Speaker and Noise Independent Voice Activity Detection

François G. Germain
Department of Music
Stanford University

Dennis L. Sun
Department of Statistics
Stanford University

Gautham J. Mysore
Adobe Research

March 26, 2013

Abstract

Voice activity detection (VAD) in the presence of heavy, non-stationary noise is a challenging problem that has attracted attention in recent years. Most modern VAD systems require training on highly specialized data: either labeled mixtures of speech and noise that are matched to the application, or, at the very least, noise data similar to that encountered in the application. Because obtaining labeled data can be a laborious task in practical applications, it is desirable for a voice activity detector to be able to perform well in the presence of any type of noise without the need for matched training data. In this paper, we propose a VAD method based on non-negative matrix factorization. We train a universal speech model from a corpus of clean speech but do not train a noise model. Rather, the universal speech model is sufficient to detect the presence of speech in noisy signals. Our experimental results show that our technique is robust to a variety of non-stationary noises mixed at a wide range of signal-to-noise ratios and outperforms a baseline.
Index Terms: non-negative matrix factorization, voice activity detection, universal models

1 Introduction

Voice activity detection (VAD) refers to the problem of identifying the speech and non-speech segments in an audio signal. It is a front-end component of many speech processing systems, including robust speech recognition [1, 2, 3] and compression systems for low-bandwidth transmission [4, 5].

Heavy and non-stationary noise pose serious challenges to VAD systems, and research in recent years has focused on developing robust systems [6]. A typical modern VAD system is trained either on mixtures of speech and noise that are matched to the application and have been labeled with voice activity (*supervised* learning) [7, 8, 9], or at the very least on noise data similar to the noise encountered in the application (*semi-supervised* learning) [10, 11, 12, 13]. In the latter case, the methods implicitly assume that noise training data is available because they require an initialization of a noise model. The semi-supervised methods listed above are also based on parametric assumptions about the noise (e.g., Gaussianity) that may be grossly violated in non-stationary noise environments.

It can be difficult and laborious to obtain such specialized training data. Thus, it is desirable to design a VAD system that is both *unsupervised*, in that it can operate without training data, and *robust*, in that it can handle a variety of noise environments over a wide range of signal-to-noise ratios. Earlier VAD systems, such as G.729B [4] and AMR [5], followed a rule-based approach and thus required no training data. They have largely been superseded by statistical and classification-based approaches, which are more robust and produce superior results [7, 8], but require labeled training data. Recently, there has been interest in developing unsupervised VAD systems that have the performance advantages of supervised systems. The usual approach has been to add an element of adaptivity to existing supervised and semi-supervised methods [14, 15].

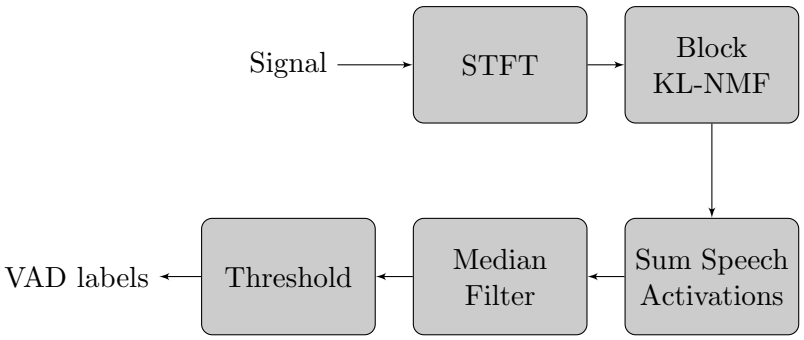

Figure 1: A schematic for the proposed method. The method is comprised of two main stages, feature extraction (first row) and classification (second row).

We propose a different approach, based on non-negative matrix factorization (NMF), a popular model in the source separation literature [16, 17]. In contrast to the aforementioned VAD approaches, we explicitly model the mixture of sounds (speech and noise). This has the advantage that if one has a reasonable general model for speech, then the approach will work in any noise environment. We will describe in detail how to obtain such a *universal speech model* in the next section, but generally speaking, this model is trained on a database of clean speech from a number of speakers. Once it is learned, it can be used to detect speech (from any unseen speaker) in any noise environment. Therefore, once the system is deployed, it is unsupervised from a user's perspective. Our approach also has the advantage of being fully interpretable—the features we use for classification correspond exactly to the relative levels of the speech and noise if we were to use this model for separation.

