REAL-LIFE VOICE ACTIVITY DETECTION WITH LSTM RECURRENT NEURAL NETWORKS AND AN APPLICATION TO HOLLYWOOD MOVIES

Florian Eyben[1]*, Felix Weninger*[1]*, Stefano Squartini*[2]*, Björn Schuller*[1]

[1]Machine Intelligence & Signal Processing Group, MMK, Techische Universität München, GERMANY
[2]Department of Information Engineering, Università Politecnica delle Marche, Ancona, ITALY

ABSTRACT

A novel, data-driven approach to voice activity detection is presented. The approach is based on Long Short-Term Memory Recurrent Neural Networks trained on standard RASTA-PLP frontend features. To approximate real-life scenarios, large amounts of noisy speech instances are mixed by using both read and spontaneous speech from the TIMIT and Buckeye corpora, and adding real long term recordings of diverse noise types. The approach is evaluated on unseen synthetically mixed test data as well as a real-life test set consisting of four full-length Hollywood movies. A frame-wise Equal Error Rate (EER) of 33.2% is obtained for the four movies and an EER of 9.6% is obtained for the synthetic test data at a peak SNR of 0 dB, clearly outperforming three state-of-the-art reference algorithms under the same conditions.

Index Terms— Voice Activity Detection, Speech Detection, Neural Networks, Long Short-Term Memory

1. INTRODUCTION

Voice activity detection (VAD), also referred to as speech activity detection, is an important first step in many speech-based systems. It is important for Automatic Speech Recognition (ASR), to avoid word insertions due to noise and background speech; it is also used in audio coding to save bandwidth, and in multi-party conference systems, for example, to reduce the amount of background noise.

Early approaches to VAD were based on simple energy thresholds or pitch and zero-crossing rate rules (cf. [1]). These approaches perform well in settings where there is little or no background noise. More recent approaches consider more advanced parameters like autoregressive (AR) model parameters [2] and line spectral frequencies (LSPs). The most promising approaches in highly corrupted conditions seem to be data-driven methods, where a classifier is trained to predict speech vs. non-speech from acoustic features (cf., e. g., [3]). Still, the performance of such approaches degrades when background noise with spectral characteristics similar to speech is present. Very recent studies suggest that the use of long-span features clearly improves the robustness in real-life and noisy settings because the decision for each frame can be performed in the context of the previous frames [4].

In this light, we propose a novel approach, which uses traditional frame-wise features, but where the classifier is capable of learning the dynamics of the inputs and adaptively using previous inputs for the decision of the current frame. In Section 2 we present three state-of-the-art statistical VAD algorithms, some of which use context information in a rule-based fashion, which we will use as baselines in our evaluation. Next, in Section 3 we introduce our proposed approach. The data-sets used for evaluations are described in Section 4. We use both synthetic data of spontaneous and read speech in controlled noise conditions, and audio tracks of Hollywood movies containing highly non-stationary noise. Results are presented in Section 5; we conclude our findings in Section 6 and discuss our work in the context of prior work in Section 7.

2. STATE OF THE ART STATISTICAL VAD ALGORITHMS

In this Section the three baseline, state-of-the-art VAD algorithms [2, 5, 6] designed for the use in noisy conditions are briefly presented. They all belong to the category of *statistical methods*, where a *Likelihood Ratio* (LR) test is applied to the hypotheses of speech presence (denoted as H_1) and speech absence (H_0), on a certain frame of the observed noisy signal $x_t = s_t + n_t$, where s_t and n_t denote the clean speech and the noise signal, respectively.

2.1. Sohn's approach ($SOHN$)

The VAD proposed in [5] is based on a statistical model in the time-frequency domain for the derivation of the LR test. Given **S**, **N**, and **X** the DFT coefficient vectors of dimension L at the current frame m, for the speech, noise and noisy speech signals, respectively, the probability density functions conditioned on H_0 and H_1 are:

$$p(\mathbf{X}|H_0) = \prod_{k=0}^{L-1} \frac{1}{\pi\lambda_N(k)} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)}\right\} \quad (1)$$

