SoftVAD in iVector-Based Acoustic Scene Classification for Robustness to Foreground Speech

Siyuan Song, Brecht Desplanques, Kris Demuynck, Nilesh Madhu
IDLab, Department of Electronics and Information Systems, Ghent University - imec, Belgium
Email: {Siyuan.Song, Brecht.Desplanques, Kris.Demuynck, Nilesh.Madhu}@ugent.be

***Abstract*—To increase the robustness of Acoustic Scene Classification (ASC) during foreground speech presence, we recently proposed a noise-floor based iVector framework exploiting the statistical *estimate* of the background signal spectrum. Thereby, ASC accuracy was greatly improved when foreground speech was predominant, at the cost of poorer performance in scenarios with low foreground speech levels. A soft Voice Activity Detector (softVAD) is introduced, here, to improve this trade-off. Three possibilities are investigated: (a) a segment-wise, weighted score fusion system, yielding a sofVAD-based weighted average of the output scores of the (classical) iVector framework and those of the noise-floor based iVector framework; (b) the introduction of weighted Baum-Welch statistics in the iVector extraction stage, with weights that emphasize the background-dominant frames and disregard speech-dominant frames in the test sequence. Based on the performance of these alternatives, a third approach (approach (c)) that performs segment-level score fusion of the frame-wise weighted statistics (approach (b)) and the noise-floor system is proposed. Experiments conclusively demonstrate that all proposals significantly improve the classification accuracy. Especially the last approach outperforms all other methods in a wide range of experimental conditions.**

***Index Terms*—Acoustic scene classification, iVector, softVAD, noise-floor estimation, foreground speech robustness**

I. Introduction

While most acoustic scene classification (ASC) systems perform well when only the ambient/background signal characterizing the acoustic scene is present, their performance degrades significantly when foreground speech is captured as well – as is typical in a real-life scenario. E.g., in telecommunications, or audio captured by hearables, it is very likely that the captured audio is dominated by the speech of the user. Practical integration of ASC into the audio processing chain of such devices, therefore, requires robustness to foreground speech. Research on this topic is still somewhat in its infancy.

A logical method to handle this condition would be to apply an operation inverse/complementary to speech *enhancement*, such that foreground speech is *removed*. There exists a wide range of speech enhancement approaches that can be adapted for this purpose (e.g. [1]–[3]). Alternatively, as recently proposed in [4], an explicit first stage is included in the ASC framework to *remove foreground speech*. A disadvantage of such methods is that they typically introduce (non-linear) artefacts in the resulting signal, which would affect the ASC system. Hence, our recent work [5] proposed to determine the acoustic scene on the basis of acoustic features extracted from the *noise-floor*, which is an estimation of the ambient signal spectrum. Along with multi-condition training (MCT), the iVector [6] ASC system utilizing noise-floor-based features significantly improved the classification accuracy in high speech-to-background ratio (SBR) scenarios. The trade-off was poorer performance in low SBR conditions. In this paper, we present different means to further improve this trade-off.

Three alternatives are investigated. First: a segment-level integration where the segment (utterance) level scores of the 'vanilla' iVector-based ASC system (acoustic features extracted directly from the microphone signal) and the noise-floor-based system [5] are combined in a weighted manner to determine the final score. The weights are obtained from the average foreground speech presence probability over the whole utterance. When speech presence is low, the scores of the vanilla system should have a weight close to 1 and that of the noise-floor based system close to 0. When foreground speech dominates, the weights are reversed, and more emphasis is given to the noise-floor based system. However, since speech signals are highly sparse, 'glimpses' of the background signal can be obtained during the speech pauses, allowing for a better exploitation of the background signal features. Thus, the second alternative incorporates *frame-level* foreground speech presence information directly into the vanilla iVector-based system at the factor analysis stage. This is done by computing weighted Baum-Welch statistics, where the extracted features are weighted according to their relevance for ASC. The higher the speech presence probability, the lower the weight allocated to the frame in the Baum-Welch statistics.

