SNORER DIARISATION BASED ON DEEP NEURAL NETWORK EMBEDDINGS

Hector E. Romero, Ning Ma and Guy J. Brown

Department of Computer Science, University of Sheffield, Sheffield S1 4DP, UK

{heromeroramirez1, n.ma, g.j.brown}@sheffield.ac.uk

ABSTRACT

Acoustic analysis of sleep breathing sounds using a smartphone at home provides a much less obtrusive means of screening for sleep-disordered breathing (SDB) than assessment in a sleep clinic. However, application in a home environment is confounded by the problem that a bed partner may also be present and snore. This paper proposes a novel acoustic analysis system for *snorer diarisation*, a concept extrapolated from speaker diarisation research, which allows screening for SDB of both the user and the bed partner using a single smartphone. The snorer diarisation system involves three steps. First, a deep neural network (DNN) is employed to estimate the number of concurrent snorers in short segments of monaural audio recordings. Second, the identified snore segments are clustered using snorer embeddings, a feature representation that allows different snorers to be discriminated. Finally, a snore transcription is automatically generated for each snorer by combining consecutive snore segments. The system is evaluated on both synthetic snore mixtures and real two-snorer recordings. The results show that it is possible to accurately screen a subject and their bed partner for SDB in the same session from recordings of a single smartphone.

***Index Terms*—** Snorer diarisation, sleep-disordered breathing, deep neural network embeddings, LSTM.

1. INTRODUCTION

The most prevalent forms of sleep-disordered breathing (SDB) are snoring, caused by a partial collapse of the upper airway during sleep, and obstructive sleep apnea (OSA) due to a complete collapse. Polysomnography (PSG) is the current gold standard for diagnosing SDB, but it is obtrusive, time-consuming and expensive [1, 2, 3, 4]. It requires the patient to sleep in a laboratory for a whole night, while multiple sensors attached to them measure physiological parameters such as oxygen saturation and respiratory effort [5]. The reliability and consistency of this test has been questioned because of the "first-night effect" [6] – due to limited movement and discomfort from wearing the sensors, and the psychological effects of being observed, sleep in a laboratory during the first night presents "more wakefulness, less total sleep time, and a lower sleep efficiency index" [7, p. 246]. Despite this, most clinical sleep evaluations are restricted to one night due to the costs involved. Improving accessibility to SDB diagnosis and treatment is, therefore, one of the key challenges that sleep medicine is currently facing [1].

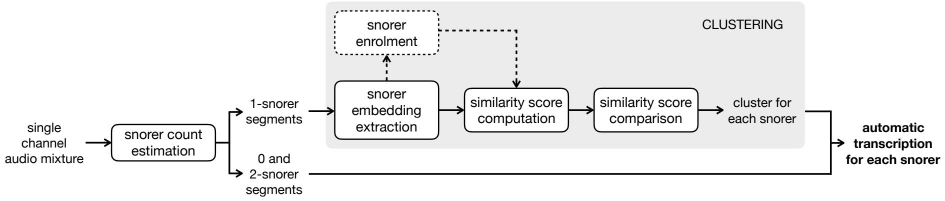
Acoustic analysis of breathing sounds using a smartphone offers an unobtrusive and inexpensive alternative to screening for SDB. Previous studies have mainly focused on detecting snore events in audio recordings of subjects sleeping on their own in a controlled condition (e.g., in a sleep clinic). Dafna et al. [1] used features such as total energy, periodicity, higher-order spectral statistics and duration with an AdaBoost classifier. Nonaka et al. [8] exploited auditory image modelling (AIM) features, and a logistic regression classifier. Emoto et al. [2] detected low intensity respiratory events using a subject-independent deep neural network (DNN). Amiriparian et al. [9] estimated the location of obstruction in the upper airway from snore events using a support vector machine (SVM) with features generated by a convolutional neural network (CNN). Romero et al. [10] proposed a DNN system using bottleneck features for robust detection of SDB events with a n-gram "language model" to better model the sequence of SDB events.