$$p(\mathbf{X}|H_1) = \prod_{k=0}^{L-1} \frac{1}{\pi\left[\lambda_N(k)+\lambda_S(k)\right]} \exp\left\{-\frac{|X_k|^2}{\lambda_N(k)+\lambda_S(k)}\right\}$$

where $\lambda_N(k)$ and $\lambda_S(k)$ are the variances of N_k and S_k, respectively, i. e., the k-th terms of vectors **N** and **S**. The LR for the k-th frequency bin is given by:

$$\Lambda_k \triangleq \frac{p(X_k|H_1)}{p(X_k|H_0)} = \frac{1}{1+\xi_k} \exp\left\{\frac{\gamma_k\xi_k}{1+\xi_k}\right\} \quad (2)$$

where $\xi_k \triangleq \frac{\lambda_S(k)}{\lambda_N(k)}$ and $\gamma_k \triangleq \frac{|X_k|^2}{\lambda_N(k)}$ are the a-priori and a-posteriori signal-to-noise ratio (SNR) [7]. The decision rule is based on the log-LR function at the current frame m, which is obtained by averaging the log-likelihood ratios for each frequency bin, as follows:

$$\mathcal{L}(m) = \log\Lambda(m) = \frac{1}{L}\sum_{k=0}^{L-1} \log\Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (3)$$

where η is the decision threshold. A hangover mechanism, based on relationship between consecutive speech frames, is also implemented to reduce the false negative occurrences.

2.2. Ramirez' approach ($RAM05$)

The algorithm proposed in [6], as the one in [8], is based on the concept that more consecutive speech frames concur into the definition of the LR function. Given the M noisy observation DFT coefficient vectors $\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_M$ involved in the two-class classification problem, the Multiple Observation - LR (MO-LR) Test over a window of $2M + 1$ frames centered on frame m is the following (assuming that the vectors $\mathbf{X}_j$ are independent and taking the logarithm):

$$\mathcal{L}(m) = \log \Lambda^M(m) = \sum_{j=m-M}^{m+M} \log \frac{p(\mathbf{X}_j|H_1)}{p(\mathbf{X}_j|H_0)} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (4)$$

where m denotes the frame on which the decision is performed. The MO-LR function can be recursively calculated. The same statistical model explicited in (1) has been considered here.

2.3. AR-GARCH based approach (ARG)

The algorithm proposed in [2] is based on the idea of modelling the speech signal by means of an AR-GARCH (*autoregressive-generalized autoregressive conditional heteroskedasticity*) model in the time domain. These are the main steps of the algorithm:

- Estimation in the time domain of the noise variance σ_t by means of the IMCRA algorithm [9] and noisy signal normalization by this value;
- For each time instant t, estimation of, first, the AR-GARCH parameter vector θ, by means of a Recursive Maximum Likelihood (RML) updating rule, and then of the clean signal in the Minimum Mean Square sense, by exploiting the knowledge of the θ vector;
- VAD decision on a frame-by-frame basis, by using the estimated clean speech over all observation windows.

Focusing on this last step, the likelihood ratio $\Lambda(m)$ is obtained by averaging the LRs calculated for each time step. In order to take the correlation between adjacent signal samples into account, and to derive the final decision rule, the vocal activity is then modelled as a first-order Markov model, so that:

$$\mathcal{L}(m) = P_{m|m} = \frac{\Lambda_m P_{m|m-1}}{\Lambda_m P_{m|m-1} + (1 - P_{m|m-1})} \underset{H_0}{\overset{H_1}{\gtrless}} \eta \quad (5)$$

where $P_{m|m}$ and $P_{m|m-1}$ are the rules obtained by using and not-using the vocal activity information of the m-th frame, respectively.

3. PROPOSED LSTM-RNN VAD

In this paper we present a novel data-driven method for voice activity detection based on (unidirectional) Long Short-Term Memory Recurrent Neural Networks (LSTM-RNN) [10]. The motivation behind the use of LSTM-RNN is their ability to model long range dependencies between the inputs. Other common data-driven VAD approaches, such as those based on GMM (cf. section 7) or Feed-Forward Neural Networks do not consider any temporal dependencies in the model. Delta features, modulation or long-span features [4] are used to overcome these issues. Standard Recurrent Neural Networks (as used in [11] for VAD), are able to model a limited amount of temporal dependency. However, LSTM-RNN go beyond simply using context information by introducing the concept of a memory cell that can be read, written and reset depending on feature context and previous outputs, by means of multiplicative input, output and forget units whose multiplicative weights are learned automatically during training. Thereby, they learn when to access which parts of past context, solving the vanishing gradient problem of traditional RNNs [12].

