A Discrete Wavelet Transform-Based Voice Activity Detection and Noise Classification With Sub-Band Selection

Salinna Abdullah, Majid Zamani, and Andreas Demosthenous
Department of Electronic and Electrical Engineering, University College London,
Torrington Place, London WC1E 7JE, UK
e-mail: salinna.abdullah.13@ucl.ac.uk; m.zamani@ucl.ac.uk; a.demosthenous@ucl.ac.uk

***Abstract*—A real-time discrete wavelet transform-based adaptive voice activity detector and sub-band selection for feature extraction are proposed for noise classification, which can be used in a speech processing pipeline. The voice activity detection and sub-band selection rely on wavelet energy features and the feature extraction process involves the extraction of mel-frequency cepstral coefficients from selected wavelet sub-bands and mean absolute values of all sub-bands. The method combined with a feedforward neural network with two hidden layers could be added to speech enhancement systems and deployed in hearing devices such as cochlear implants. In comparison to the conventional short-time Fourier transform-based technique, it has higher F_1 scores and classification accuracies (with a mean of 0.916 and 90.1%, respectively) across five different noise types (babble, factory, pink, Volvo (car) and white noise), a significantly smaller feature set with 21 features, reduced memory requirement, faster training convergence and about half the computational cost.**

Keywords—Discrete wavelet transform, mel-frequency cepstral coefficients, multilayer perceptron, noise classification, sub-band selection, voice activity detection.

I. Introduction

Speech enhancement algorithms such as those employed in cochlear implants have been shown to perform well in noisy conditions to a limited extent. Conventional speech enhancement algorithms that utilize spectral subtraction [1] and statistical-models [2] generally achieve significant improvement in speech intelligibility in stationary noise, but only modest improvement in non-stationary noise. The success of these algorithms has been limited partly because although they have been created to accommodate all acoustic environments, they only show optimal speech enhancement over a limited range of background noise scenarios [1]. Real-world auditory environments exhibit a large variety of temporal and spectral characteristics that require a more adaptable approach to speech enhancement. This has been demonstrated by noise adaptive speech processing pipelines using supervised learning such as in [3], where speech enhancement performance was improved by adjusting the denoising requirement according to different acoustical conditions.

In this paper, a method is proposed as shown in Fig. 1 for achieving a compact and robust acoustic noise classification system that could potentially be implemented in hearing devices, where a small, low-power and robust system is desired. This study extends [4] by proposing an adaptive wavelet-based voice activated detector (VAD) which uses a more computationally efficient method for wavelet sub-band selection through energy examination of the sub-bands instead of Hurst exponents and ℓ_2-norm. Mel-frequency cepstrum coefficients (MFCCs) extracted from the selected sub-bands are used in combination with mean absolute values (MAVs) from all sub-bands for feature extraction to further account for the global variation across the sub-bands for classification training and testing.

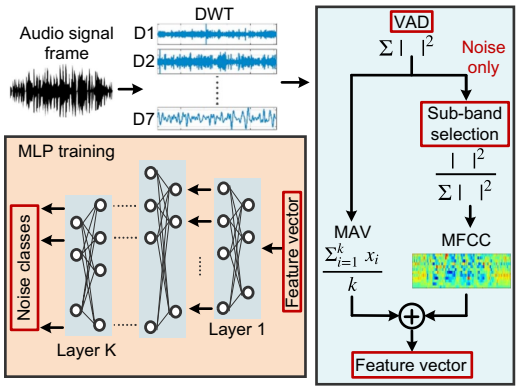

Fig. 1. Architecture of the proposed pre-enhancement system within a speech processing pipeline that uses wavelet parameters for VAD and sub-band selection, resulting in more effective feature extraction.