2 Proposed Method

Like most approaches to voice activity detection, our approach proceeds in two stages: feature extraction, followed by classification. The two stages are shown in the first and second rows, respectively, of Figure 1. Both the feature extraction and classification come naturally out of models for source separation. We describe each stage in turn in the subsections below.

2.1 Feature Extraction

Because humans tend to perceive spectral features of audio—at least on short time scales—it is natural to use frequency-domain rather than time-domain features in audio processing. This is well-known in speech processing, where mel-frequency cepstral coefficients (MFCCs) have long been standard features. In source separation, it is typical to work with invertible transforms, such as the Short-Time Fourier Transform (STFT), because it is necessary to recover the time-domain signals.

Audio signals are additive, so each frame of a magnitude spectrogram is roughly the sum of the spectral features that comprise it. If we think of a magnitude spectrogram as a matrix $V := (V_{ft})$ of non-negative numbers so that each column is the spectrum at time t, then this is saying that each column of the matrix can be written as:

$$V_{\cdot t} \approx \sum_k H_{kt} W_{\cdot k}$$

where $W_{\cdot k}$ denotes a spectral feature (indexed by k) and H_{kt} is the activation of that feature at time t. The critical assumption we will make is that these spectral features remain the same across all time. Since all sounds must be generated from this fixed set of spectral features,

we say that $(W_{\cdot k})_{k=1}^{K}$ is a model for the sound class. If we define matrices $W := (W_{fk})$ and $H := (H_{kt})$, then the above statement can be restated in matrix form as

$$V \approx WH. \tag{1}$$

Non-negative matrix factorization (NMF) [18] is a method for uncovering these spectral features W and the corresponding activations H from a magnitude spectrogram V [16]. It solves the optimization problem

$$\underset{W,H\geq 0}{\text{minimize}} \;\; D(V||WH) \tag{2}$$

for some measure of divergence D between V and WH. The non-negativity constraint ensures that the factors W and H can be interpreted as energies and activations.

Turning to the problem at hand, if we have a mixture of speech and noise, then W is comprised of a model for speech W_S and a model for noise W_N, i.e. we can partition (1) as:

$$V \approx \begin{bmatrix} W_S & W_N \end{bmatrix} \begin{bmatrix} H_S \\ H_N \end{bmatrix} \tag{3}$$

where H_S and H_N are matrices containing the activations of the speech and noise features, respectively.

However, applying NMF directly to the mixture spectrogram will not yield the representation (3), since it is impossible to identify the speech features W_S apart from the noise features W_N. However, if one is able to learn either W_S or W_N from clean training data and fix these quantities in applying NMF to the mixture spectrogram, then there is enough structure to distinguish the two sources. This is known as semi-supervised (if one of W_S and W_N is fixed) or supervised learning (if both are fixed) in the source separation literature [19].

In source separation, one also faces the problem of obtaining clean training data of the sources to be separated. Because existing algorithms depend on clean examples of the specific speaker and/or noise encountered in the mixture, they have difficulty generalizing to unseen speech and noise. We recently proposed a source separation technique [20], which leverages the knowledge that one of the sources is speech to perform source separation. The idea is to learn a model from clean speech examples from many different speakers (but not necessarily the speaker in the recording) and then incorporate this so-called "universal" speech model into the source separation pipeline. This is accomplished by learning a model $W^{(g)}$ for each speaker $g = 1, ..., G$ in the speech corpus and then adding a penalty in the optimization criterion to encourage the activation coefficients $H^{(g)}$ of most of the speakers to be zero. In other words, we now have the model:

$$V \approx \begin{bmatrix} W^{(1)} & \cdots & W^{(G)} & W_N \end{bmatrix} \begin{bmatrix} H^{(1)} \\ \vdots \\ H^{(G)} \\ H_N \end{bmatrix} \tag{4}$$

where many of the $H^{(g)}$ are entirely zero so that the corresponding speaker model $W^{(g)}$ is effectively not used. This captures the intuition that only a few models should be necessary to explain any given speaker and ensures robustness against poorly fitting speaker models in the speech corpus.