The networks we use for VAD have an input layer which matches the size of the low-level acoustic feature vectors, one or more hidden layers, and an output layer with a single linear unit. The networks are trained as regressors to output a voicing score for every frame in the range [-1; +1]; +1 indicating voicing, -1 indicating silence or noise. Two neural network topologies are investigated:

- $N1$: 1 recurrent hidden layer (4 blocks with 50 LSTM cells each)
- $N2$: 3 recurrent hidden layers (50 LSTM cells in one block; 10 sigmoid neurons; 20 LSTM cells)

On the input side of the networks we use a standard RASTA-PLP [13] frontend with cepstral coefficients 1–18 and their first order delta coefficients. The frame size is 25 ms and the frame step is 10 ms. It is important to highlight that the 36 dimensional feature vector does not contain an energy coefficient (e. g., 0-th cepstral coefficient). We decided for this to make the networks input level invariant. Features were extracted with our openSMILE toolkit [14] and z-normalization was applied to all features (mean zero, variance 1). The means and variances for the z-normalization are computed from the training set only. The LSTM-RNNs were trained and evaluated with the rnnlib by Alex Graves [15].

Training is performed with the backpropagation through time (BPTT) algorithm; the weights are updated using the gradient descent algorithm with a learning rate of 10^{-5} and momentum 0.9. This requires weights to be initialized with non-zero values, thus we initialize the weights with uniform random values sampled from $]0; 0.1]$. To increase robustness against convergence into a suboptimal local minimum of the weight space, we train three networks with different random weight initialisations. Network predictions for the test and validation set are then computed by averaging the predictions for all three networks. To further enhance generalisation, we added Gaussian noise with zero mean and standard deviation of 0.3 to all inputs. To avoid over-adaptation, a maximum of 40 training epochs was run. Further, training was stopped early if there was no error improvement over 10 epochs. The frame-wise root mean quadratic error between the targets and the network predictions is used as evaluation criterion during network training.

The computational complexity for evaluating the networks is linear with respect to the number of input frames. For every frame a constant number of operations needs to be performed. Many computations can be run in parallel, which is ideal for implementation on embedded hardware such as Digital Signal Processors (DSPs) or Field Programmable Gate Arrays (FPGAs). The asymptotically quadratic complexity wrt. the network size can be drastically reduced in practice by the chosen block structure of the hidden layers.

4. DATA SETS

To obtain a large amount of labelled and diverse speech data for training and validating the networks, we synthesise data by building random utterance sequences overlaid with additive noise. The speech data is taken from the Buckeye [16] and the TIMIT corpus [17]. The Buckeye corpus consists of 26 h of spontaneous speech from 40 speakers (20 male, 20 female) recorded in informal interview situations. Only the subjects' speech is used, and speaker turns corresponding to utterances between silence segments of at least 0.5 s are extracted according to the automatic alignment delivered with the Buckeye corpus. The corpus is split subject-independently into a

training, validation and test set, respecting stratification by age and gender. The segmentation and subdivision is exactly equal to the one used in [18]. The original TIMIT training set is split speaker-independently into a training and validation set. Speech for the VAD test set is taken from the original TIMIT and Buckeye test sets. Four types of noise are used: $babble$, $city$, white and pink $noise$, and $music$. The babble noise recordings are taken from the *freesound.org* website. Samples from the categories pub-noise, restaurant chatter, and crowd noise are joint. The music recordings are instrumental and classical music pieces from the *last.fm* website. The city recordings were recorded at TUM in Munich, Germany with smartphones while people where cycling and walking through the city. White and pink noise samples were generated with pseudo random number generators and a bandpass filter.

The noise samples used for synthesising the VAD training, validation, and test samples are fully disjunctive (i.e., different pieces of music, different babble samples, etc.). Noise samples for the test and validation sets are 30 minutes each, the remaining noise audio is used for the training set. The lengths of these samples varies from 94 minutes (babble) to 176 minutes (music).

