DNN-based Causal Voice Activity Detector

Ivan J. Tashev
Microsoft Research
One Microsoft Way, Redmond,
WA 98051, USA
ivantash@microsoft.com

Seyedmahdad Mirsamadi
University of Texas at Dallas
800 West Campbell Road, Richardson,
TX 75080, USA
mirsamadi@utdallas.edu

***Abstract*—Voice Activity Detectors (VAD) are important components in audio processing algorithms. In general, VADs are two way classifiers, flagging the audio frames where we have voice activity. Most of them are based on the signal energy and build statistical models of the noise background and the speech signal. In the process of derivation, we are limited to simplified statistical models and this limits the accuracy of the classification. Using more precise, but also more complex, statistical models makes the analytical derivation of the solution practically impossible. In this paper, we propose using deep neural network (DNN) to learn the relationship between the noisy speech features and the correct VAD decision. In most of the cases we need a causal algorithm, i.e. working in real time and using only current and past audio samples. This is why we use audio segments that consist only of current and previous audio frames, thus making possible real-time implementations. The proposed algorithm and DNN structure exceeds the classic, statistical model based VAD for both seen and unseen noises.**

***Index Terms*—voice activity detection, deep neural networks, speech statistical model, noise statistical model.**

I. Introduction

Voice Activity Detectors (VAD) are algorithms for detecting the presence of speech signal in the mixture of speech and noise. They are part of noise suppressors, double talk detectors, codecs, and automatic gain control blocks, to mention a few. The VAD output can vary from simple binary decision (yes/no), to soft decision (probability of speech presence in the current audio frame), to probability of speech presence in each frequency bin of each audio frame. The commonly used VAD algorithms are based on the assumption of quasi-stationary noise, i.e. the noise spectrum changes much slower than the speech signal. A classic VAD algorithm works in real time and makes the decisions based on the current and previous samples, i.e. it is causal. Most of these algorithms work in frequency domain for better integration in the audio processing chain and provide estimation for each frequency bin separately. One of the approaches frequently used as a baseline VAD algorithm is standardized as ITU-T Recommendation G.729-Annex B [1]. An improved and generalized VAD is described in [2], where authors create a soft decision VAD assuming Gaussian distribution of the noise and speech signals. A simple HMM is added to create a hangover scheme in [3] and to finalize the decision utilizing the timing of switching the states. This algorithm can be generalized and optimized for better performance as described in [4].

Most of the VAD algorithms assume Gaussian distribution of the noise and speech signals. It is well known that while the distribution of noise amplitudes in time domain is well modelled with the Gaussian distribution, the distribution of the amplitudes of the speech signal has higher kurtosis than the Gaussian distribution. Gazor and Zhang [5] published a study for the speech signal distribution in time domain, later in [6] this study was extended with models of the Probability Density Functions (PDF) of the speech signal magnitudes in frequency domain. Several attempts are published in the literature to utilize the non-Gaussianity of the speech signal for better noise suppression rules [7], [8] and [9], or for better VAD [10] and [11]. In most of the cases it is very difficult to find analytical form of the suppression rules, or speech presence probability, and the proposed solutions are either approximate or computationally expensive.

The statistical audio signal processing also assumes that the frequency bins in one audio frame are statistically independent, which allows processing these bins individually. The same assumption is in force for the consecutive audio frames, which allows processing of the audio signal frame–by–frame. In reality there are noise signals that change faster than the speech signal (clapping, clanks, etc.), the consecutive audio frames are highly correlated, and the frequency bins in the same frame contain information that can be utilized by processing them together. Still, the assumptions above led to working VAD algorithms, which serve well in pretty much every audio processing system.

In this paper we propose an algorithm for causal VAD based on deep neural networks (DNN). The DNN is trained on segments of several consecutive audio frames, and with all frequency bins together to utilize the correlation between the frames and bins. We do not assume any prior distribution of the noise and speech signals and expect the DNN to learn the dependency between the input features and the VAD decision. In Section II, we formulate the problem and present the statistical model-based VAD. Sections III and IV describe the proposed neural network structure and the evaluation dataset. In Section V, we describe the experimental results and we conclude in VI.