In the proposed method, 25-ms frames of the audio signal obtained by using a Hamming window are distinguished either as noisy speech or noise-only frames through a wavelet-based VAD system. The wavelet-based VAD employs thresholds that are adaptively tuned according to the sound and noise level variations over time to more accurately track speech with noise and noise-only segments. Thereafter, energy examination of wavelet decomposed noise-only frames is used to identify the most informative sub-bands for feature extraction. MFCC extraction is employed to obtain useful signatures from the selected sub-bands and the MFCCs are then combined with a vector of MAVs of the wavelet coefficients of all sub-bands to form the overall feature vector used for classification training and testing. With the MAVs, an insight into the global variation between sub-bands is obtained whereas the MFCCs further decompose the localized energy to provide important signatures of the selected sub-bands with only 13 coefficients. Audio signal frames and sub-band level features are selectively used for the classification training and testing using a 4-layer perceptron neuron network. The simulation results obtained using clean speech utterances mixed with different types of noise show that the proposed method is capable of accurately distinguishing between speech and non-speech frames. This achieves high classification accuracy for the trained noise types, providing more discriminative features for training and smaller memory requirement with a small set of features.

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant EP/R512400/1.

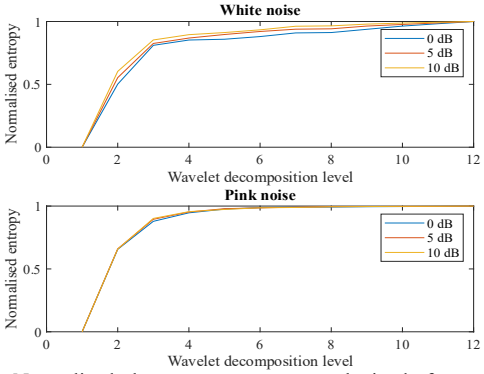

Fig. 2. Normalized log energy entropy obtained from wavelet decomposition levels 1 to 12 when speech utterances contaminated with white and pink noise at 0, 5 and 10 dB SNR were assessed.

II. Proposed Noise Classification and Voice Activity Detection Method

A. Discrete Wavelet Transform

Discrete wavelet transform (DWT) [5] represents a signal as series-approximations where the low pass version of the decomposition corresponds to the coarse approximations and the high pass version corresponds to the detail information. In this paper, the Daubechies wavelet of order 4 (db4) is used as the mother wavelet as it is shown in [6] to optimally capture key information present in speech. The number of decomposition levels is chosen to be 7 based on the log energy entropy assessment, of which sample results are presented in Fig. 2, conducted to find the optimal number of decomposition levels for use that would provide a desirable balance between classification accuracy and computational cost. Fine frequency resolution for low frequencies is required since most of the noise types used in the experiment exhibit greater energy in frequencies below 500 Hz. However, fine frequency resolution at low frequencies would mean a greater number of decomposition levels, thus demanding more computations. In Fig. 2, the sample entropy assessment results show that with a speech utterance contaminated with white and pink noise at 0, 5 and 10 dB SNR, the resulting entropy begins to saturate when a decomposition level of 7 is used. Therefore, utilizing 7 decomposition levels is a reasonable compromise, where a sufficiently fine frequency resolution is achieved without introducing a significantly greater computational cost. Table 1 lists the frequency range for each sub-band for a 7-level decomposition on signals sampled at 8 kHz.

B. Voice Activity Detection

To increase the robustness of the noise classifier, the captured audio frame is identified as speech plus noise or noise-only. Feature extraction is activated for a noise-only frame, and the resulting features are fed to the classifier to establish the noise class of the noise signal. Enhancement settings can be then be adjusted accordingly to reduce the noise present in the subsequent noisy speech frames. The VAD is also used to ensure that the latter stages of the speech processing pipeline are deactivated for silent frames. In this paper, a VAD based on DWT is considered since this transform is already computed as part of the noise classification pipeline, thus limiting the computational burden on the overall system. DWT-based VADs have been explored in the literature such as [7] but in this work, it is demonstrated that a fairly robust VAD can be achieved from simple energy comparisons at the frame and sub-band levels.

TABLE 1. Frequency range for each sub-band (detail sub-bands D1-D7 and approximation sub-band A7)

Sub-band	Frequency Range (Hz)
D1	4000 – 8000
D2	2000 – 4000
D3	1000 – 2000
D4	500 – 1000
D5	250 – 500
D6	125 – 250
D7	62.5 – 125
A7	0 – 62.5