Each synthetic utterances in the VAD training set is composed of $N \in \{1, \ldots, 5\}$ original speech utterances, which are randomly selected either from TIMIT or Buckeye. A pause before the first utterance, pauses between all utterances, and a pause after the last utterance are inserted with a uniformly random length of 0.5 to 5 seconds. Each of the original utterances is normalised to have a peak amplitude of 0 dB and then all N normalised utterances are multiplied with a uniformly random gain factor $g_{s,lin} = 10^{\frac{g_s}{20.0}}$ where $g_s \in [+3\,dB; -20\,dB]$. For 80 % of the synthetic utterances, a random noise sample, which matches the total length of the N speech utterances and the $N + 2$ pauses, is selected from the training noise pool and normalised to a peak amplitude of 0 dB. A multiplicative gain $g_{n,lin}$ according to equation (6) is applied to the noise segment:

$$g_{n,lin} = 10^{(log(g_{s,lin}) - \frac{SNR}{20.0})} \tag{6}$$

The SNR is randomly chosen for each mixed instance as $SNR \in [-6dB; +25dB]$. The remaining 20% of all synthetic utterances are not overlaid with noise. 1 948 instances are created with speech from Buckeye. This corresponds to 15 h of total audio, where 6:43 h are non-speech and 8:17 h are speech. From TIMIT speech, 3 493 instances are generated. This corresponds to 19:45 h of total audio, where 12:54 h are non-speech and 6:51 h are speech. In total there is 34:54 h of audio in the VAD training set.

The validation set is built in a similar way, however one single mixed instance each with a total length of 22.5 minutes is generated from Buckeye speech and TIMIT speech. The gain of each of the original utterances is varied randomly over the same range as for the training set, and pauses are added using the same parameters. This same sequence of speech utterances and pauses is overlaid with four continuous 30 minute segment of babble, music, city, and white+pink noise (all normalised to 0 dB peak amplitude). A fixed gain $g_{n,lin}$ is chosen for this noise segment as $g_{n,lin} = 0.5(g^{\mu}_{s,lin} + g^{min}_{s,lin})$, where $g^{\mu}_{s,lin}$ and $g^{min}_{s,lin}$ are the average and minimum multiplicative gain factors of the speech utterances. In total, the VAD validation set has 3 h of audio, where 1:22 h are speech and 1:38 h are non-speech.

For the VAD test set, 15 minute long mixed instances are created each from TIMIT and Buckeye speech. Thus, the total length of each test instance is 30 minutes. The clean version of the 30 minute test audio contains 12 minutes of speech and 18 minutes of silence. A single fixed gain of -6 dB for the clean speech is applied and noise is added with a peak SNR (noise gain relative to speech gain) of 0 dB.

Table 1: Frame-level results for validation and test set of nets $N1$ and $N2$ and the $RAM05$, ARG, and $SOHN$ algorithms. Area under ROC curve (AUC), Equal Error Rate (EER), and combined error rate (false negative rate (FNR) + false positive rate (FPR)) computed with a threshold estimated from the validation set. Test set: -6 dB gain applied to original speech signal, average SNR is 0 dB.