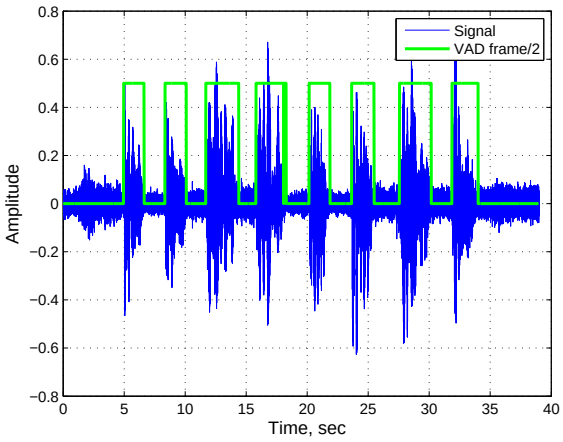

Fig. 1. Noisy signal in time domain with SNR=10 dB.

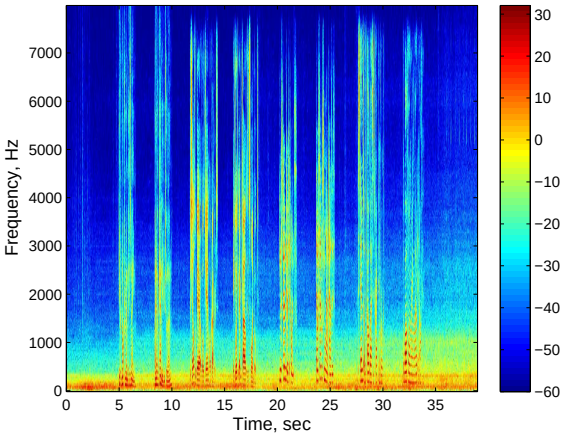

Fig. 2. Noisy signal in frequency domain with SNR=10 dB.

II. Problem definition

A. Modelling

We have a limited discrete signal in time domain $x(lT)$ where $l \in [0, L-1]$, $x_{\min} \leq x(lT) \leq x_{\max}, \forall l$, and T is the sampling period. An example of such signal is shown in Figure 1. This signal is a mixture of two limited, discrete, and uncorrelated signals $x(lT) = n(lT) + s(lT)$, noise $n(lT)$ and speech $s(lT)$, respectively. After framing, windowing, and converting to frequency domain we have $X_k^{(n)} = N_k^{(n)} + S_k^{(n)}$, where $k \in [0, K-1]$ is the frequency bin, K is the number of frequency bins, $n \in [0, N-1]$ is the frame number, and N is the number of audio frames. The same can be written in matrix form $\mathbf{X} = \mathbf{N} + \mathbf{S}$, where all are $K \times N$ complex matrices representing the spectra of the signal, the noise, and the speech components. This representation is visualized in Figure 2, where the magnitudes are in decibel scale.

In each frame and/or frequency bin we consider two hypotheses:

$$H_0\text{: speech is absent, } \mathbf{X} = \mathbf{N}$$
$$H_1\text{: speech is present, } \mathbf{X} = \mathbf{N} + \mathbf{S}.$$

The goal of the VAD algorithm is to produce the presence probability $P_k^{(n)}(H_1)$ for each frequency bin, and $P^{(n)}(H_1)$ for each frame (column in the above matrices above). An example of the expected VAD decision per frame is shown in Figure 1.

B. Voice Activity Detectors

Let us assume that noise and speech signals are zero mean and fully characterized by their respective variances σ_n^2 and σ_s^2, and we have a prior knowledge of the PDFs of these two signals, $p_n\left(a\left|\sigma_n^2\right.\right)$ and $p_s\left(a\left|\sigma_s^2\right.\right)$ respectively. The PDF of a mix of two uncorrelated signals is the convolution of the PDFs of the two signals:

$$p_x\left(a\left|\sigma_n^2, \sigma_s^2\right.\right) = p_n\left(a\left|\sigma_n^2\right.\right) * p_s\left(a\left|\sigma_s^2\right.\right). \tag{1}$$

Note that this equation has analytical solution for a small number of distribution pairs, it has to be solved numerically for most of the cases.

