCASE, September 2016.

[24] W. D. Duckitt, S. K. Tuomi, and T. R. Niesler, "Automatic detection, segmentation and assessment of snoring from ambient acoustic data," *Physiological Measurement*, vol. 27, pp. 1047–1056, 2016.

[25] J. Schröder, J. Anemüller, and S. Goetze, "Peformance comparison of GMM, HMM and DNN based approaches for acoustic event detection within task 3 of the DCASE 2016 challenge," in *DCASE proceedings*. DCASE, September 2016.

[26] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, "A bag-of-audio-words approach for snore sounds' excitation localisation," in *Speech Communication; 12. ITG Symposium*, 2016, pp. 230–234.