In order to encourage many of the blocks $H^{(g)}$ to be zero, we add a regularization term to the NMF problem (2) that encourages block sparsity.

$$\underset{W,H\geq 0}{\text{minimize}} \;\; D(V||WH) + \lambda \sum_{g=1}^{G} \log(\beta + ||H^{(g)}||_1) \tag{5}$$

where $H = \begin{bmatrix} H_S & H_N \end{bmatrix}^T = \begin{bmatrix} H^{(1)} & \cdots & H^{(G)} & H_N \end{bmatrix}^T$, leaving the user with the choice of λ, which controls the tradeoff between separation and artifacts. We consider the case where D is Kullback-Leibler divergence, denoted D_{KL}.

The algorithm for solving (5) with KL divergence is called Block KL-NMF and presented in Algorithm 1. We refer the reader to [20] for the derivation.

Algorithm 1 Block KL-NMF

inputs V, W_S
initialize H, $W = \begin{bmatrix} W_S & W_N \end{bmatrix}$ (assuming $1^T W = 1$)
repeat
  $R \leftarrow V./(WH)$
  $H \leftarrow H.^*(W^T R)$
  for $g = 1 : M$ **do**
    $H_g \leftarrow \dfrac{1}{1 + \lambda/(\beta + ||H_g||_1)} H_g$
  end for
  $W_N \leftarrow W_N.^*(RH_N^T)$
  $W_N \leftarrow W_N./(11^T W_N)$ (renormalize W)
until convergence **return** H

$.*$ and $./$ denote componentwise multiplication and division.

2.2 Classification

After solving (5), classifying each time frame as either speech or non-speech is straightforward. We simply sum up the speech activations $a_t = \sum_{k=1}^{K_S} H_{kt}$, where K_S is the total number of speech features, to produce a single activity number for each frame. After median filtering a_t to produce a smoothed estimate $\tilde{a}_t$, we classify a frame as speech if $\tilde{a}_t > c$ and non-speech otherwise. The user can adjust the threshold c depending on the desired true-positive and false-positive tradeoff.

Note that our classification algorithm depends only on the speech activations and not on the noise activations. This ensures that our algorithm is robust to non-stationary noise environments where the signal-to-noise ratio may be fluctuating.

3 Experiments

In this section, we determine parameter settings for our method and evaluate its performance relative to existing methods.

3.1 Data

We trained universal models with $N = 10, 20, 30, 40, 50, 60$ speakers (half male, half female) from the TIMIT speech database and $K = 5, 10, 20, 30, 40, 50$ features per speaker.

We then formed a synthetic data set using speech from held-out speakers in the TIMIT database, mixed with a variety of stationary and non-stationary noise samples from two different sources: the NOISEX-92 database [21] and the noise examples used in Duan et al., which we will refer to as the Duan data set [22]. Whereas the former contains primarily stationary noise examples, the latter is comprised of highly non-stationary noise examples. We considered signal-to-noise ratios of $-12, -6, 0$, and 6 dB. The duration of each mixture signal was 30 seconds, with several speech segments interspersed throughout the examples. Each speech segment is a TIMIT sentence, which is approximately 3-seconds long.

The sampling rate of all examples was 16kHz, and the signals were processed using a Hann window of length 64ms and a hop size of 16ms.

3.2 Parameter Determination

To determine optimal parameter settings, we divided the data set of speech and noise mixtures into a development and a test set. For each parameter setting, we applied the pipeline shown in Figure 1 to the examples in the development set. As we vary the decision threshold c for classifying a time frame as speech, we obtain a tradeoff between the false positive and false negative rates. We used the accuracy at the equal error rate (EER), for comparing the different parameter settings. This is the error rate at which the false positive and false negative rates are equal.

This parameter sweep uncovered $N = 20$ and $K = 10$ as the optimal parameters for the universal model. Although in principle it is possible to choose the number of noise spectral features K_N depending on the noise environment, in the interest of automating the VAD system, we also conducted a sweep over K_N, finding the optimal number over a wide class of noises to be $K_N = 10$. Also, although the optimal group sparsity parameter λ ideally should depend on the SNR, for simplicity we also determine a single optimal value over all the examples, finding $\lambda = 4096$. Finally, we found a median filter on blocks of 70 frames to work best. This set of parameters was used on the test set in the experiments below.