Applying acoustic analysis of breathing sounds in a natural sleep environment (e.g., in a bedroom at home) must deal with various levels of interfering noise and room acoustics. In particular, the breathing sounds of the bed partner is a main source that can negatively impact upon the performance of the acoustic analysis system, due to the proximity to the microphone. With a single smartphone, the system needs to distinguish the breathing sounds of both snorers, in order to correctly diagnose SDB for each individual. This situation resembles that of speaker diarisation [11], which aims to identify "who speaks when" in a multi-speaker audio recording. However, the current task has some particular constraints. First, speaker diarisation systems generally employ acoustic beamforming with a microphone array. Some studies have also attempted speaker diarisation using single-channel audio recordings, e.g., the deep clustering approach [12, 13], but in general the microphone is not in alignment with all the speakers. In contrast, in the SDB diagnosis task a single smartphone is used and typically the smartphone is positioned on a bedside table, resulting in an arrangement of both snorers and the phone in a straight line. This arrangement makes separation of snorers' breathing sounds more challenging. Second, for speaker diarisation the speakers are usually aware of each other speaking, and thus there is less cross-talk and the main task is to identify the speaker. In comparison, two snorers do not consciously interact, and overlapping breathing sounds are more likely to occur.

To address the challenges faced in designing a smartphone-based solution for assessing SDB at home, we propose a novel single-channel approach to snorer diarisation which can handle interfering snore sounds from the bed partner. First, the number of concurrent snorers in each acoustic segment is estimated. Next, segments containing a single snorer are clustered based on snorer embeddings, which enable different snorers to be discriminated. Finally, an automatic transcription is generated for each snorer by combining adjacent snore segments. Such a framework allows the screening of both snorers for SDB during the same session.

The remainder of this paper is organised as follows. Section 2 describes the snorer diarisation system and the data used to develop it. Section 3 presents the evaluation framework. The results are discussed in Section 4 and concluding remarks are made in Section 5.

H. E. Romero was funded by a PhD studentship from Passion for Life Healthcare and the University of Sheffield. N. Ma and G. J. Brown were supported by Innovate UK Open Grant 26767.



snorer enrolment
CLUSTERING
single channel audio mixture
snorer count estimation
1-snorer segments
snorer embedding extraction
similarity score computation
similarity score comparison
cluster for each snorer
0 and 2-snorer segments
automatic transcription for each snorer

Fig. 1. System diagram of the proposed snorer diarisation system.

2. SYSTEM DESCRIPTION

An overview of the proposed snorer diarisation system is given in Fig. 1. The system takes as input a single-channel sleep audio recording containing two snorers, and generates an automatic transcription of snore events for each snorer. This is carried out by two subsystems: (i) snorer count estimation, and (ii) clustering of single-snorer events. The first subsystem divides the audio recording into non-overlapping 250 ms segments and estimates the number of snorers present in every segment: 0 if there are no snore events, 1 if there are snore events that do not concurrently occur, and 2 if there are concurrent snore events. Then, the identified single-snorer segments are passed to the clustering subsystem which extracts snorer embeddings from the segments using a DNN framework. The snorer embeddings are a feature representation that enables different snorers to be clustered based on their similarity scores. After clustering, the single-snorer segments are combined with the 0- and 2-snorer segments to generate snore event transcriptions for each snorer. In this way, the proposed snorer diarisation system does not require separation of each subject's breathing sounds.

2.1. Snorer count estimation

The single-channel snore mixtures are first divided into non-overlapping segments of 250 ms. The short-time Fourier transform (STFT) is computed with a frame rate of 10 ms and a frame size of 25 ms using a Hann window. Therefore, each segment consists of 25 frames of STFT magnitude feature vectors. The feature vectors are log-compressed before being standardised with zero means and unit variances with respect to the training dataset.

We treat snorer count estimation as a classification problem, and therefore use a DNN in this subsystem. Two DNN architectures are investigated. The first DNN is a fully-connected dense network. The STFT features of all the 25 frames within a segment are flattened as input to the network. The hidden layers consist of three fully connected dense layers. Each layer has 512 sigmoid activation units with a 50% dropout rate. The second network is a bidirectional long short-term memory (BLSTM) system. As shown in Fig. 2, it employs three BLSTM layers with 30, 20 and 40 units, respectively, followed by a max-pooling layer. A similar architecture has proven successful in speaker count estimation [14]. For both networks, the output layer is a dense layer with three softmax activation units, one for each of the "classes": 0, 1 and 2 snorers.

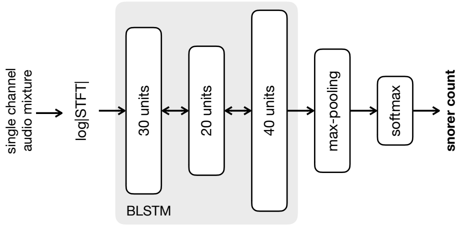


Fig. 2. The BLSTM architecture for snorer count estimation.