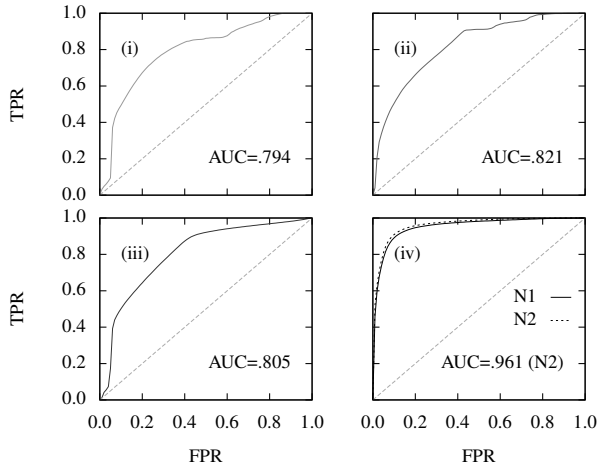
	AUC				
set	$N1$	$N2$	$RAM05$	ARG	$SOHN$
validation	.814	**.838**	.713	.685	.709
test clean	.980	**.985**	.955	.962	.959
test babble	.909	**.932**	.877	.875	.826
test music	.921	**.940**	.725	.675	.677
test city	.968	**.972**	.928	.935	.931
test noise	.941	**.949**	.878	.773	.878
test ALL	.951	**.961**	.821	.794	.805
[%]	**FNR + FPR**				
validation	53.69	**52.33**	72.19	67.95	68.77
test clean	12.99	**11.65**	14.64	13.67	13.52
test babble	36.73	**26.21**	66.55	59.23	76.01
test music	31.03	**27.13**	79.96	88.94	82.21
test city	17.61	**16.63**	28.11	31.48	29.67
test noise	23.26	**23.14**	66.75	64.45	61.31
test ALL	24.18	**20.95**	52.82	51.54	53.22
[%]	**EER**				
ALL	10.41	9.55	26.58	25.85	26.99

To test the VAD in challenging real-life conditions, we use a second test set consisting of the full English audio tracks of four Hollywood movie DVDs. The choice of these movies was inspired by the official development set of the 2012 MediaEval campaign's violence detection task [19]. Speech and non-speech segments in those four movies were manually annotated. The list of movies and statistics on speech / non-speech segments are found in Table 2.

5. RESULTS

Results for the synthetic test and validation sets are given in Tab. 1. We use two evaluation metrics: the area under receiver operating characteristic (ROC) curves (AUC) and the combined error rate (false positive rate (FPR) + false negative rate (FNR)). Fixed thresholds which correspond to the thresholds at the Equal Error Rate (EER) on the validation set are used for all test set evaluations to ensure the test data is fully unknown to the system. For nets $N1$ and $N2$ the thresholds are -0.268 and -0.071, respectively. The same thresholds are used for the DVD movie test set. For FPR and FNR computation, the thresholded predictions (both for reference and LSTM) are smoothed with a silence hysteresis of 5 frames (i. e., non-speech segments shorter than 5 frames are joined with the adjacent speech segments). We observe that both the $N1$ and $N2$ network topologies outperform all baseline algorithms in terms of both AUC and FNR + FPR, and notably also for clean speech. The largest margin of improvement is found for music, babble, white and pink noise. On city noise, the baselines are relatively robust, which can be attributed to the fact that the average energy of these noise samples is much lower than the peak amplitude (e. g., loud cars passing by). The ROC curves for the proposed and the baseline algorithms are shown in Figure 1. The 'smoothness' of the curves for the proposed approach compared

Fig. 1: Receiver operating characteristic (ROC) curves for VAD on synthetic test set: True-positive-ratio (TPR) vs. false-positive-ratio (FPR) and area under curve (AUC) for AR-GARCH [2] (i), Ramirez' approach [6] (ii), Sohn's approach [5] (iii) and the proposed LSTM-RNN approach (iv) using network topologies N1 and N2.



to the baselines is due to the modeling as a regression task in training, which delivers a 'continuum' of scores in testing. As to ROC, the behaviour of the two network topologies is nearly identical. The EER across validation and test sets is around 10 % for both network topologies as opposed to 25 % and above for the baseline algorithms.

The results for the movie test set are given in Tab. 3. Compared to the synthetic test set, the performance of our method and $SOHN$ on the movie test set is much lower. However, the networks still clearly outperfom $SOHN$. Note, that the results for $RAM05$ and ARG could not be obtained due to the high computational complexity of these algorithms, but given the test set results we estimate their performance to be similar to $SOHN$. One main reason for the reduced performance on the movie set might be that many noise types occur that have not been seen in training, such as gunshots, fighting, etc., and noises that are easy to confuse with speech, such as animal sounds. Another reason is the coarse annotation style of speech segments; for the sake of efficiency, longer conversations were labelled as continuous speech segments, even though they included small pauses. In the evaluations this results in a higher miss rate than is actually given. Compared to [3] (25.3% EER on YouTube videos) our EERs are very competitive, considering that their system was trained on in-domain data. Next, a comparison of both approaches on YouTube videos and Hollywood movies would be highly interesting.