3.3 Baselines

We compare the proposed method to two existing methods [4, 14]. Both are natural candidates for comparison to our method because they neither require training data from the user, nor assume that the beginning of the signal contains no speech. The first method, the G.729B VAD [4], is a classical algorithm that extracts several acoustical features combined together by fuzzy rules to produce a single decision for each frame. The second method is a recent unsupervised technique based on sequential Gaussian mixture models (SGMM) [14]. We used the standard C implementation of G.729B and an implementation of SGMM provided by the authors. As shown in Section 3.4, the proposed method significantly outperforms both baselines.

3.4 Experimental results

Figures 2 and 3 show the filtered activity curves for two different noise environments: keyboard noise (non-stationary) and jet fighter noise (stationary). The black line at the NOISEX-92 top shows the decision at the EER threshold (dotted line), and the gray line below shows the ground truth.

To obtain ROC curves, we vary the decision threshold c on the median-filtered activity curve estimated from the signal. For each value of the threshold, we compute the *true positive rate* (TPR) and *false positive rate* (FPR). We also vary a decision threshold to compute the ROC curve for the SGMM model. These curves are shown in Figures 4 and 5 for three different noises each from the NOISEX-92 and Duan data sets at three different SNRs: -6 dB, 0 dB, and 6 dB. We also show the TPR and FPR for the G.729B VAD as a single point on these plots.

To facilitate comparison with G.729B VAD, we also tabulated the accuracy (the percentage of correctly labeled frames) at the EER threshold for our method and the SGMM. These numbers are shown in Tables 1 and 2. Both the ROC curves and tables confirm that our method significantly outperforms existing approaches in a wide variety of noise environments, even in challenging heavy noise environments.

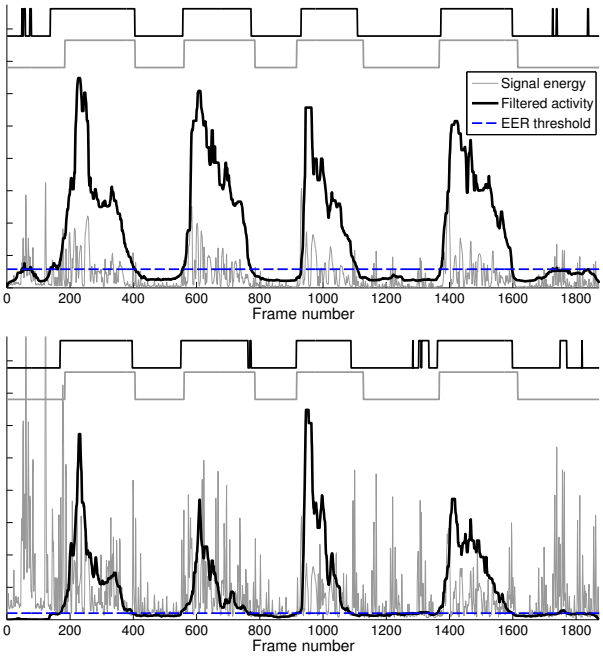

Figure 2: Median-filtered activity curve for keyboard background noise from the Duan data set for 6dB SNR (top) and -6dB SNR (bottom). The VAD decision at the EER threshold (black) and ground truth (gray) are shown at the top.

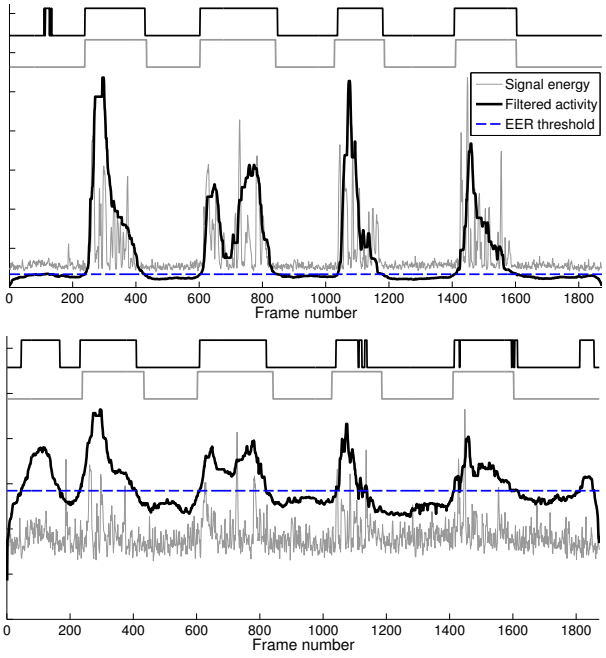

Figure 3: Median-filtered activity curve for the Buccaneer figher jet noise from NOISEX-92 for 6dB SNR (top) and -6dB SNR (bottom). The VAD decision at the EER threshold (black) and ground truth (gray) are shown at the top.