Experimental results demonstrate that this second system significantly improves the ASC performance in the presence of foreground speech, except at very high SBRs. To further improve performance, we propose a weighted score fusion, as described above, but of the modified Baum-Welch statistics system and a noise-floor based system that primarily operates on the disregarded speech frames. This last method thereby makes full use of the information in the whole segment, as compared to only focussing on the glimpses in speech pauses or relying entirely on the noise-floor based system. Consequently it yields the best performance over all considered SBRs.

This work is presented as follows. Section II summarizes the key aspects of the two baseline iVector-based ASC frameworks. Section III introduces the softVAD as a means to estimate the speech presence probability and presents the different ways of integrating it within the ASC framework. Sections IV and V describe the experimental setup and present and discuss the evaluation. Section VI concludes the paper.

II. Baseline iVector-based ASC frameworks

The acoustic features used in the iVector framework are the Mel-Frequency Cepstral Coefficients (MFCCs). For the vanilla system, these are extracted from short, overlapping frames of the *microphone* signal. Using factor analysis, the information in the variable length recording, described by a sequence of MFCC feature vectors, is compressed to a fixed length representation called the iVector. A regularized Gaussian backend classifier with class-specific covariance models is then used to process the estimated iVector and perform scene classification.

In the noise-floor-based iVector framework, the MFCCs are derived from an estimate of the background signal power spectral density. The latter may be obtained by any of several well-known statistical methods employed in single-microphone speech enhancement (e.g., [7], [8]), where this quantity is often termed the 'noise-floor'. The remainder of the framework is identical to the iVector framework mentioned above. Using the noise-floor based framework, significant classification accuracy gain is seen in most conditions with higher speech-to-background ratios (SBRs). However, when no speech is present, or the SBR is relatively low, the vanilla iVector framework achieves a better performance. For more details, the reader is referred to [5]. Note that neither of these baselines *differentiate* between non-speech and speech frames. We hypothesize that incorporating such distinction will yield a better performance trade-off over a wide range of SBRs.

III. Incorporating speech presence probability

Foreground speech presence probability is estimated using soft voice activity detection (softVAD). SoftVAD has been applied with success in fields such as speaker recognition, speaker segmentation and language recognition [9], [10]. Specifically, [11] demonstrates that a softVAD integrated into an iVector based speaker recognition system is better than a regular VAD that indicates speech active frames by means of a binary label. This idea was later incorporated in [12] for a speaker segmentation system, yielding a significant performance benefit. This is also the framework we adopt here.

A. SoftVAD for ASC

A softVAD assigns, to each signal frame t, a weight ($\in [0,1]$) that may be interpreted as a posterior measure of foreground speech dominance in that frame. In ASC, therefore, this measure may be used to *de-emphasize* frames dominated by foreground speech (and containing little to no background information) and emphasize frames with little or no speech. SoftVAD systems are typically implemented using GMM/HMM frameworks or neural networks [13]. We choose a GMM-based approach because the components of the GMM are also an integral part of the iVector system, allowing for a holistic incorporation of softVAD within the framework.

To implement the GMM-based softVAD, we require a *background signal* GMM θ_B and a *speech* GMM θ_S. These are trained on the same acoustic features $\mathbf{o}_t$ and development dataset used for training the ASC system. During the evaluation phase, the log-likelihoods (LLs) of the acoustic features $\mathbf{o}_t$ of each signal frame t are calculated using θ_B and θ_S. A softmax function transforms the LLs to *background* posteriors (i.e., dominance of background) $p(\theta_B|\mathbf{o}_t)$ [12]:

$$p(\theta_B|\mathbf{o}_t) = \frac{p_B e^{\rho \log p(\mathbf{o}_t|\theta_B)}}{p_B e^{\rho \log p(\mathbf{o}_t|\theta_B)} + p_S e^{\rho \log p(\mathbf{o}_t|\theta_S)}}\,, \tag{1}$$

where ρ is a hyperparameter that can be tuned on validation data and p_B and p_S are, respectively, the background and speech *priors*. We assume $p_S = p_B = 0.5$ in our work.