6. CONCLUSION AND OUTLOOK

In this paper we have presented a novel VAD approach based on LSTM-RNN. We further presented a method for synthesising training data for the LSTM-RNN to approximate real-life settings without the need for in-domain data. We demonstrated the feasibility of this approach on real-life noisy speech data from Hollywood movies, and we showed that LSTM-RNN outperforms all three statistical VAD baselines. This is all the more notable since our method does not require future context, unlike the RAM05 and SOHN methods.

Future work will investigate the performance of LSTM-RNNs in more detail, analysing the context learning behaviour in comparison to GMMs, MLPs and RNNs using time-frequency or modulation spectrum features. Furthermore, using semi-supervised and active learning to efficiently adapt the generic models presented in this paper to specific use cases such as movies or web videos will be attempted.

Table 2: Movie test set. Movie length and percentage of parts with speech; min., mean, max. duration of continuous speech segments.

Title	[hh:mm]	% sp.	min/mean/max [s]
I Am Legend	1:36	39.2	0.5/21.4/174.9
Kill Bill Vol. 1	1:46	33.9	0.4/39.3/321.2
Saving Private Ryan	2:42	48.6	0.5/25.2/230.4
The Bourne Identity	1:53	40.7	0.6/32.6/185.6

Table 3: Frame-wise results for the movie test set of nets $N1$ and $N2$ and the $SOHN$ algorithm. Area under ROC curve (AUC), Equal Error Rate (EER), and combined error rate (false negative rate (FNR) + false positive rate (FPR)) computed with a threshold estimated from the validation set. Results for RAM05 and ARG are not included due to their heavy computational load on the large DVD test set.

	AUC		
movie	$N1$	$N2$	$SOHN$
I Am Legend	.704	.676	.567
Kill Bill 1	.627	.601	.554
Saving P.	.743	.680	.577
Bourne Id.	.685	.647	.603
ALL	.722	.676	.556
[%]	**FNR + FPR**		
I Am Legend	76.65	75.57	94.90
Kill Bill 1	94.14	94.41	102.88
Saving P.	67.03	81.70	92.46
Bourne Id.	70.83	80.10	90.95
ALL	69.87	78.03	95.52
[%]	**EER**		
ALL	33.18	36.76	45.73

7. RELATION TO PRIOR WORK

Many previous approaches to VAD rely on Gaussian mixture modeling and adaptation as typical for ASR, to adapt the VAD models to speakers [20] and background noise [21–23], in contrast to the proposed discriminative approach. [24] adapts GMMs to channel and noise conditions. [25] proposes to couple VAD with the acoustic models in the recogniser, whereas the proposed approach does not rely on phonetic modeling. Use of temporal context in data-based approaches has been proposed, e. g., by [26] who use PLP based and similar, more advanced temporal features combined with GMMs. [4] compares a standard GMM system using 14 PLP cepstral coefficients with a Multi-Layer Perceptron (MLP) based system using Long-Span acoustic features computed over .5 seconds windows. MLP based speech/non-speech posteriors are then decoded with two ergodic Hidden Markov Models (HMMs). However, these systems do not use adaptive context learning as by LSTM-RNN. [3] compares GMM with a discriminative classifier and proposes novel features instead of standard MFCC/PLP frontends. Real noisy, manually labelled YouTube videos are used for evaluation, but only in-domain training is considered.

8. REFERENCES

[1] K. Woo, T. Yang, K. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *IET Electronics Letters*, 2000.

[2] S. Mousazadeh and I. Cohen, "AR-GARCH in Presence of Noise: Parameter Estimation and Its Application to Voice Activity Detection," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 4, pp. 916–926, 2011.

[3] A. Misra, "Speech/nonspeech segmentation in web videos," in *Proc. of INTERSPEECH 2012, Portland, USA*. September 2012, ISCA.

[4] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesel, and P. Matjka, "Developing a speech activity detection system for the darpa rats program," in *Proc. of INTERSPEECH 2012, Portland, USA*. September 2012, ISCA.

[5] J. Sohn and N. Kim, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, 1999.