4 Conclusion

We have presented a method based on non-negative matrix factorization for performing voice activity detection that requires no training data from the user and is robust to changes in the

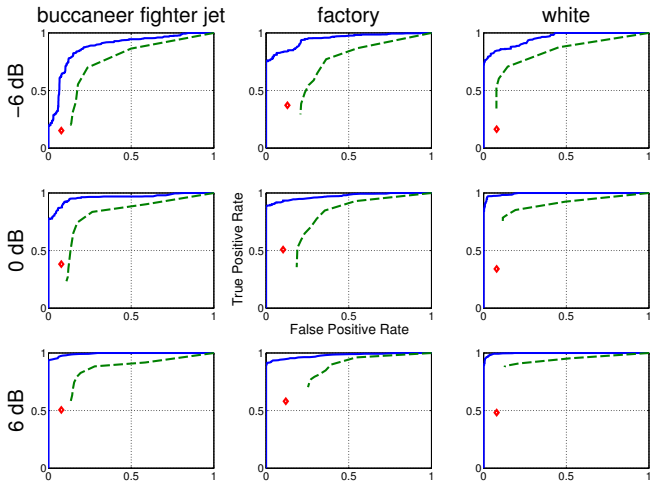

Figure 4: ROC curves for 3 examples of noise background from the NOISEX-92 data set mixed at 3 SNRs. For comparison, the result of SGMM (dashed) and the G.729B VAD (◊) are shown.

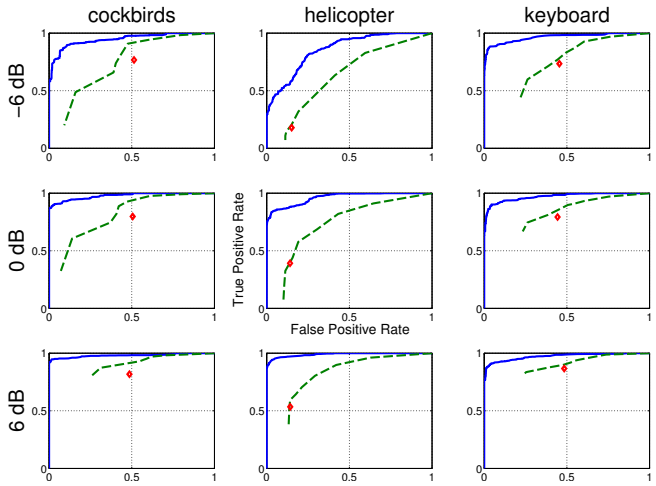

Figure 5: ROC curves for 3 examples of noise background from the Duan data set mixed at 3 SNRs. For comparison, the result of SGMM (dashed) and the G.729B VAD (◊) are shown.

noise environment. In particular, our method is able to handle a variety of non-stationary noises at low signal-to-noise ratios. Our experiments show that this approach significantly outperforms existing approaches.

However, it is important to note that the proposed approach is a batch algorithm, whereas in many applications an online method that performs real-time voice activity detection is desired. We believe that recent work on online extensions of NMF-based source separation [22] can be adapted to the universal speech model, making an online version of the proposed approach possible. However, we defer this and other extensions to future work.

5 Acknowledgements

We are grateful to Dongwen Ying for sharing code and Bhiksha Raj for helpful references.

	Accuracy (%)		
SNR	Proposed	SGMM [14]	G.729B [4]
6dB	94.1	78.2	69.9
0dB	91.5	76.0	65.4
−6dB	84.7	69.1	58.5
−12dB	76.0	63.3	53.6

Table 1: Average accuracy of the proposed method and of the baseline methods with the NOISEX-92 background noises. For our method and SGMM, the accuracy is computed at the EER.