B. Weighted score fusion system using softVAD

For a given signal segment, denote by $\mathbf{LL}_{\text{v}}$ the M- dimensional *vector* of log-likelihood values for the M available acoustic scenes, obtained from the vanilla iVector-based ASC framework. Similarly, denote by $\mathbf{LL}_{\text{nf}}$ the log-likelihoods similarly obtained from the noise-floor-based iVector framework. A direct application of the softVAD would be to calculate a weighted average of these segment-level log-likelihoods. This fused log-likelihood vector $\mathbf{LL}_{\text{sf}}$ may be expressed as:

$$\mathbf{LL}_{\text{sf}} = \alpha \mathbf{LL}_{\text{v}} + (1-\alpha)\mathbf{LL}_{\text{nf}} \tag{2}$$

where the weight α is the average *background occupancy* or dominance in that segment. If N is the total number of frames in the segment, a straightforward computation of α is:

$$\alpha = \frac{1}{N}\sum_{t=0}^{N-1} p(\theta_B|\mathbf{o}_t) \tag{3}$$

The most likely acoustic scene then corresponds with the index of the maximum value within the fused LL vector $\mathbf{LL}_{\text{sf}}$. This system is denoted as SF_{VAD} in the following.

Exploiting weighted score fusion as in (2) can be expected to improve ASC performance over a wide range of SBRs. However, since softVAD provides a background posterior per frame and score fusion can only be implemented per segment, we next explore the possibility of a frame-wise integration.

C. Modified Baum-Welch statistics system using softVAD

Frame-wise integration of the softVAD information can be carried out directly within the factor analysis stage by modifying the estimation of the zeroth and first-order Baum-Welch statistics N^c and $\boldsymbol{f}^c$ for each component c of the background GMM. Thus, during iVector extraction, each frame is weighted by the background posterior [11], [12] and the weighted Baum-Welch statistics are obtained as:

$$\begin{aligned} N^c &= \sum_{t=0}^{N-1} p(\theta_B|\mathbf{o}_t)\gamma\big(\theta_{B,c}\big|\boldsymbol{o}_t\big) \\ \boldsymbol{f}^c &= \sum_{t=0}^{N-1} p(\theta_B|\mathbf{o}_t)\gamma\big(\theta_{B,c}\big|\boldsymbol{o}_t\big)\boldsymbol{o}_t\,, \end{aligned} \tag{4}$$

where $\gamma\big(\theta_{B,c}\big|\boldsymbol{o}_t\big)$ is the occupation probability of component c of the background GMM θ_B and $\boldsymbol{o}_t$ represents the feature vector for frame t. Given these modified Baum-Welch statistics, the iVector can then be extracted as in [14]. This system is denoted as BW_{mod} in the following.

D. Late fusion of BW$_{mod}$ and noise-floor based system

The BW$_{\text{mod}}$ system mainly extracts information relevant for ASC during the speech pauses, where the background signal is 'glimpsed'. In speech dominant frames, however, some background information is still present in the *noise floor*. Thus, we investigate a suitable integration of this information (from the noise-floor based system) alongside the BW$_{\text{mod}}$ system.

First, in order to extract relevant information in the speech dominant frames, so as to be truly complementary to the BW$_{\text{mod}}$ system of Section III-C, the noise-floor based system is modified. Chiefly, after extracting the acoustic features $\mathbf{o}_{t,\text{NF}}$ from the noise-floor and training the background GMMs $\theta_{B_{\text{NF}}}$, *complementarily weighted* Baum-Welch statistics are computed in the factor analysis stage:

$$\begin{aligned} N^c_{\text{NF}} &= \sum_{t=0}^{N-1} \Big(1 - p(\theta_B|\mathbf{o}_t)\Big)\gamma\big(\theta_{B_{\text{NF}},c}\big|\boldsymbol{o}_{t,\text{NF}}\big) \\ \boldsymbol{f}^c_{\text{NF}} &= \sum_{t=0}^{N-1} \Big(1 - p(\theta_B|\mathbf{o}_t)\Big)\gamma\big(\theta_{B_{\text{NF}},c}\big|\boldsymbol{o}_{t,\text{NF}}\big)\boldsymbol{o}_{t,\text{NF}}\,. \end{aligned} \tag{5}$$

The weighting by $\Big(1 - p(\theta_B|\mathbf{o}_t)\Big)$ emphasizes background information in the noise-floor during speech-dominant frames. Following this, iVector extraction is similar to Section III-C.