The probability $P(H_1|a)$ of signal with amplitude a to contain speech is derived after directly applying the Bayesian rule:

$$P(H_1|a) = \frac{p(a|H_1)P(H_1)}{p(a|H_1)P(H_1) + p(a|H_0)P(H_0)}. \tag{2}$$

Here $P(H_1)$ and $P(H_0) = 1 - P(H_1)$ are the prior probabilities for speech and noise presence respectively. After dividing by $p(a|H_0)P(H_0)$ we have:

$$P(H_1|a) = \frac{\varepsilon\Lambda}{1+\varepsilon\Lambda}, \tag{3}$$

where $\varepsilon = P(H_1)/P(H_0)$, and Λ is the likelihood ratio:

$$\Lambda = \frac{p_x\left(a\left|\sigma_n^2, \sigma_s^2\right.\right)}{p_n\left(a\left|\sigma_n^2\right.\right)}. \tag{4}$$

The proportion of the prior probabilities for speech and noise ε can be assumed constant and known. Then if we can estimate the noise and speech variances - we can estimate the speech presence probability in each frame and/or frequency bin.

Under the assumption of zero mean Gaussian distribution of both speech and noise signals, [3] provides analytical solution of (4) for the likelihood for speech signal presence in frequency bin k of audio frame n:

$$\Lambda_k = \frac{1}{1+\xi_k}\exp\left(\frac{\gamma_k\xi_k}{1+\xi_k}\right), \tag{5}$$

where $\xi_k = \frac{\lambda_S(k)}{\lambda_N(k)}$ and $\gamma_k = \frac{|X_k|^2}{\lambda_N(k)}$ are the prior and posterior SNRs respectively, $\lambda_S(k) = \sigma_s^2(k)$ and $\lambda_N(k) = \sigma_n^2(k)$. The decision directed approach [12] is used for estimation of the prior SNR:

$$\xi_k^{(n)} = \alpha\frac{\tilde{\lambda}_S^{(n-1)}}{\lambda_N^{(n-1)}} + (1-\alpha)\cdot\max\left[0, \left(\gamma_k^{(n)} - 1\right)\right]. \tag{6}$$

Note that this approach utilises partially the fact that the consecutive speech frames are correlated. Here α is a constant typically in the range of $0.95 - 0.98$.

In [3] is also proposed smoothing of the estimated likelihood using a first order HMM. After the derivation the smoothed likelihood for speech presence in the current frequency bin is estimated as:

$$\tilde{\Lambda}_k^{(n)} = \frac{a_{01} + a_{11}\tilde{\Lambda}_k^{(n-1)}}{a_{00} + a_{10}\tilde{\Lambda}_k^{(n-1)}}\Lambda_k^{(n)}, \tag{7}$$

where a_{01} and a_{10} are the prior probabilities for changing the state. Then the probability for speech presence in the current frequency bin is:

$$P_k^{(n)}\left(H_1|X_k^{(n)}\right) = \frac{\tilde{\Lambda}_k^{(n)}}{1 + \tilde{\Lambda}_k^{(n)}}. \tag{8}$$

Note that ε from equation (3) conveniently cancels out.