Wavelet energy is useful in emphasizing the amplitude variation between regions and the presence of homogeneity within a region. Therefore, wavelet energy features are extensively used in this work. For silence detection, the total sub-band energy E_{tot} is

$$E_{tot} = \sum_{l=1}^{L+1} E_l, \quad (1)$$

where E_l is the total energy of one sub-band and L is the number of decomposition levels. The frame will be classified as silent by setting the binary flag $f_{silence}$ to 1 if the E_{tot} is smaller than a fixed threshold, T_1

$$f_{silence} = \begin{cases} 1, & if\ E_{tot} < T_1 \\ 0, & \text{otherwise.} \end{cases} . \quad (2)$$

A non-silent frame will then be classified as either noisy speech or noise-only. A noisy speech frame usually contains a E_{tot} greater than a noise-only frame. An unvoiced, noise-only segment often shows some energy concentration in relatively higher frequency sub-bands, while a noisy speech segment shows energy concentration in lower frequency sub-bands. Clean speech in particular was found to exhibit significantly higher energy in the 5, 6 and 7 detail sub-bands (D5, D6 and D7), representing the 62.5 – 500 Hz frequency range. This is related to the sound pressure level and vocal loudness that are often higher in the fundamental frequency and lower formants of speech [8]. Therefore, a frame is classified as noisy speech if E_{tot} is greater than another fixed threshold, T_2, and the accumulated average energy of D5, D6 and D7 makes up more than a fixed ratio, R_1, of the accumulated average energy of all sub-bands. Thus, the flag for indicating a noise-only frame (f_{noise}) is set according to the following conditions:

$$f_{noise} = \begin{cases} 1, & if\ (E_{tot} < T_2)\ \&\ \left(\sum_{l=5}^{L} E_l < R_1\right) \\ 0, & \text{otherwise} \end{cases} . \quad (3)$$

In this study, thresholds T_1, T_2 and R_1 are adaptively tuned to provide the optimal VAD performance according to the sound and noise level variations over time. T_1 ranges between 0.001 and 0.01, with an increment of 0.001, and it becomes larger when the average E_{tot} is found to be relatively larger for a prolonged period of time (2500 × 25 ms frames long). Similarly, values between 0.01 and 0.1, with an increment of 0.005 are used for T_2 and values between 0.1 to 0.8, with an increment of 0.05 are used for R_1. R_1 is tuned according to $E_{l=5\ to\ L}$ variation over time. However, larger average $E_{l=5\ to\ L}$ over time often comes with a larger E_{tot}. It was found that for noisy speech samples presented at 10 dB SNR, values 0.005, 0.055 and 0.25 for T_1, T_2 and R_1, are often employed for the optimal VAD output.

C. Feature Extraction of Informative Sub-bands

The sub-band selection also exploits wavelet energy features because energies of different types of noise are often concentrated on different frequency bands. The sub-band selection process compares the sub-band energy ratio (SER), which identifies the relative energy distribution in the sub-bands, and selects the top 3 sub-bands with the highest normalized SER for feature extraction. The SER is given by

$$\text{SER} = \frac{\sum_{n=1}^{N}(C_n)^2}{\sum_{l=1}^{L+1}\sum_{n=1}^{N}(C_n^l)^2}, \quad (4)$$

where C_n represents the coefficient vector of a sub-band, N is the length of the coefficient vector and L is the number of decomposition levels. The SERs are then normalized according to the bandwidth of each sub-band.

The feature vector used for classification training and testing is a combination of MAVs calculated from the coefficients of each wavelet sub-band and MFCCs extracted from the three selected sub-bands. The MAVs can provide an insight into the global variation between the sub-bands whereas the MFCCs can further decompose the localized energy and provide more important signatures of the selected sub-bands. The MAV is given by

$$\text{MAV} = \frac{1}{N}\sum_{n=1}^{N}|C_n|, \quad (5)$$

where, C_n is the coefficient vector of a sub-band and N is the length of the coefficient vector. The resulting MAV feature vector will contain a total of 8 features since there are 7 detail sub-bands and 1 approximation sub-band. In the calculation of the MFCC features, a 24-channel mel-scale was used to obtain 13 MFCCs from the selected sub-bands that were concatenated together prior to the extraction. The decision to use 24 channels was predicated from conventional cochlear implants having between 12-24 electrodes (i.e., stimulation channels) [9] and only lower-order coefficients, in this case 13 coefficients, are kept because they contain the most information about the overall spectral shape of the signal.

