

Unsupervised Speech Activity Detection Using Voicing Measures and Perceptual Spectral Flux

Seyed Omid Sadjadi and John H. L. Hansen, *Fellow, IEEE*

Abstract—Effective speech activity detection (SAD) is a necessary first step for robust speech applications. In this letter, we propose a robust and unsupervised SAD solution that leverages four different speech voicing measures combined with a perceptual spectral flux feature, for audio-based surveillance and monitoring applications. Effectiveness of the proposed technique is evaluated and compared against several commonly adopted unsupervised SAD methods under simulated and actual harsh acoustic conditions with varying distortion levels. Experimental results indicate that the proposed SAD scheme is highly effective and provides superior and consistent performance across various noise types and distortion levels.

Index Terms—Clarity, Combo-SAD, harmonicity, perceptual spectral flux, speech activity detection, voicing measures.

I. INTRODUCTION

SPEECH activity detection (SAD) has applications in a variety of contexts such as speech coding [1], automatic speech recognition (ASR) [2], speaker and language identification [3], and speech enhancement. In addition, for surveillance and monitoring applications that involve listening to long conversational audio recordings with small *a priori* speech presence probability, SAD, including segmentation techniques [6], [7], can help mitigate the excessive cognitive load on human listeners by removing long and often noisy non-speech intervals. In this study, our goal is to develop a robust and unsupervised SAD system for such severe noisy communication applications [4], [5].

State-of-the-art SAD techniques include both supervised and unsupervised approaches. Supervised methods, which are often based on either Gaussian mixture models (GMM) [5], hidden Markov models (HMM) [2], or multi-layer perceptrons (MLP) [5] work well given that pre-trained models for both speech and non-speech classes broadly match the acoustic characteristics of the test environment. Hence, they are limited to applications where a large amount of training data is available and the

acoustic properties of the test environment are consistent and known *a priori* as well.

On the other hand, unsupervised SAD techniques assume no *a priori* knowledge about the acoustic characteristics of the test environment, and can be categorized as feature-based [1], [8], or statistical model-based [9]–[12]. Feature-based methods perform well under stationary noise conditions at relatively high signal-to-noise ratio (SNR) levels, however their performance degrades rapidly as the noise level increases. Statistical model-based techniques employ a likelihood ratio test (LRT) of speech presence and absence hypotheses in the short-time Fourier transform (STFT) domain, assuming that an estimate of the noise power spectrum is available. The LRT based techniques are generally robust and effective, however their performance is dependent on the accuracy of the noise spectrum estimate, which is assumed to be uncorrelated and additive, making them vulnerable to the presence of non-stationary and rapid changing noise.

In this study, we propose a robust and unsupervised SAD system solely based on features that convey fundamental traits of speech, which are governed by the speech production process. The effectiveness of the proposed technique is evaluated in both simulated and actual noisy conditions using data from the SPINE2 corpus [13] and dry-run speech material from Phase-I of the DARPA program Robust Automatic Transcription of Speech (RATS) [4], respectively. Performance of the proposed SAD method is benchmarked against that of commonly adopted feature-based (e.g., ITU G729 Annex B [1]) and statistical model-based (e.g., single and multiple observation LRT [9], [10], [12]) approaches.

II. SYSTEM DESCRIPTION: COMBO-SAD

In this section, we describe the procedure for extraction of a 1-dimensional feature vector that is used in our system as soft-decision for the speech/non-speech discrimination task. This “combo” feature is *efficiently* obtained from a linear combination of four different voicing measures as well as a perceptual spectral flux feature. The voicing measures include harmonicity, clarity, prediction gain, and periodicity. The perceptual spectral flux and periodicity are extracted in the frequency domain, while the harmonicity, clarity, and prediction gain are all time domain features. For feature extraction, the audio signal is blocked into 32 ms frames with a 10 ms skip rate. In order to extract the periodicity, harmonicity, and clarity, an approximate knowledge of the plausible pitch range in human speech is required. Here, we choose a pitch period duration within the interval of [2, 16] ms (or equivalently [62.5, 500] Hz in the frequency domain), where the lower limit is imposed by the analysis frame length, and the fact that each frame should at least cover two pitch periods for a reliable voicing estimate.