To combine the likelihoods from all frequency bins to compute the likelihood for speech signal presence in the entire frame we can use geometric mean or arithmetic mean. The geometric mean assumes the speech signal has energy in all frequency bins, i.e. reflects the fact that the speech is a wideband signal, but speech is also a sparse signal and absence of speech in several frequency bins will drive the likelihood very low. On the other hand the arithmetic mean will have high likelihood even if we have high energy only in a few frequency bins, which is definitely not speech. In [10] authors propose using a weighted sum to combine the likelihoods from the frequency bins:

$$\begin{aligned}\Lambda^{(n)} = \quad & \beta \exp\left(\frac{1}{(k_{end}-k_{beg})}\sum_{k=k_{beg}}^{k_{end}} \log\left(\Lambda_k^{(n)}\right)\right) \\ & + (1-\beta)\frac{1}{(k_{end}-k_{beg})}\sum_{k=k_{beg}}^{k_{end}} \Lambda_k^{(n)}.\end{aligned} \tag{9}$$

Here the parameter β is adjusted for achieving best accuracy. Also note the implicit bandpass filtering by processing only the frequency bins between k_{beg} and k_{end}.

We can apply likelihood smoothing in the same way as in equation (7) by introducing b_{01} and b_{10}, which are the prior probabilities for switching the state on frame level. We can compute the speech presence probability $P^{(n)}$ after using equation (8).

The binary flag $V^{(n)}$ for speech presence (1) or absence (0) can be obtained by comparing the likelihood $\Lambda^{(n)}$ or the speech presence probability $P^{(n)}$ with fixed threshold η:

$$V^{(n)} = \begin{cases} 1 & \text{if} \quad P^{(n)}(H_1) > \eta \\ 0 & \text{if} \quad P^{(n)}(H_1) \leq \eta \end{cases} \tag{10}$$

For practical purposes a small hysteresis is added to prevent "ringing" of the flag when the probability is close to the threshold.

At the end of processing of each frame we can update the noise model:

$$\begin{aligned}\lambda_N^{(n)}(k) = \quad & \lambda_N^{(n-1)}(k) + \\ & P_k\left(H_0|X_k^{(n)}\right)\frac{T}{\tau_p}\left(\left|X_k^{(n)}\right|^2 - \lambda_N^{(n-1)}(k)\right)\end{aligned} \tag{11}$$

where $P_k\left(H_0|X_k^{(n)}\right) = 1 - P_k\left(H_1|X_k^{(n)}\right)$ is the speech absence probability, T is the frame shift time, and τ_p is the time constant for updating the model.

The introduced VAD parameters (time constants, prior probabilities, etc.) can be optimized for given dataset using the approach described in [4].

III. Deep Learning Approach

The challenges for VAD increase with the proliferation of mobile devices and infotainment systems in cars. In both cases the noise levels are higher and SNRs are lower. Far field sound capture also adds higher reverberation, compared to close talking microphones in smartphone devices. The consumer of the enhanced speech shifts from telecommunications to speech recognition. While taking a phone call, people try to find a quieter place simply because there are limitations of how much power we can put in the headphone, before starting to harm the users' hearing. In the case of speech enabled dialog system for a mobile device, the user speaks and the system typically responds by showing the results on the screen. This lifts the limits at how noisy conditions the system should work – the user will be happy if the system can understand when asking with a normal tone in a noisy stadium.

In general the speech presence probability is a function of the magnitudes of the frequency bins in the current and several previous audio frames. The question is can a neural network learn that function, without assumptions for the statistical distribution of the speech and noise signals, without explicitly handling the temporal and spectral contexts, and with adding the capability for distinguishing between speech and fast varying non-stationary noises.

In this paper, we propose to use a fully connected deep neural network (DNN) as shown in Figure 3. The input features are the magnitudes of all frequency bins in the current and several previous audio frames, forming the current segment. The output is the probabilities for speech presence in each frequency bin and the probability for speech presence in the entire frame. The performance of the DNN will be evaluated both against seen noise (i.e. this type of noise is presented in the training set) and unseen noise (this type of noise has not been presented in the training set).