Next, the log-likelihoods of this modified noise-floor system and that of the BW$_{\text{mod}}$ system are fused according to (2). We term this last system SF$_{\text{BW}_{\text{mod}}\text{-NF}_{\text{mod}}}$.

E. Data selection for training the softVAD

The softVAD requires training a background GMM and a speech GMM. One simple way of choosing the training dataset is to use the background signals (without foreground speech) to train the background GMM and the clean foreground speech signals for the speech GMM. Note that silence frames are removed when training the speech GMM. Theoretically, models trained on such data can yield a good softVAD to obtain the background posterior during test. However we found it best, when training the speech GMM, to also include speech mixed with background signals at different SBR levels. See Section V for this analysis.

IV. Experimental setup

A. Dataset

The dataset consists of a background ASC dataset and a foreground speech dataset. The background dataset is taken from DCASE 2016 Task 1 dataset [15]. This includes approximately 10 hours of development data and 3 hours of evaluation data, distributed evenly over 15 different acoustic scenes. The development data contains 78 segments of every acoustic scene, each of 30 seconds duration. The evaluation data contains contains 26 segments of each acoustic scene, each again 30 seconds long. For the foreground speech 78 speakers are randomly selected from LibriSpeech dataset [16] for the development. For the *evaluation*, we choose 26 speakers from the PTDB-TUG pitch-tracking dataset [17] and the Multilingual Speech Dataset of NTT-AT[1]. Note, thus, that the foreground speech data is selected from completely different datasets in training and evaluation. The DCASE data is downsampled to 16 kHz mono for use with the speech data.

1) ASC datasets: To generate scenarios with foreground speech, 30 seconds of speech from a speaker is concatenated into one file. In the development dataset, 78 speakers are used as foreground speakers, with a new speaker being used for each background ASC signal segment. Note, however, that the same 78 speakers are used for all the 15 different scenes. The complete development dataset is obtained by merging the original, clean development ASC data with the mixed ASC data (at an SBR of -5 dB). This development data is used to train the noise-floor background GMM, the iVector extractors and the Gaussian backend classifiers.

The same method is applied to generate the foreground-mixed scenes for evaluation. Here, 26 different speakers are used. There is no overlap between the speaker set in development dataset and evaluation dataset. The speech is added on top of the background signal at the following SBRs: $\{-5\,\text{dB}, 0\,\text{dB}, 5\,\text{dB}, 10\,\text{dB}, 15\,\text{dB}, 20\,\text{dB}\}$.

2) softVAD datasets: As mentioned in Section III-E, ASC signals without foreground speech are used to train the background GMM. The training dataset for the speech GMM consists of pure speech signals as well as speech mixed with the background with an SBR of 10 dB. However, in both cases, the acoustic features are only extracted from frames where the speech has high energy. This is because speech is highly sparse and conversational speech has several pauses between words and sentences. Such frame selection, thus, avoids the leakage of frames containing silence (when using pure speech) or background-only (when using the mixed signals) into the speech GMM training. We perform a limited ablation study of this softVAD training dataset configuration in Section V-B.

B. System parameters

The iVector extraction system is identical to that in [5], with all hyper-parameters being determined by four fold cross-validation on the development dataset. Each GMM consists of 256 components and the rank of the iVector extractor T is 150. In the softVAD module, frames with an energy level above the median are considered as speech frames during training of the speech GMM. The energy level is always calculated on the clean speech signal and not the mixed one. The factor ρ in (1), used to calibrate the posteriors, is 2 (empirically chosen).

V. Results and discussion

A. Performance of the proposed speech-robust ASC systems

The benefit of the proposed improvements are evaluated by computing the classification accuracy at different SBRs. Five systems are compared: (a) the vanilla iVector framework, (b) the noise-floor based iVector framework, (c) SF$_{\text{VAD}}$, (d) BW$_{\text{mod}}$ and (e) SF$_{\text{BW}_{\text{mod}}\text{-NF}_{\text{mod}}}$. The first two systems are existing baselines against which we benchmark the improvements obtained by the proposed systems (c), (d) and (e).