Manuscript received December 02, 2012; accepted December 07, 2012. Date of publication January 04, 2013; date of current version January 23, 2013. This work was supported by the Defense Advanced Research Projects Agency (DARPA) under Contract D10PC20024. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA or its Contracting Agent, the U.S. Department of the Interior, National Business Center, Acquisition and Property Management Division, Southwest Branch. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Stephen McLaughlin.

The authors are with the Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, Department of Electrical Engineering, Richardson, TX 75080-3021 USA (e-mail: sadjadi@utdallas.edu; John.Hansen@utdallas.edu).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/LSP.2013.2237903

A. Time Domain Features

All time domain voicing measures in this section, directly or indirectly, use the normalized autocorrelation proposed in [14] for noise robust pitch estimation. The deterministic autocorrelation of a short-time windowed segment $x(n)$ is computed as,

$$r_{xx}(t, k) = \frac{\sum_{j=0}^{N-1} x(j)w(j)x(j+k)w(j+k)}{\sum_{j=0}^{N-1} w(j)w(j+k)}, \quad (1)$$

where $w(j)$ is a Hanning window, and t and k are frame and autocorrelation lag indices, respectively. It has been shown that normalization by autocorrelation of the window function in (1) effectively mitigates the impact of strong formants on the maximum autocorrelation peak in the pitch range, and obviates the need for low-pass filtering and/or center-clipping [15]. In addition, it also compensates for the windowing effect which tapers the autocorrelation function towards zero for larger lags.

1) *Harmonicity*: Harmonicity (a.k.a. harmonics-to-noise ratio) is defined as the relative height of the maximum autocorrelation peak in the plausible pitch range. Mathematically, it can be expressed as,

$$h(t) = \frac{r_{xx}(t, k_{\max})}{r_{xx}(t, 0) - r_{xx}(t, k_{\max})}, \quad (2)$$

$$k_{\max} = \arg \max_{2 \text{ ms} \leq k \leq 16 \text{ ms}} r_{xx}(t, k).$$

Note that the autocorrelation of a periodic signal is also periodic with the same period, and its maximum takes on values close to the autocorrelation at zero lag. Accordingly, for voiced segments which have periodic structure, the harmonicity shows sharp peaks.

2) *Clarity*: We define clarity as the relative depth of the minimum average magnitude difference function (AMDF) valley in the plausible pitch range. Computing the AMDF from its exact definition is costly; however, it has been shown that the AMDF can be derived (analytically) from the autocorrelation as [15], $D(t, k) \approx \beta(k) \cdot \sqrt{2[r_{xx}(t, 0) - r_{xx}(t, k)]}$, where $\beta(k)$ is a scale factor that can vary between 0.6 and 1.0. We have found that the clarity feature is not very sensitive with respect to the value of this parameter; therefore, we set $\beta(k)$ to 0.8 in our experiments. The clarity is then extracted as,

$$c(t) = 1 - \frac{D(t, k_{\min})}{D(t, k_{\max})}, \quad (3)$$

$$k_{\min} = \arg \min_{2 \text{ ms} \leq k \leq 16 \text{ ms}} D(t, k).$$

Subtracting the term from 1 in (3) simply converts the minimum to a maximum, which is more desirable for our application. In this manner, the clarity exhibits large values for voiced and speech-like segments, while maintaining a minimum for background sounds.

3) *Prediction Gain*: The prediction gain is defined as the ratio of the signal energy to the linear prediction (LP) residual signal energy. The signal energy can be obtained from the autocorrelation at zero lag. In order to calculate the LP residual signal energy, however, the Levinson-Durbin recursion [15] is applied and the error from the last step yields the energy of the residual signal. The prediction gain is then computed as, $G_p(t) = \log(r_{xx}(t, 0)/\epsilon^p)$, where ϵ^p is the error in the last step of the recursion, and p is the order of LP analysis which

is set to 10 in this study (assuming a sampling rate of 8 kHz). In short-time frames, there is a high correlation among speech samples, making it easier to predict, or in other words the denominator in $G_p(t)$ becomes smaller. Therefore, the average prediction gain reaches its highest value for voiced and speech-like frames.