Both neural networks were developed using TensorFlow [15]. They were trained with a learning rate of 0.001, a batch size of 128 segments, a dropout rate of 50%, and categorical cross-entropy as the loss function. These hyperparameters were set heuristically initially, and then optimised using a validation dataset. 25 epochs were required to achieve convergence.

2.2. Clustering of snore events

The 0- and 2-snorer segments identified by the snorer count estimation step are directly included in the transcription for both snorers, whereas the 1-snorer segments are passed to the clustering subsystem for allocation of a snorer. Clustering of 1-snorer segments is achieved by first extracting snorer embeddings with a deep neural network. Snorer embeddings can be seen as a learned feature representation (similar to bottleneck features [10]) that distinguish one snorer from another. There are in total 41 snorers in the dataset used in this study. As illustrated in Fig. 3, the neural network for extracting the snorer embeddings consists of one layer of 512 long short-term memory (LSTM) units, followed by a dense layer of 128 linear units and an output layer of 41 softmax activation units, one for each of the snorers. The input to the network is also the log-compressed STFT magnitude features, as used in the snorer count estimation step. During training, the objective of the network is to classify the 41 snorers in the training data. At testing time, the output softmax layer of the trained network is discarded, and the snorer embedding is the output of the last hidden layer. The embeddings are then used as features for clustering. Similar techniques have been shown to be effective for speaker verification using speaker embeddings [16].

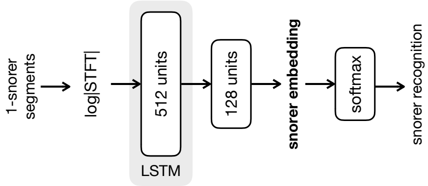


Fig. 3. The LSTM network for snorer embedding extraction.

The LSTM neural network was trained with a learning rate of 0.001, and a batch size of 128. Batch normalisation was applied after each layer except for the last one. Categorical cross-entropy was used as the loss function. Using a validation dataset, convergence was reached with 30 epochs. We also investigated a standard DNN architecture similar to one used for speaker recognition [16]. It replaces the LSTM layer described above with four fully connected layers of 512 sigmoid activation units, and batch normalisation is applied to the output of each of these layers. The input features are the flattened log-compressed STFT magnitudes.

Snorer embeddings are then obtained from the last hidden layer for snorer clustering. Five snorer embeddings for each snorer in the

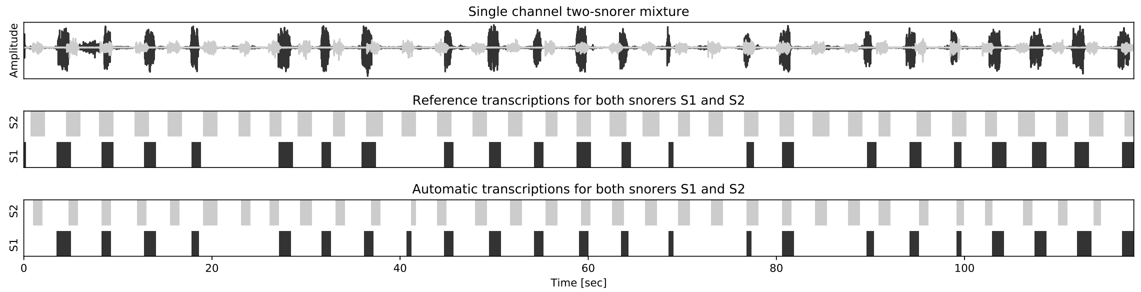


Fig. 4. Snorer diarisation example for a 2-snorer mixture. Snorer 1 and snorer 2 are shown in black and grey, respectively.

evaluation set are used for snorer enrolment, which correspond to one or two snore events each. These five embeddings per snorer are used as the reference embeddings. Clustering is done by computing the average similarity score (i.e., cosine similarity) between a given snore segment and the reference embeddings of each snorer. The segment is assigned to the snorer with the largest similarity score:

$$\hat{s} = \underset{s}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \frac{f^\top f(s_i)}{\|f\| \, \|f(s_i)\|} \tag{1}$$

where $\hat{s}$ is the snorer that the test segment is assigned to, f is the embedding of the test segment, s is one of the snorers in the mixture, and $f(s_i)$ is the reference embedding $i \in [1, 5]$ of snorer s. Finally, an automatic transcription is obtained for each snorer by combining consecutive snore segments. An output example is shown in Fig. 4.