D. Classification

The multilayer perceptron (MLP) employed in this work is a 4-layer perceptron neuron network (including 1 linear input and output layer) consisting of 2 hidden layers with 10 neurons each. The neurons employ the sigmoid activation functions, and the network uses the mean square error as a cost function and is trained with the Levenberg-Marquardt [10, 11] learning algorithm. The training is stopped when the magnitude of the gradient used to adjust the network weights and biases is less than 1e-5 or when the maximum training epoch, a measure of the number of times all training data are used once to update the network weights, of 1000 is reached.

III. Simulations

A. Datasets

1000 randomly chosen utterances from the TIMIT [12] training set were used as the training utterances and 100 utterances from the TIMIT core test set, consisting of 192 utterances from unseen speakers of both genders, were used as the test utterances. For the training and testing noises, 5 noises from the NOISEX [13] dataset were used. The noises are a mix of 4-minute long stationary and nonstationary noises that include a babble noise, factory noise, pink noise, Volvo (car) noise and white noise. A sampling frequency of 8 kHz was used throughout the experiment. For the training sets, random cuts of the first 2 minutes of each noise were used to mix with the training utterances at 10, 5 and 0 dB SNR. The test mixtures were in turn a mix of random cuts of the last 2 minutes of each noise and the test utterances at same SNRs.

B. Evaluation Methods

To assess the efficacy of the proposed VAD, the outcome from the VAD was visually compared with the unsupervised robust voise activity detection (rVAD) proposed in [14]. For the evaluation metric for the task of noise classification, the classic classification accuracy (CAcc) and F_1 [15] score were used. The CAcc is higher every time the trained DNN model correctly predicts the noise label for the test set. The F_1 score is calculated from the precision and recall of the test where the precision is the ratio of true positive (TP) detections over all positive detections including those not correctly identified (TP+FP); and the recall is ratio of TP detections over all detections that should have been identified as positive (TP+FN). The F_1 score is obtained by

$$F_1\ \text{score} = \frac{TP}{TP + \frac{1}{2}(FP + FN)}. \quad (6)$$

The computational cost was estimated in terms of the number of additions (or subtractions) and the number of weighted multiplications (or divisions) needed to execute the algorithm. The computational cost of the DWT, VAD, sub-band selection, feature extraction and MLP were combined to give the approximate overall cost for executing the proposed pre-enhancement pipeline within speech processing. The approximate inference time to process a single frame of 25-ms and the time taken to train an epoch were also evaluated on a CPU. The results obtained from the proposed method were compared with the results obtained when the following feature extraction methods were used: (1) STFT; (2) multi-resolution cochleagram (MRCG) [16], which encodes power distributions of an audio signal also in the time-frequency representation at different resolutions; and (3) MAVs combined with MFCCs extracted from all frames and wavelet decomposition levels (i.e., without VAD and sub-band selection).

C. Results and Discussion

An example result from the proposed wavelet-based adaptive VAD is shown in Fig. 3. It shows that the proposed VAD is more robust in noisy conditions when compared to the rVAD. When the VADs were tested on a speech sentence contaminated with babble noise at 10 dB SNR, the rVAD failed to distinguish between noisy speech and noise-only segments. In contrast, the proposed adaptive VAD was able to do the segmentation effectively albeit with some slight truncation and extended error, where noisy-speech frames are misjudged as noise-only and noise-only frames are misjudged as noisy speech. The proposed method continued to outperform the rVAD at lower SNRs (i.e., 5 dB and 0 dB SNR).

Fig. 4 shows the average CAcc achieved by the different feature extraction methods described in Section III.B when three different SNR values (10, 5, and 0 dB) were used for testing. The proposed method led to a higher mean

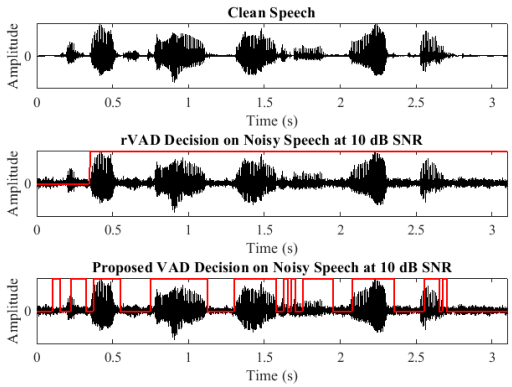

Fig. 3. Evaluation of the proposed VAD: (a) waveform of the original clean speech, (b) the rVAD output when tested on the clean speech contaminated with babble noise at 10 dB SNR and (c) the proposed wavelet-based VAD output on the same noisy speech with thresholds $T_1 = 0.005$, $T_2 = 0.055$ and $R_1 = 0.25$. 1's on the solid red line represent noisy speech segments. 0's represent either silent or noise-only segments.