IV. Dataset and Evaluation

A. Dataset

A multi-condition training corpus with different noise types, signal-to-noise ratios (SNRs), and reverberant properties was created based on the TIMIT training set [13]. We used a collection of 100 different noise signals from [14], which includes a variety of different noise types (crowd noise, traffic and car noise, etc.). We also used a set of 60 different room

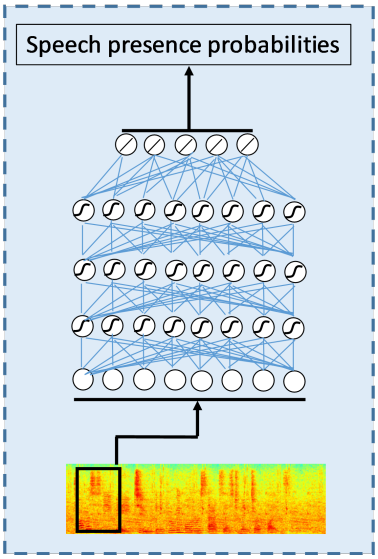

Fig. 3. Proposed DNN-based VAD.

impulse responses (RIRs) recorded at multiple distances (from 1 to 4 meters) in a room with reverberation time (T_{60}) of approximately 300 ms.

The training corpus was created as follows: speech and noise sound pressure levels (SPL) in a room were assumed to be normally distributed with means $\mu_s = 60$ dBA SPL and $\mu_n = 55$ dBA SPL, and standard deviations $\sigma_s = 8$ dB and $\sigma_s = 10$ dB respectively. An utterance is randomly selected from the TIMIT training set, and scaled to a level that is randomly selected according to the assumed distribution for speech levels. Similarly, a randomly selected signal from the noise dataset is scaled to a level chosen from the noise power distribution. Correction for the Lombard effect is performed on speech signal level. The scaled speech signal is convolved with a randomly selected RIR, and the scaled noise is added to the result. This noisy signal is then synchronized with the clean speech signal to remove the delay introduced by the RIR. Such a temporal alignment of the noisy and clean reference signals is necessary so that the subsequent framing and feature extraction steps will produce feature pairs which correspond to the same section of the speech signal. The final SNRs were limited to $[-5, 30]$ dB.

This procedure is used to create a dataset of clean/noisy pairs for training. In a similar fashion, we generated two different test datasets based on the TIMIT test set, each containing 200 utterances. The first test dataset uses noise signals from the noise corpus used to generate the training dataset, and the second uses a completely disjoint set of noise samples from NOISEX-92 corpus [15]. We call these seen and unseen test datasets, respectively.

B. Labeling and Evaluation

The ground truth is binary (speech signal presented or absent) and was obtained by running a simple threshold-based VAD on the clean speech utterances. TIMIT utterances are recorded with very high quality and simple comparison with given threshold provides a flag for presence or absence of speech signal for both per-bin and per-frame labels.

The evaluation criterion is the root-mean-squared (RMS) error between the VAD output and the ground truth obtained above:

$$E_f = \sqrt{\frac{1}{N}\sum_n \left(P^{(n)}\left(H_1\right) - G^{(n)}\left(H_1\right)\right)}. \qquad (12)$$

Here $G^{(n)}\left(H_1\right)$ is the ground truth.

TABLE I
VAD CLASSIFICATION ERRORS

Dataset	Per bin	Per frame
Baseline, average	0.46328	0.68949
Development	0.32068	0.24669
Test	0.31418	0.41633
Test, unseen	0.32560	0.44601

V. EXPERIMENTAL RESULTS

All of the voice and noise files were converted to 16 kHz sampling rate. To convert from absolute sound pressure levels to the signal on the output of the ADC convertor the clipping levels of the microphone was assumed 120 dB SPL – typical for most of the widely used MEMS microphones. We have generated 400 files for training, 200 files for testing, and 200 files for testing with unseen noises. The total duration of the dataset was 2 hours.

The frame size was 512 samples, weighted with Hann window before converting to frequency domain. This results in 256 frequency bins for each audio frame. Frame shift was 256 samples (50%). Overlap and add procedure was used to synthesize the signal back to time domain as described in [16].

Each segment consisted of seven frames, which means input feature vector of 1792 magnitudes. The neural network had four hidden layers of 512 nodes each. The output layer had 257 neurons: one for the speech presence probability for the entire frame and 256 for each frequency bin. For training and evaluation we used CNTK toolkit [17].