	Accuracy (%)		
SNR	Proposed	SGMM [14]	G.729B [4]
6dB	92.7	67.7	62.7
0dB	90.1	64.1	61.0
−6dB	85.4	60.3	58.2
−12dB	76.6	55.5	55.7

Table 2: Average accuracy of the proposed method and of the baseline methods with the Duan background noises. For our method and SGMM, the accuracy is computed at the EER.

References

[1] L. Karray and A. Martin. Towards improving speech detection robustness for speech recognition in adverse conditions. *Speech Communication*, 40(3), 261-276.

[2] J. Ramirez, J. C. Segura, M. C. Bentez, A. de la Torre, A., and A. Rubio. A new adaptive long-term spectral estimation voice activity detector. In *Proceedings of Eurospeech*, 2003.

[3] A. Misra. Speech/Nonspeech Segmentation in Web Videos. In *Proceedings of Interspeech*, 2012.

[4] ITU-T Recommendation G.729-Annex B. A silence compression scheme for G.729 optimized for terminals conforming to recommendation V.70.

[5] ETSI EN 301 708 Recommendation. Voice Activity Detector (VAD) for Adaptive Multi-Rate (AMR) Speech Traffic Channels.

[6] J. Ramirez, J. M. Gorriz, and J. C. Segura. Voice Activity Detection. Fundamentals and Speech Recognition System Robustness. In M. Grimm and K. Kroschel. *Robust Speech Recognition and Understanding*, 1-22.

[7] E. Dong, G. Liu, Y. Zhou, and X. Zhang. Applying Support Vector Machines to Voice Activity Detection. In *Proceedings of the International Conference on Signal Processing (ICSP)*, 2002.

[8] T. Kinnunen, E. Chernenko, M. Tuononen, P. Franti, and H. Li. Voice activity detection using MFCC features and support vector machine. In *Proceedings of the International Conference on Speech and Computer*, 2007.

[9] P. Harding and B. Milner. On the use of Machine Learning Methods for Speech and Voicing Classification. In *Proceedings of Interspeech*, 2012.

[10] J. Sohn, N. Soo, and W. Sung. A statistical model-based voice activity detection. *IEEE Signal Processing Letters* 6(1), 1999.

[11] Y. Cho, K. Al-Naimi, and A. Kondoz. Improved voice activity detection based on a smoothed statistical likelihood ratio. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001.

[12] J. Ramirez, J. Segura, C. Bentez, L. Garca, and A. Rubio. Statistical voice activity detection using a multiple observation likelihood ratio test. *IEEE Signal Processing Letters 12(10)*, 2005.

[13] J. Ramirez, J. Segura, J. Gorriz, and L. Garcia. Improved voice activity detection using contextual multiple hypothesis testing for robust speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 15(8), 2007.

[14] D. Ying, Y. Yan, J. Dang, F. K. Soong. Voice Activity Detection Based on an Unsupervised Learning Framework. *IEEE Transactions on Audio, Speech, and Language Processing* 19(8), 2011.

[15] M. K. Omar. Speech Activity Detection for Noisy Data using Adaptation Techniques. In *Proceedings of Interspeech*, 2012.

[16] P. Smaragdis and J. C. Brown. Non-Negative Matrix Factorization for Polyphonic Music Transcription. In *IEEE Workshop of Applications of Signal Processing to Audio and Acoustics (WASPAA*, 2003.

[17] T. Virtanen. Monaural Sound Source Separation by Nonnegative Matrix Factorization with Temporal Continuity and Sparseness Criteria. *IEEE Transactions on Audio, Speech, and Language Processing* 15(3), 2007.

[18] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature* 401 (6755), 1999.

[19] P. Smaragdis, B. Raj, and M. V. Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Proceedings of the International Conference on Independent Component Analysis and Signal Separation*, 2007.

[20] D. L. Sun and G. J. Mysore. Universal Speech Models for Speaker Independent Single Channel Source Separation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[21] A. Varga and H. J. M. 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* 12(3), 1993.

[22] Z. Duan, G. J. Mysore, and P. Smaragdis. Online PLCA for real-time semi-supervised source separation. In *Proceedings of the International Conference on Latent Variable Analysis and Source Separation (LVA/ICA)*, 2012.