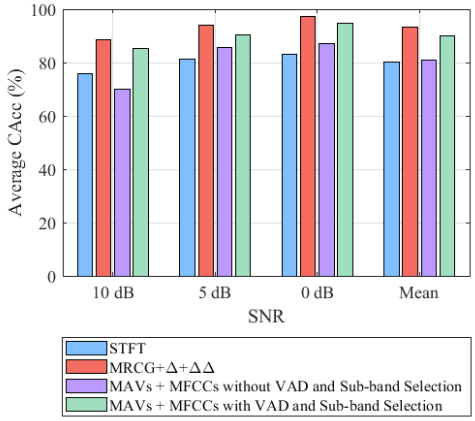

Fig. 4. Classification accuracy comparison. The addition of Δ and ΔΔ, (first and second-order derivatives, respectively) to yield the MRCG + Δ + ΔΔ feature set was suggested in [16] to better capture temporal dynamics of the signal. A 24-channel MRCG + Δ + ΔΔ feature set resulted in a dimensionality of 288 (24×4×3) for each frame.

classification accuracy of 90.1% than when STFT (80.1%) or MAVs + MFCCs without VAD and sub-band selection (81.0%) is used for feature extraction. This implies that the addition of the VAD and sub-band selection, and the feature extraction in the wavelet domain have significantly contributed to improving the noise CAcc. The CAcc obtained with MRCG + Δ + ΔΔ is the highest for every SNR scenario tested. However, the high classification accuracy of the MRCG has a much higher computational cost, and its feature size (288 features as shown in Table 2) for a single frame is larger than that needed by the proposed method by a factor of 17.1. Such a large feature size, when fed into the same neural network configuration means having more nodes at the input layer and thus, more parameters to take into consideration during training and testing. In addition to a wider network structure, this will lead to much longer training times and larger hardware memory. The observation that all the feature extraction methods assessed performed better when tested with 0 dB SNR than when tested with 5 dB or 10 dB SNR demonstrates their ability to learn from the noise components rather than the speech. The F_1 scores obtained were in agreement with the CAcc scores. A higher number of false negatives and false positives were obtained in higher SNR conditions. This again indicates that the systems have been conditioned to better classify noises when they are prevalent.

TABLE 2. COMPUTATIONAL COST, INFERENCE AND TRAINING TIME COMPARISON.

Feature	Feature vector size	Comp. Cost*	Inference time **	Train time***
STFT	101	72 053	186.79	29.39
MRCG + Δ + ΔΔ	288	628 590	1475.00	135.02
MAVs + MFCCs without VAD and Sub-band Selection	21	36 347	106.81	3.73
MAVs + MFCCs with VAD and Sub-band Selection	21	36 857	106.96	3.74

* Estimated cost per 25-ms frame based on ComputComp = $N_{add(sub)} + 10N_{mult(div)}$, where $N_{add(sub)}$ is the number of additions (or subtractions), and $N_{mult(div)}$ is the number of multiplications (or divisions) [17].
** Approximate time taken in milliseconds (ms) to obtain the output from the trained model given an input. This includes the time needed to extract the relevant features.
*** Approximate time taken in seconds (s) to train one epoch in MLP with 20,000 frame samples.

Similar to the CAcc results, the MRCG method gave the highest mean F_1 score of 0.916. This is followed by the proposed approach with the VAD and sub-band selection (mean F_1 score of 0.916).