B. Frequency Domain Features

Both features from this section are extracted in the STFT domain which is formed by taking a 2048-point DFT from Hanning windowed frames after zero padding. Magnitude information is only used and the phase response is discarded.

1) *Periodicity*: In the STFT domain, the harmonics of the pitch frequency are apparent in the magnitude spectrum of speech during voiced and speech-like segments. This observation serves as the basis for the harmonic product spectrum (HPS) technique [15] which has been widely applied for pitch detection in noisy environments. The HPS in the log-spectral domain is defined as, $P(t, \omega) = \sum_{l=1}^R \log |X(t, l\omega)|$, where R is the number of frequency-compressed copies of the original spectrum, which is fixed to 8 in this study. The frequency-compressed copies coincide at the fundamental frequency and reinforce the amplitude, while other harmonics are cancelled or attenuated in the final product. The periodicity is computed as the maximum peak of $P(t, \omega)$ in the plausible pitch range,

$$P_{hps}(t) = P(t, \omega_{\max}), \quad (4)$$

$$\omega_{\max} = \arg \max_{62.5 \text{ Hz} \leq \omega \leq 500 \text{ Hz}} P(t, \omega).$$

The periodicity is especially impervious to noise and other background sounds, since their spectral harmonics cannot combine coherently in the HPS. The periodicity can thus be used to effectively discriminate speech from non-speech sounds.

2) *Perceptual Spectral Flux*: Over short-time frames, speech is a quasi-stationary and slowly varying signal, meaning that its spectrum does not change rapidly from one frame to another. Hence, one can effectively exploit this quality to develop a feature capable of discriminating speech from other more rapidly varying sounds. An example of such a feature is the SF [16], which measures the degree of variation in the spectrum across time. Given the benefits of incorporating perceptual models into speech processing frameworks (e.g., MFCC), in this study we define the perceptual SF as,

$$SF_p(t) = \left\| X_m(t, \omega) - X_m(t-1, \omega) \right\|_1, \quad (5)$$

where $\|\cdot\|_1$ denotes the L^1 -norm, and $X_m(t, \omega)$ is the energy normalized mel-spectrum at frame t which is calculated using an 80-channel mel-filterbank spanning the frequency range from 0 to the Nyquist frequency. The perceptual SF exhibits relatively deep valleys for speech segments, while maintaining a maximum value for background sounds/silence. Accordingly, we employ the negative of this parameter as a feature for speech/non-speech discrimination.

After extracting the above noted features, a 5-dimensional vector is formed by concatenating the voicing measures along with the perceptual SF. Each feature dimension f_i is then normalized according to, $f'_i = (f_i - \mu_i)/\sigma_i$, where the mean μ_i and standard deviation σ_i are computed over the entire waveform. The normalized 5-dimensional feature vectors are linearly mapped into a 1-dimensional feature space represented by the most significant eigenvector of the feature covariance matrix. This is realized through a principal component analysis

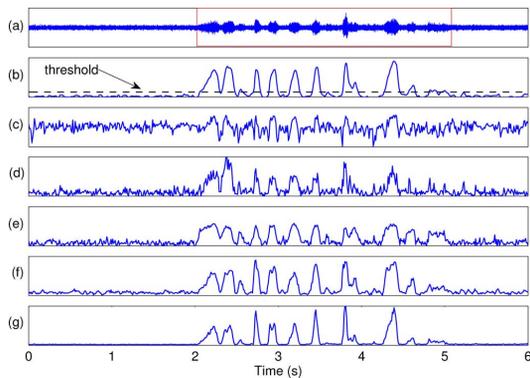


Fig. 1. Individual features as well as their combination for a sample waveform with an average segmental SNR of 2.59 dB. (a) noisy speech signal, (b) the combo feature along with the decision threshold (dashed), (c) the perceptual SF, (d) the periodicity, (e) the clarity, (f) the prediction gain, and (g) the harmonicity.