The VAD classification errors, according to equation (12) are shown in Table I. Note that this is the RMS of the error, compared with a binary classifier. This means that a well working VAD with output probability of 0.1 when speech is not present and 0.9 when speech is present will have error of 0.2. The baseline is the VAD, described in Section II. As expected the results against the test dataset degrade. There also noticeable degradation, but less than expected in the results against the test dataset with unseen noise. The classification error as function of the SNR is shown in Figure 4. The trends are consistent with the numerical results.

VI. CONCLUSIONS

In this paper, we proposed using a deep neural network to overcome shortages in the models used by the statistical VAD. We achieved substantial reduction of the classification error for

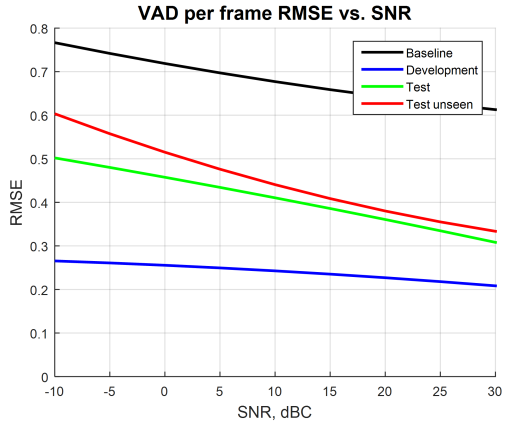

Fig. 4. VAD error per frame.

both seen and unseen noises. The reduction of the performance against unseen noises was less than expected. As a reasonable next step, we consider experimenting with different neural networks, for example RNN with LSTM for preserving the state and reducing the size of the input vector.

REFERENCES

[1] "Recommendation G.729 Annex B: a silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," 1997.

[2] J. Sohn and W. Sung, "A voice activity detection employing soft decision based noise spectrum adaptation," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1998.

[3] J. Sohn, N. Kim, and W. Sung, "A statistical model based voice activity detector," *IEEE Signal Processing Letters*, vol. 6, pp. 1–3, 1999.

[4] Ivan Tashev, Andrew Lovitt, and Alex Acero, "Unified framework for single channel speech enhancement," in *Proceedings of IEEE Pacific Rim Conference on Communications, Computers and Signal Processing*, 2009.

[5] S. Gazor and W. Zhang, "Speech probability distribution," *IEEE Signal Processing Letters*, vol. 10, pp. 204–207, 2003.

[6] Ivan Tashev and Alex Acero, "Statistical modeling of the speech signal," in *Proceedings of International Workshop on Acoustic, Echo, and Noise Control (IWAENC)*, 2010.

[7] R. Martin, "Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.

[8] T. Lotter, *Speech Enhancement*, chapter Single- and Multi-Microphone Spectral Amplitude Estimation Using a Super-Gaussian Speech Model, pp. 67–95, Springer-Verlag, 2005.

[9] Ivan J. Tashev and Malcolm Slaney, "Data driven suppression rule for speech enhancement," in *Information Theory and Applications Workshop*. University of California in San Diego, 2013.

[10] Ivan Tashev, Andrew Lovitt, and Alex Acero, "Dual stage probabilistic voice activity detector," in *NOISE-CON 2010 and 159th Meeting of the Acoustical Society of America*, 2010.

[11] Ivan J. Tashev, "Offline voice activity detector using speech supergaussianity," in *Information Theory and Applications Workshop*. University of California in San Diego, 2015.

[12] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[13] John S. Garofolo et al., "TIMIT acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, 1993.

[14] G. Hu, "100 nonspeech environmental sounds," in *http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html*, 2004.

[15] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

[16] Ivan J. Tashev, *Sound Capture and Processing: Practical Approaches*, Wiley, July 2009.

[17] Dong Yu and et al, "An introduction to computational networks and the computational network toolkit," Tech. Rep., Microsoft MSR-TR-2014-112, 2014.