Table 2 lists the estimated computational cost, inference time and training time per epoch of each feature extraction method. The proposed classification method is the second most efficient to compute and the addition of the VAD and sub-band selection did not significantly increased the computational cost. The computational cost for the proposed method is only around 1.4% more than the computational cost for the MAVs + MFCCs without VAD and sub-band selection. The increment in inference and train time introduced from the addition of the adaptive VAD and sub-band selection is also negligibly small. Overall, the proposed method is effective for reducing inference and training time, computational cost, and memory requirement whilst boosting the robustness of acoustic noise classification in comparison to the STFT method. Future work will include assessing the generalization performance of the proposed method and improving its VAD performance in much lower SNR, and CAcc and F_1 scores in higher SNR conditions.

IV. CONCLUSION

In this paper, a wavelet-based adaptive VAD has been explored and selection criteria have been formulated for efficient selection of the wavelet sub-bands for feature extraction. A simple feedforward MLP has been used for the recognition process. Results show that classification accuracies are improved with the VAD and sub-band selection by a mean of 9.1% (mean CAcc of 90.1% for the proposed method versus mean CAcc of 81.0% when VAD and sub-band selection were not used). The proposed approach led to a mean F_1 score of 0.916, indicating an acceptable proportion of false positive and false negative assignments. This is accompanied by a decrease in computational cost by a factor of around 1.95 and 17.1 when compared to a conventional STFT-based and the MRCG classification method, respectively. The low-computation wavelet and neural network-based framework are suitable for implementation in compact speech enhancement algorithms.

REFERENCES

[1] L. Yang and Q. Fu, "Spectral subtraction-based speech enhancement for cochlear implant patients in background noise", *The Journal of the Acoustical Society of America*, vol. 117, no. 3, pp. 1001-1004, Mar. 2005.

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

[3] S. Abdullah, M. Zamani and A. Demosthenous, "Towards more efficient DNN-based speech enhancement using quantized correlation mask", *IEEE Access*, vol. 9, pp. 24350-24362, Feb. 2021.

[4] S. Abdullah, M. Zamani and A. Demosthenous, "Acoustic noise classification using selective discrete wavelet transform-based mel-frequency cepstral coefficient", *International Journal of Simulation Science and Technology*, vol. 21, no. 2, Mar. 2020.

[5] S. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, no. 7, pp. 674-693, Jul. 1989.

[6] L. Lei and S. Kun, "Speaker recognition using wavelet cepstral coefficient, i-vector, and cosine distance scoring and its application for forensics", *Journal of Electrical and Computer Engineering*, vol. 2016, pp. 1-11, Nov. 2016.

[7] S. Joseph and A. Babu, "Wavelet energy based voice activity detection and adaptive thresholding for efficient speech coding", *International Journal of Speech Technology*, vol. 19, no. 3, pp. 537-550, Jun. 2016.

[8] B. M. Gelfer and Q. Bennett, "Speaking fundamental frequency and vowel formant frequencies: Effects on perception of gender*", Journal of Voice*, vol. 27, no. 5, pp. 556-566, Sept. 2013.

[9] N. Croghan, S. Duran and Z. Smith, "Re-examining the relationship between number of cochlear implant channels and maximal speech intelligibility", *The Journal of the Acoustical Society of America*, vol. 142, no. 6, pp. Dec. 2017.

[10] K. Levenberg, "A method for the solution of certain non-linear problems in least squares", *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164-168, Jul. 1944.

[11] D. W. Marquardt, "An algorithm for the least-squares estimation of nonlinear parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, no. 2, pp. 431-441, Jun. 1963.

[12] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallet and N. Dahlgren, *DARPA TIMIT: Acoustic-phonetic continous speech corpus*. Washington, DC, USA: US Department of Commerce, 1993.

[13] A. Varga and H. 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, pp. 247-251, 1993.

[14] Z. Tan, A. Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," *Computer Speech and Language*, vol. 59, pp. 1–21, Jan. 2020.

[15] N. Chinchor and B. Sundheim, "Muc-5 evaluation metrics", *Proceedings of the 5th Message Understanding Conference*, Aug. 1993, pp. 69-78.

[16] J. Chen, Y. Wang and D. Wang, "A Feature Study for Classification-Based Speech Separation at Low Signal-to-Noise Ratios", *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1993-2002, Sept. 2014.

[17] M. Zamani, J. Sokolic, D. Jiang, F. Renna, M. R. Rodrigues, and A. Demosthenous, "Accurate, very low computational complexity spike sorting using unsupervised matched subspace learning," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 14, no. 2, pp. 221–231, Feb. 2020.