2.3. Alternative enrolments

Ideally, the reference embeddings should be extracted from real snore recordings of each snorer. In practice, obtaining such enrolment snore sounds may not be always possible. This section investigates two alternative snorer enrolment methods: (i) using speech sounds from the subject, and (ii) using snore sounds simulated by the subject while awake.

Previous studies have investigated the correlation between speech and SDB. Fiz et al. [17] studied the harmonics of vowels vocalised by healthy individuals and those with OSA, and found differences in vocalisation between the two groups. Robb et al. [18] analysed the formant frequencies and bandwidths of prolonged vowels uttered by subjects with and without OSA, and reported lower formant values and wider formant bandwidth for the OSA group in comparison with the healthy one. Fernandez et al. [19] created a speech corpus of OSA and healthy Spanish-speaking subjects, and described differences in nasalisation between both groups. Glodshtein et al. [20] used features extracted from vowel and nasal phonemes to distinguish between OSA and healthy Hebrew-speaking subjects. Botelho et al. [21] used features computed from read speech, elongated vowels, and spontaneous speech to classify between OSA and healthy Portuguese-speaking individuals. These studies suggest that enrolling snorers using speech is plausible, as speech and SDB are correlated. We therefore propose to use a read sentence that highlights relevant vowel and nasal phonemes for snorer enrolment. An example of such sentences is: *"why women and men are on my main ammonium moon"*.

If only one of the snorers is available to produce the enrolment recordings, a similarity score threshold can be used in the clustering step. Specifically, clustering can be done by computing the average similarity score between a given 1-snorer segment and the reference embedding from the enrolled snorer. A 1-snorer segment is assigned to the enrolled snorer if the similarity score is above a threshold optimised on the validation dataset, and to the other snorer otherwise.

3. EXPERIMENTS AND EVALUATION

3.1. Synthetic snore mixtures

The manually annotated sleep breathing sound corpus from our previous study [10] was used to generate synthetic two-snorer mixtures. The corpus consists of audio recordings from six snorers (50 minutes per snorer), collected with a smartphone in domestic environments. The signal to noise ratio (SNR) of the recordings is relatively low as they were recorded using smartphones that are designed for close talking. An adaptive noise suppression algorithm [1] was applied and a bandpass filter (20 Hz – 6 kHz) was further employed to attenuate low and high frequency noise that might be present. All recordings were normalised to the same root mean square (RMS) level before mixing. Since snore recordings often contain large portions of silence, only the non-silence portions were considered for computing the RMS level. Three positive SNRs were used: 5 dB, 10 dB or 20 dB to simulate the scenario where a smartphone is usually placed on a bedside table next to one of the snorers.

For snorer count estimation, we mixed every possible pair of audio recordings from different snorers. Four snorers were used for training and the remaining two snorers were split between validation (48%) and testing (52%). The training dataset included 76,800 (250 ms long) segments extracted for each class, totalling 16 hours. The validation and test datasets contained 9,600 segments for each class (2 hours). The labels were automatically generated from the manual annotations of snore events.

For snore event clustering, a large number of snorers was needed to learn the embeddings that can discriminate between snorers. There are in total 44 snorers in the entire database [10]. We selected 41 snorers for training the embedding extraction network. For each snorer, 3,000 snore segments of duration 250 ms were extracted according to automatically generated annotations. Among them, 2,400 segments were used for training and 300 segments were used for validation and testing. The 41 snorers used to develop this subsystem were different from the two snorers used to evaluate the complete snorer diarisation system (i.e., the task is "snorer-independent").

3.2. Real snore mixtures

In addition to the synthetic snorer mixtures, we collected real two-snorer sleep audio recordings using a smartphone in a domestic bedroom. Annotating two-snorer sleep audio recordings is a complex task as one has to assign each snore event to one (or both) of the snorers. Given the variability of breathing sounds and the subjectivity in the annotation process, a proper reference is difficult to obtain

Table 1. Results for the proposed snorer diarisation systems.