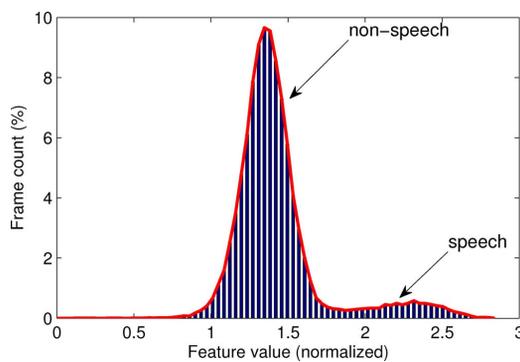


Fig. 2. Distribution of the combo feature values for a sample waveform.

(PCA) and retaining the dimension that corresponds to the largest eigenvalue. The 1-dimensional “combo” feature is smoothed via a 3-point median filter to serve as soft-decisions for the SAD (i.e., called “Combo-SAD”). Fig. 1 shows sample time domain plots of individual features as well as their combination for part of a noisy speech waveform with an average segmental SNR of 2.59 dB. It is seen that, although the individual features might not be as discriminative for the SAD task, their combination exhibits a great potential for noise robust speech/non-speech detection. Distribution of the combo feature values for a 366-second noisy speech signal is depicted in Fig. 2 (note that this is an example for illustration; the distribution of the combo feature can vary across different waveforms based on the distortion level and also the proportion between speech and non-speech frames, although it still remains bimodal). It is evident that the combo feature has a bimodal distribution in which speech and non-speech classes are well separated. We exploit this property for hard-decision making by fitting a 2-mixture GMM to the feature and estimating a detection threshold (Th) from a weighted average of the mixture means as, $Th = \alpha \mu_{sp} + (1 - \alpha) \mu_{sil}$, where μ_{sp} and μ_{sil} are the speech and non-speech mixture means, respectively. The weight parameter α can be tuned to achieve the desired false accept/reject rate (P_{fa} and P_{miss}). To increase robustness of the GMM fit over two distinct modes of the feature distribution, after random initialization, the EM algorithm is repeated 5 times and the model with the largest likelihood is selected.

III. EXPERIMENTAL SETUP

The proposed unsupervised SAD system is evaluated using speech material from the SPINE2 evaluation set [13] and

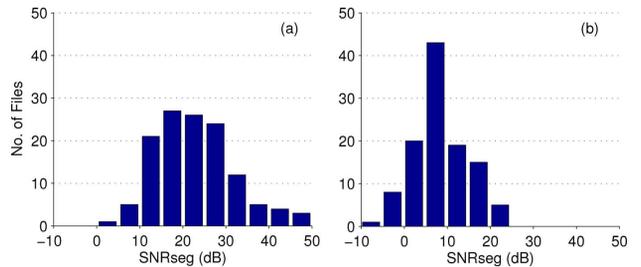


Fig. 3. Distribution of files in terms of the segmental SNR (dB) in (a) SPINE2 corpus, and (b) RATS Phase-I dry-run data.

RATS Phase-I dry-run data. The SPINE2 evaluation set consists of 64 talker-pair conversations in stereo format (128 mono waveforms) recorded in simulated military background noise environments. On average, each mono waveform is 180 seconds long containing only 78 seconds of speech activity. Background noise types include quiet, office, Humvee, aircraft carrier, E3A, and MCE field shelter. RATS dry-run data consists of a total of 111 conversational telephone speech (CTS) waveforms that were retransmitted and recorded over 8 extremely degraded communication channels with distinct noise characteristics. The distortion type is non-linear (similar to clipping) and to some extent correlated with speech. Each CTS file is 900 seconds long with very sparse speech activity. As previously noted in Section I, our goal here is to develop a robust unsupervised SAD system for long audio recordings with small *a priori* speech presence probability, in order to assist human listeners avoid auditing long noisy non-speech intervals. This justifies our choice of the aforementioned datasets for evaluations. However, our system has the potential to be adapted for automatic speech applications such as speaker and language identification, with no or minimal modifications.