[1] https://www.ntt-at.com/product/speech2002/

Figure 1 depicts the classification accuracy of each system for different SBRs. The accuracy of the baseline vanilla framework (**red**) strongly deteriorates with increasing SBR. At an SBR of 20 dB, where the foreground speech is very prominent, the accuracy drops to 43.2% (down from 80.3% in the absence of foreground speech). This degradation is significantly limited (flattened) when using the noise-floor based iVector framework (**dark blue**). However, the performance in the absence of foreground speech and in low SBR conditions is worse, since the statistical estimate of the noise-floor spectrum unavoidably removes information useful for ASC.

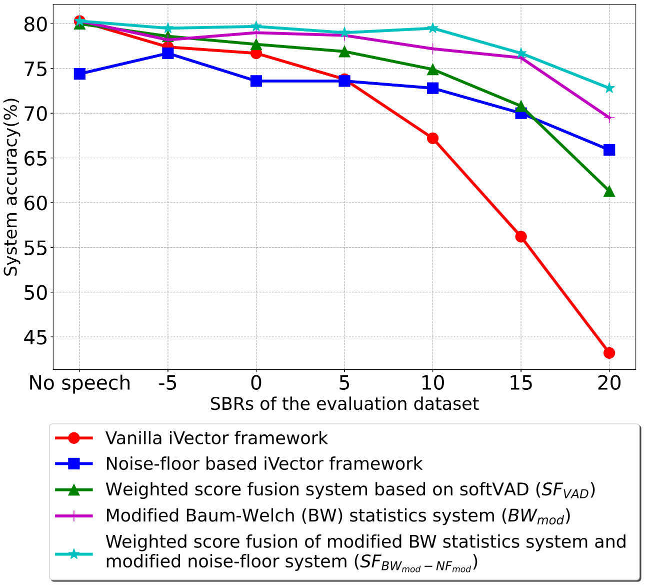

Fig. 1. Classification accuracy of the five systems for different SBRs.

The last three curves present the performance of the systems proposed in this paper. It may be seen that, for most conditions, $\mathrm{SF}_{\mathrm{VAD}}$ (**green**) improves with respect to the two baselines. However, at high SBRs, the accuracy drops to 61.3%, which is *worse* than the noise-floor based iVector framework. We hypothesise this is because, in such conditions, the vanilla iVector system scores are heavily impacted by foreground speech. Further, since we use a softVAD, the weights are rarely binary. Thus, a weighted combination of these scores with the noise-floor based system serves to degrade the final result.

We expected $\mathrm{SF}_{\mathrm{VAD}}$ to be better than $\mathrm{BW}_{\mathrm{mod}}$ (**purple**) at high SBRs and vice versa at low SBRs. Surprisingly, however, the results indicate that the second alternative consistently outperforms the first alternative and shows relatively consistent performance over all testing conditions. This is an interesting result as it indicates that, with conversational foreground speech, sufficient background signal information can be extracted during speech pauses to enable a good ASC.

To further analyze this implication, Table I presents the average background occupancy in the evaluation dataset, separately for each SBR condition. This indicates the average percentage of frames judged to be background dominant for that SBR condition. For this evaluation, a frame is considered to be background dominant if its background posterior $p(\theta_B|\mathbf{o}_t) \geq 0.3$. This decision threshold was chosen since it corresponds to the minimum in between the speech mode and background mode in the histogram plot of the frame-based background posteriors. We see that background occupancy *decreases* with increasing SBR and, in the condition of pure speech, background occupancy is 32.68%. This corresponds to the length of the typical silences present in conversational speech (indeed, speech enhancement literature typically assumes a speech *absence* probability of 0.2, see e.g., [18]).

TABLE I
AVERAGE BACKGROUND OCCUPANCY(%) IN THE EVALUATION DATASET OBTAINED BY SOFTVAD FOR DIFFERENT SBRS.

SBR (evaluation data)	Average background occupancy(%)
No speech	80.34
−5 dB	56.10
0 dB	48.29
5 dB	43.16
10 dB	40.56
15 dB	39.11
20 dB	37.96
Pure speech*	32.68

* Clean speech signals (Section IV-A), without background mixing.