	STANDARD DNN				**LSTM**			
Test Data	**Synthetic Mixtures**	**Real Mixtures**			**Synthetic Mixtures**	**Real Mixtures**		
Enrolment	**Real Snore**	**Real Snore**	**Sim. Snore**	**Speech**	**Real Snore**	**Real Snore**	**Sim. Snore**	**Speech**
Precision	74.86%	54.80%	49.38%	55.04%	71.30%	52.11%	51.44%	51.34%
Sensitivity	67.63%	66.81%	60.52%	67.46%	67.68%	59.00%	58.13%	58.35%
Specificity	89.95%	92.47%	91.52%	92.47%	87.95%	92.58%	92.49%	92.44%
Accuracy	83.10%	89.38%	87.79%	89.46%	81.73%	88.54%	88.36%	88.34%
F-measure	71.06%	60.22%	54.39%	60.62%	69.44%	55.34%	54.58%	54.62%
Segment DER	16.90%	10.62%	12.21%	10.54%	18.27%	11.46%	11.64%	11.66%
Event DER	20.52%	51.67%	56.67%	49.17%	20.99%	49.17%	50.83%	50.83%

from single channel audio recordings alone. For this reason, during the recording session, two-channel recordings were additionally made with microphones placed on the bedhead close to each snorer. In this way a proper reference could be obtained to manually annotate the real two-snorer sleep audio recordings.

3.3. Evaluation metrics

The performance of the snorer diarisation system was evaluated at both segment-level and event-level. Segment-level evaluation allows the computation of the precision, sensitivity, specificity, F-measure and the segment-level diarisation error rate (DER):

$$\text{DER}_{\text{segment}} = \frac{\text{FP} + \text{FN}}{\text{reference segment count}} \qquad (2)$$

where FP and FN are the false positives and the false negatives, respectively. This provides information on the quality of the segmentation, which is relevant if the time spent snoring is to be reported by the system. Event-level evaluation is achieved by merging consecutive segments with the same snorer count into a single event, and aligning these events with the reference transcription. Since there are only two kinds of events (i.e., snore and non-snore), no substitutions are observed. The event-level DER takes into account the insertion and deletions errors, and is defined as:

$$\text{DER}_{\text{event}} = \frac{\text{insertions} + \text{deletions}}{\text{reference event count}} \qquad (3)$$

4. RESULTS AND DISCUSSION

The subsystems for snorer count estimation and clustering of snore events attained an average sensitivity of 74% and 80%, respectively, when evaluated separately. Table 1 presents the results produced by the proposed snorer diarisation systems. For synthetic snore mixtures, both the standard DNN system and the LSTM system performed well using real snores as enrolment data, achieving a specificity (true negative rate) above 88%. This shows that the systems are able to effectively discriminate snore from non-snore events. The overall performance of the standard DNN system is slightly better than the LSTM system, which is likely due to the limited amount of data available to train the systems, i.e., more complex network architectures require a greater amount of data than standard architectures.

When evaluated on real snore mixtures, using real snoring as enrolment data, the systems achieved significantly lower precision, which suggests the classifier generated many false positives. This shows a limitation of the proposed systems, as they were trained only on the synthetic mixtures, which introduces a mismatch between the training data and the test data. The large difference in event DERs could be due to the number of snore events in the audio data. The audio recordings used to generate the synthetic mixtures contain a large amount of snore events, whereas the real two-snorer audio recordings contain fewer snore events, so the number of false positives increases.

Enrolment using real snore sounds from individual snorers may not be convenient in practice. We investigated two alternatives: (i) asking the patient to simulate snoring while awake, and (ii) using speech of the patient for enrolment. Initial analysis shows that there are clear differences in the acoustic characteristics between a real and a simulated snore produced by the same subject. With respect to real snores, the average duration of simulated snores was increased by 42%, the average pitch was decreased by 53%, and the average spectral centroid was increased by 62%. Enrolment with speech sounds would ideally require the snorer embedding extractor network (Section 2.2) to be trained with both real snoring and speech. In this way, the network could learn a vocal tract embedding. However, the results show that, even for a snorer embeddings obtained only from real snores, it is possible to cluster 1-snorer events by providing a simulated snore or speech signal for enrolment with the same performance as using real snore sounds.

Although the snorer diarisation system that we introduce here resembles a speaker diarisation system, it is worth noting some important differences. First, speaker diarisation systems do not normally take into account overlapping speech for evaluation [22], whereas our system does consider concurrent snoring. This is because, unlike a conversation [23] where speakers are aware of each other speaking, there could be a significant number of snores that overlap during sleep. Second, our system assumes that there is a maximum of two snorers, while in speaker diarisation systems the maximum number of speakers is not known. Third, speaker diarisation systems typically output a single transcription for all the speakers. In contrast, the proposed snorer diarisation system produces a separate transcription for each snorer, in order to assess both snorers in the same session.