Fig. 3 illustrates the SNR differences between the two data corpora used in our evaluations. It can be seen that RATS dry-run set is of comparatively much lower quality with the bulk of the data exhibiting an average segmental SNR below 10 dB (note, however, that due to its non-linear nature, the distortion in RATS data may not be strictly characterized by SNR). It is expected that a robust SAD should perform equally well on both datasets. Here, segmental SNR is reported because there is no access to the original clean speech or background noises for these databases, and the SNR is estimated based on speech/non-speech information from manual transcripts.

Given that majority of the features described in Section II can primarily detect voicing, the proposed SAD may perform poorly in detecting short unvoiced frames surrounding a voiced segment. To alleviate this issue, similar to [5] and [8], the boundaries of each detected speech segment are extended by 0.1 s (i.e., 10 frames). The same post-processing is applied for other techniques used in our evaluations.

IV. RESULTS AND CONCLUSION

We evaluate the proposed unsupervised Combo-SAD system on the two datasets, and compare its performance against that of five commonly adopted SAD solutions, namely ITU G729B [1], single-observation LRT (SOLRT) [9], SOLRT paired with an HMM-based hangover smoothing scheme [9], multiple-observation LRT (MOLRT) [10], and a recently introduced modification to MOLRT that incorporates harmonicity information into the LRT framework [12].

Receiver operating characteristic (ROC) curves obtained from the evaluations are shown in Fig. 4. These curves indicate

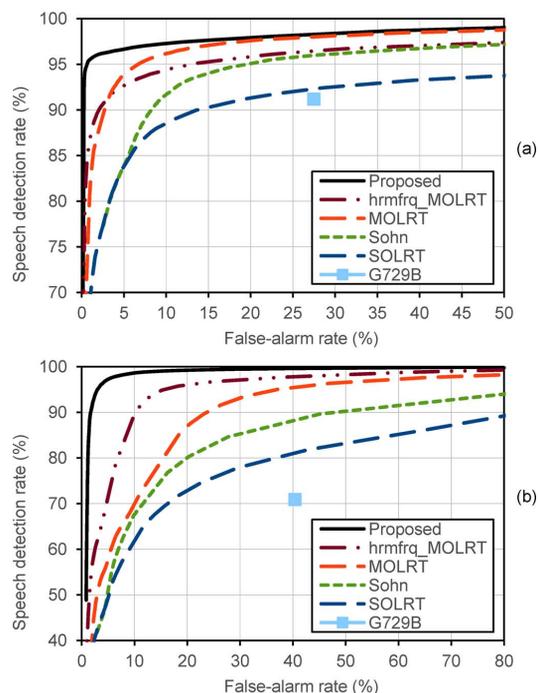


Fig. 4. Comparison of ROC curves for the proposed Combo-SAD method versus other feature-based and LRT-based techniques considered in the evaluations using (a) SPINE2 evaluations set, and (b) RATS dry-run data for Phase-I.

that: (i) as expected, the RATS dry-run data is much more challenging than the SPINE2 task for SAD. All SAD schemes considered perform reasonably well on the SPINE2 evaluation set, however except for the proposed Combo-SAD system and the modified MOLRT (hrmfrq_MOLRT), we observe a significant drop in speech detection performance on the RATS dry-run data. (ii) Our unsupervised system performs equally well on both high and low SNR data, which points to its robustness across diverse environmental noise and non-linear channel distortions. The basic assumption in formulating different flavors of statistical LRT based SAD is that the noise is additive and uncorrelated, however, it should be noted that the type of distortion seen in RATS data is non-linear and correlated with speech. This is one of the major reasons for the poor performance of this class of SAD methods on RATS data. In addition, it should be noted that many existing SAD systems (including those considered in this study) are developed for scenarios where higher energy is strongly correlated with speech presence (i.e., additive noise scenarios). However, due to the transmission failures, RATS data contains non-transmission (NT) regions characterized with high-amplitude static noise which adversely impact parameter/noise estimation and eventually performance of the existing methods. It is worth remarking here that according to the RATS program guidelines for the SAD task the NT regions are excluded from scoring.