[6] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[7] I. Cohen, "On the decision-directed estimation approach of Ephraim and Malah," in *Proc. of ICASSP*. IEEE, 2004, vol. I, pp. 1–293.

[8] J. Ramirez, J. Segura, M. Benitez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Communication*, vol. 42, no. 3, pp. 271–287, 2004.

[9] I. Cohen, "Noise spectrum estimation in adverse environment: Improved minima controlled recursive averaging," *IEEE Trans. Audio Speech Processing*, vol. 11, no. 5, pp. 466–475, 2003.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9(8), pp. 1735–1780, 1997.

[11] R. Gemello, F. Mana, and R.D. Mori, "Non-linear esimation of voice activity to improve automatic recognition of noisy speech," in *Proc. of INTERSPEECH 2005, Lisbon, Portugal*. September 2005, pp. 2617–2620, ISCA.

[12] S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber, "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies," in *A Field Guide to Dynamical Recurrent Neural Networks.*, S.C. Kremer and J.F. Kolen, Eds. 2001, IEEE Press.

[13] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, Apr. 1990.

[14] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE – the munich versatile and fast open-source audio feature extractor," in *Proc. ACM Multimedia (MM), Florence, Italy*. 2010, pp. 1459–1462, ACM.

[15] A. Graves, S. Fernández, and J. Schmidhuber, "Multidimensional recurrent neural networks," in *Proc. of the 2007 International Conference on Artificial Neural Networks, Porto, Portugal*, September 2007.

[16] M.A. Pitt, L. Dilley, K. Johnson, S. Kiesling, W. Raymond, E. Hume, and E. Fosler-Lussier, *Buckeye Corpus of Conversational Speech (2nd release)*, Department of Psychology, Ohio State University (Distributor), Columbus, OH, USA, 2007, [www.buckeyecorpus.osu.edu].

[17] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, N.L. Dahlgrena, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," 1993.

[18] F. Weninger, B. Schuller, M. Wöllmer, and G. Rigoll, "Localization of non-linguistic events in spontaneous speech by non-negative matrix factorization and Long Short-Term Memory," in *Proc. of ICASSP*, Prague, Czech Republic, 2011, pp. 5840–5843.

[19] C.H. Demarty, C. Penet, G. Gravier, and M. Soleymani, "The MediaEval 2012 Affect Task: Violent scenes detection in Hollywood Movies," in *Proc. of MediaEval 2012 Workshop*, Pisa, Italy, 2012.

[20] S. Matsuda, N. Ito, K. Tsujino, H. Kashioka, and S. Sagayama, "Speaker-dependent voice activity detection robust to background speech noise," in *Proc. of INTERSPEECH 2012, Portland, USA*. September 2012, ISCA.

[21] S. Deng, J. Han, T. Zheng, and G. Zheng, "A modified MAP criterion based on hidden Markov model for voice activity detecion," in *Proc. of ICASSP*. may 2011, pp. 5220–5223, IEEE.

[22] Y. Suh and H. Kim, "Multiple acoustic model-based discriminative likelihood ratio weighting for voice activity detection," *Signal Processing Letters*, vol. 19, no. 8, pp. 507–510, aug 2012.

[23] M. Fujimoto, S. Watanabe, and T. Nakatani, "Frame-wise model re-estimation method based on gaussian pruning with weight normalization for noise robust voice activity detection," *Speech Communication*, vol. 54, no. 2, pp. 229–244, 2012.

[24] M.K. Omar, "Speech activity detection for noisy data using adaptation techniques," in *Proc. of INTERSPEECH 2012, Portland, USA*. September 2012, ISCA.

[25] K. Thambiratnam, W. Zhu, and F. Seide, "Voice activity detection using speech recognizer feedback," in *Proc. of INTERSPEECH 2012, Portland, USA*. September 2012, ISCA.

[26] S. Thomas, S.H. Mallidi, T. Janu, H. Hermansky, N. Mesgarani, X. Zhou, S. Shamma, T. Ng, B. Zhang, L. Nguyen, and S. Matsoukas, "Acoustic and data-driven features for robust speech activity detection," in *Proc. of INTERSPEECH 2012, Portland, USA*. September 2012, ISCA.