5. CONCLUSIONS

This paper has introduced the problem of single channel snorer diarisation, which arises when screening for SDB in a home environment where both a patient and their bed partner may be present. We have described a solution to this problem by applying deep learning to overcome the challenges posed by single channel sleep audio recordings. Our proposed solution does not require separation of each subject's breathing sounds, and allows two people to be screened for SDB in the same session.

Currently the snore count estimation and the clustering stages are trained separately and employed in sequence. Future work will investigate if the two stages can be more tightly coupled and optimised within an unified framework. We also intend to use this system as a component in a system for OSA monitoring and diagnosis.

6. 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] 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.

[3] 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.

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

[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] K. N. Hutchison, Y. Song, L. Wang, and B. A. Malow, "Analysis of sleep parameters in patients with obstructive sleep apnea studied in a hospital vs. a hotel-based sleep center," *Journal of Clinical Sleep Medicine*, vol. 4, no. 2, pp. 119–122, 2008.

[7] W. Lu, J. Cantor, R. N. Aurora, M. Nguyen, T. Ashman, L. Spielman, A. Ambrose, J. W. Kerllman, and W. Gordon, "Variability of respiration and sleep during polysomnography in individuals with TBI," *Neurorehabilitation*, vol. 35, pp. 245–251, 2014.

[8] 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.

[9] 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.

[10] H. E. Romero, N. Ma, G. J. Brown, A. V. Beeston, and M. Hasan, "Deep learning features for robust detection of acoustic events in sleep-disordered breathing," in *Proceedings of ICASSP 2019*. 2019, IEEE.

[11] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 356–370, 2012.

[12] Y. Isik, J. Le Roux, Z. Chen, S. Watanabe, and J. R. Hershey, "Single-channel multi-speaker separation using deep clustering," in *Proceedings of INTERSPEECH 2016*. ISCA, 2016.

[13] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, "Deep clustering: discriminative embeddings for segmentation and separation," in *Proceedings of ICASSP 2016*. IEEE, 2016.

[14] F. R. Stoter, S. Chakrabarty, B. Edler, and E. A. P. Habets, "Classification vs. regression in supervised learning for single channel speaker count estimation," in *Proceedings of ICASSP 2018*. 2018, IEEE.

[15] 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.

[16] E. Marchi, S. Shum, K. Hwang, S. Kajarekar, S. Sigtia, H. Richards, R. Haynes, Y. Kim, and J. Bridle, "Generalised discriminative transform via curriculum learning for speaker recognition," in *Proceedings of ICASSP 2018*. 2018, pp. 5324–5328, IEEE.

[17] J. A. Fiz, J. Morera, J. Abad, A. Belsunces, M. Haro, J. I. Fiz, R. Jane, P. Caminal, and D. Rodenstein, "Acoustic Analysis of Vowel Emission on Obstruvtive Sleep Apnea," *Chest*, vol. 104, no. 4, pp. 1093–1096, 1993.

[18] M. P. Robb, J. Yates, and E. J. Morgan, "Vocal Tract Resonance Characteristics of Adults with Obstructive Sleep Apnea," *Acta Oto-Laryngologica*, vol. 117, no. 5, pp. 760–763, 1997.

[19] R. Fernandez, J. L. Blanco, L. Hernandez, E. Lopez, J. Alcazar, and D. T. Toledano, "Assessment of Severe Apnoea through Voice Analysis, Automatic Speech, and Speaker Recognition Techniques," *EURASIP Journal on Advances in Signal Processing*, 2009.

[20] E. Goldshtein, A. Tarasiuk, and Y. Zigel, "Automatic Detection of Obstructive Sleep Apnea Using Speech Signals," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 5, pp. 1373–1382, 2011.

[21] M. C. Botelho, I. Trancoso, A. Abad, and T. Paiva, "Speech as a Biomarker for Obstructive Sleep Apnea Detection," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5851–5855.

[22] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker diarization with LSTM," in *Proceedings of ICASSP 2018*. 2018, IEEE.

[23] E. Kurtić, G. J. Brown, and B. Wells, "Resources for turn competition in overlapping talk," *Speech Communication*, vol. 55, no. 5, pp. 721–743, 2013.