DARPA has set specific targets for the SAD task within the RATS program that should be satisfied for acceptable system performance. A system that passes the Phase-I target should achieve a challenging missed detection rate (P_{miss}) of 5% or less at a false-alarm rate (P_{fa}) of 3%. Table I summarizes performances of the different SAD techniques in terms of P_{miss} for the datasets considered. It is observed that our unsupervised technique consistently outperforms the commonly adopted LRT-based methods, satisfying the Phase-I targets for the DARPA

TABLE I
PERFORMANCE COMPARISON OF THE PROPOSED COMBO-SAD TECHNIQUE VERSUS THE LRT-BASED APPROACHES IN TERMS OF P_{miss} AT $P_{\text{fa}} = 3\%$

Data	Proposed	hrmfrq_MOLRT	MOLRT	Sohn	SOLRT
SPINE2	3.7	8.6	8.3	20.9	20.9
RATS	4.6	37.5	49.3	57.3	56.7

TABLE II
PERFORMANCE COMPARISON OF THE INDIVIDUAL FEATURES USED IN THE PROPOSED COMBO-SAD ON RATS DATA, IN TERMS OF P_{miss} AT $P_{\text{fa}} = 3\%$

Feature	harmonicity	clarity	prediction gain	periodicity	pSF
P_{miss}	10.76	7.43	13.47	13.22	33.14

RATS program. For the sake of comparison, performance of the proposed Combo-SAD framework with the individual features used are reported in Table II. Clearly, the clarity and perceptual SF are the best and worst performing features, respectively, while the other 3 features show similar performances.

REFERENCES

- [1] A. Benyassine, E. Shlomot, H.-Y. Su, D. Massaloux, C. Lamblin, and J.-P. Petit, "ITU-T recommendation G.729 Annex B: A silence compression scheme for use with G.729 optimized for V.70 digital simultaneous voice and data applications," *IEEE Commun. Mag.*, vol. 35, pp. 64–73, Sep. 1997.
- [2] R. Sarikaya and J. H. L. Hansen, "Robust detection of speech activity in the presence of noise," in *Proc. ICSLP*, Dec. 1998, pp. 1455–1458.
- [3] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, "The delta-phase spectrum with application to voice activity detection and speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, pp. 2026–2038, Sep. 2011.
- [4] K. Walker and S. Strassel, "The RATS radio traffic collection system," in *Proc. ISCA Odyssey*, Jun. 2012.
- [5] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Vesely, and P. Matejka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. INTERSPEECH*, Sep. 2012.
- [6] R. Huang and J. H. L. Hansen, "Advances in unsupervised audio classification and segmentation for the broadcast news and NGSW corpora," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 3, pp. 907–919, May 2006.
- [7] B. Zhou and J. H. L. Hansen, "Efficient audio stream segmentation via the combined T^2 statistic based Bayesian information criterion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 467–474, Jul. 2005.
- [8] J. Ramirez, J. C. Segura, C. Benitez, A. de la Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech Commun.*, vol. 42, pp. 271–287, Apr. 2004.
- [9] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE Signal Process. Lett.*, vol. 6, pp. 1–3, Jan. 1999.
- [10] J. Ramirez, J. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Process. Lett.*, vol. 12, pp. 689–692, Oct. 2005.
- [11] A. Davis, S. Nordholm, and R. Togneri, "Statistical voice activity detection using low-variance spectrum estimation and an adaptive threshold," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, pp. 412–424, Mar. 2006.
- [12] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Proc. IEEE ICASSP*, Mar. 2010, pp. 4466–4469.
- [13] *Linguistic Data Consortium (LDC) Catalog Number: LDC2001S08*, [Online]. Available: <http://www.ldc.upenn.edu/Catalog/>
- [14] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of the sampled sound," in *Proc. Inst. Phonetic Sci.*, 1993, vol. 17, pp. 97–110.
- [15] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Upper Saddle River, NJ: Prentice-Hall, 2010.
- [16] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *Proc. IEEE ICASSP*, Apr. 1997, pp. 1331–1334.