The occupancy metric gives an indication of the amount of data effectively utilized by $\mathrm{BW}_{\mathrm{mod}}$ for ASC, thereby yielding a better understanding of how the accuracy is affected as a function of the amount of usable data. At an SBR of 20 dB, only 38% of the frames are effectively used in computing the Baum-Welch statistics. This implies that from the 30s length segments considered in our evaluation, only 12s of data is effectively used for ASC. To validate this conclusion, we consider an experiment where we take the vanilla iVector system and evaluate it on a 30-second and 10-second segments of the evaluation dataset, in the absence of speech. The classification accuracy is, respectively, 80.3% and 74.6% where the latter score is close to the 69.5% reported in Fig. 1.

Though $\mathrm{BW}_{\mathrm{mod}}$ extracts information relevant for ASC mainly from speech-absent frames, it still shows competitive performance. This can be improved if we can further leverage information gleaned during speech-active frames. This leads to the $\mathrm{SF}_{\mathrm{BW}_{\mathrm{mod}}\text{-}\mathrm{NF}_{\mathrm{mod}}}$ system where, by computing complementarily weighted Baum-Welch statistics in the noise-floor based iVector framework, we exploit information in frames which are not actively used in $\mathrm{BW}_{\mathrm{mod}}$ and combine this information using (weighted) score fusion. This system (**cyan** in Fig. 1) consistently outperforms all other systems in the tested conditions. This further indicates that complementary information is extracted from the noise-floor and from the $\mathrm{BW}_{\mathrm{mod}}$ system.

B. Training dataset configuration of the speech GMM

Lastly, we evaluate the benefit of different datasets used for training the speech GMM in the softVAD. Two sets are considered: a first set where the speech GMM is trained only on pure speech signals and our *default* second set where, *in addition* to pure speech signals, speech mixed with background at an SBR of 10 dB is also used. All other settings are identical to those described in Section V-A. Figure 2 shows the classification accuracy of $\mathrm{BW}_{\mathrm{mod}}$ for the speech GMM trained on these datasets. We observe that by augmenting the data used for training the speech GMM, the ASC performance is more steady over all testing SBRs. We believe that using multi-condition training for the speech GMM allows for a more reliable discrimination between background dominant and speech-dominant frames. Especially, we expect a better detection of weaker/lower energy speech frames in the higher SBR conditions. Such frames, if not detected as speech dominant, may negatively affect the ASC performance.

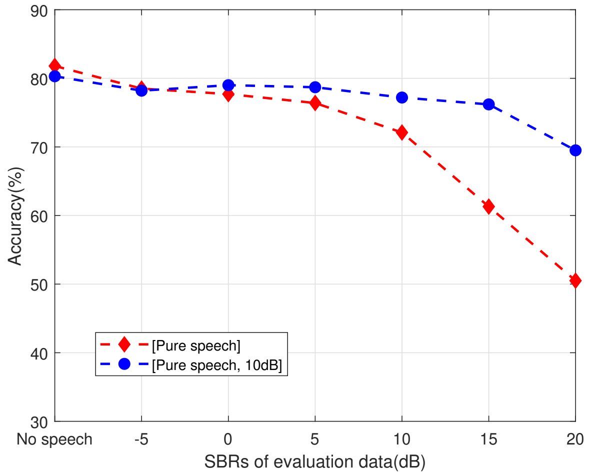

Fig. 2. Impact of different training datasets for the speech GMM in softVAD

VI. Conclusions

Our previous work [5], based on noise-floor features, improved the ASC performance for high SBR conditions, but at the cost of poorer performance in low SBRs. To improve this trade-off, we propose to incorporate speech presence information within the ASC framework. We have investigated three possibilities for this: a weighted score fusion system ($\mathrm{SF}_{\mathrm{VAD}}$), a modified Baum-Welch statistics system ($\mathrm{BW}_{\mathrm{mod}}$) and, lastly, late fusion of $\mathrm{BW}_{\mathrm{mod}}$ with a complementarily weighted, modified noise-floor based system, which we term $\mathrm{SF}_{\mathrm{BW}_{\mathrm{mod}}\text{-}\mathrm{NF}_{\mathrm{mod}}}$. Speech presence probability is estimated by a GMM-based softVAD, which can be holistically incorporated in our iVector framework.

Results demonstrate that $\mathrm{BW}_{\mathrm{mod}}$ already outperforms our previous system [5] as well as $\mathrm{SF}_{\mathrm{VAD}}$. This indicates that, in conversational foreground speech, sufficient background signal information can be extracted during speech pauses to enable a good scene classification. Thus, training a system to explicitly perform ASC in speech pauses may already provide sufficient foreground-speech robustness. However, the noise-floor based system stills yields information relevant to ASC in the speech-dominant frames and integrating this information can further improve the ASC system performance. This is evident from the superior performance of $\mathrm{SF}_{\mathrm{BW}_{\mathrm{mod}}\text{-}\mathrm{NF}_{\mathrm{mod}}}$ compared to all other approaches, over the entire range of tested SBRs.

Lastly, we note that the comparative analysis is limited to the baselines in our previous work. The only other paper we find on the topic is [4], which uses two separate DNNs in cascade, one for speech removal and the other for ASC. However it would be more efficient to train a single end-to-end DNN for ASC, implicitly robust to foreground speech. Our experiments with the modular and highly controllable iVector framework provide extra insights into this aspect. Based on our results, attentive statistics pooling [19] could help in steering DNN-based ASCs to focus on background 'glimpses'. Whether such frame-level attention weights can be solely based on self-attention or if these weights should be steered by a softVAD module is subject of future work.

References

[1] R. C. Hendriks, T. Gerkmann, and J. Jensen, *DFT-Domain based single-microphone noise reduction for speech enhancement*, ser. Synthesis Lect. Speech and Audio Proc. Morgan & Claypool, 2013.

[2] S. Elshamy and T. Fingscheidt, "DNN-based cepstral excitation manipulation for speech enhancement," *IEEE/ACM Trans. Audio, Speech, and Lang.*, vol. 27, no. 11, pp. 1803–1814, 2019.

[3] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Speech enhancement by LSTM-based noise suppression followed by CNN-based speech restoration," *EURASIP Journal on Advances in Signal Processing*, vol. 2020, no. 1, pp. 1–26, 2020.

[4] S. Liu, A. Triantafyllopoulos, Z. Ren, and B. W. Schuller, "Towards speech robustness for acoustic scene classification," *Proc. Interspeech 2020*, pp. 3087–3091, 2020.

[5] S. Song, B. Desplanques, C. De Moor, K. Demuynck, and N. Madhu, "Robust acoustic scene classification in the presence of active foreground speech," *European Signal Processing Conference (EUSIPCO)*, 2021.

[6] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Lang.*, vol. 19, no. 4, pp. 788–798, 2010.

[7] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech and Lang.*, vol. 20, no. 4, pp. 1383–1393, 2012.

[8] S. Rangachari and P. C. Loizou, "A noise-estimation algorithm for highly non-stationary environments," *Speech Communication*, vol. 48, no. 2, pp. 220–231, 2006.

[9] L. Ferrer, M. McLaren, N. Scheffer, Y. Lei, M. Graciarena, and V. Mitra, "A noise-robust system for NIST 2012 speaker recognition evaluation," SRI International, USA, Tech. Rep., 2013.

[10] L. Ferrer, M. McLaren, A. Lawson, and M. Graciarena, "Mitigating the effects of non-stationary unseen noises on language recognition performance," in *Proc. Interspeech*, 2015, pp. 3446–3450.

[11] M. McLaren, M. Graciarena, and Y. Lei, "Softsad: Integrated frame-based speech confidence for speaker recognition," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 4694–4698.

[12] B. Desplanques, K. Demuynck, and J.-P. Martens, "Soft VAD in factor analysis based speaker segmentation of broadcast news," in *Odyssey2016*, 2016, pp. 158–165.

[13] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program." in *Interspeech*, 2013, pp. 3497–3501.

[14] O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of i-vector extraction," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4516–4519.

[15] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in *24th European Signal Processing Conference (EUSIPCO)*, 2016, pp. 1128–1132.

[16] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an ASR corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[17] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, "A pitch tracking corpus with evaluation on multipitch tracking scenario," in *Proc. INTERSPEECH*, 2011, pp. 1509–1512.

[18] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 33, no. 2, pp. 443–445, 1985.

[19] K. Okabe, T. Koshinaka, and K. Shinoda, "Attentive Statistics Pooling for Deep Speaker Embedding," in *Interspeech*, 2018, pp. 2252–2256.