Noise-robust speech triage

Anthony L. Bartos,[1] Tomas Cipr,[2] Douglas J. Nelson,[3,a)] Petr Schwarz,[2] John Banowetz,[4] and Ladislav Jerabek[1]

[1]*Suzanne R. Miller Associates, Marriotsville, Maryland 21104, USA*
[2]*Phonexia Limited and Brno University of Technology, Brno, Czech Republic*
[3]*United States Department of Defense, 9800 Savage Road, Fort Meade, Maryland 20755, USA*
[4]*Naval Research Laboratory, Washington, DC 20375, USA*

(Received 21 September 2017; revised 22 February 2018; accepted 26 March 2018; published online 23 April 2018)

A method is presented in which conventional speech algorithms are applied, with no modifications, to improve their performance in extremely noisy environments. It has been demonstrated that, for eigen-channel algorithms, pre-training multiple speaker identification (SID) models at a lattice of signal-to-noise-ratio (SNR) levels and then performing SID using the appropriate SNR dependent model was successful in mitigating noise at all SNR levels. In those tests, it was found that SID performance was optimized when the SNR of the testing and training data were close or identical. In this current effort multiple i-vector algorithms were used, greatly improving both processing throughput and equal error rate classification accuracy. Using identical approaches in the same noisy environment, performance of SID, language identification, gender identification, and diarization were significantly improved. A critical factor in this improvement is speech activity detection (SAD) that performs reliably in extremely noisy environments, where the speech itself is barely audible. To optimize SAD operation at all SNR levels, two algorithms were employed. The first maximized detection probability at low levels ($-10\,\text{dB} \leq \text{SNR} < +10\,\text{dB}$) using just the voiced speech envelope, and the second exploited features extracted from the original speech to improve overall accuracy at higher quality levels ($\text{SNR} \geq +10\,\text{dB}$). https://doi.org/10.1121/1.5031029

[JFL] Pages: 2313–2320

I. INTRODUCTION

We address the problem of efficient speaker, language, and gender identification (SID, LID, and GID) and diarization of speech in degraded conditions. Our goal is to triage potentially massive amounts of data to efficiently extract relevant information. A variety of approaches have been proposed to improve the performance of SID,[1–5] LID,[6–8] GID,[9] and diarization[10,11] in degraded channels. Generally, these methods rely on modifying extracted features or analysis methods to make performance more robust. We have previously demonstrated[12] that several eigen-channel SID algorithms[13,14] could be significantly improved by testing on speaker models trained on a lattice of signal-to-noise-ratio (SNR) levels. It was found that the best classification performance occurred when the SNR of the test data nearly matched the SNR of the training data. An advantage of this approach was that no modification of the recognition process was required.

Our present effort is based on the i-vector approach developed at the 2008 JHU workshop[15] as simplification of Kenny's Joint Factor Analysis model.[16] We apply methods similar to our previous eigen-channel SID efforts to improve the performance of i-vector SID, LID, GID, and diarization. In the present tests, SID, GID, LID, and diarization were applied to speech at SNR levels ranging from barely audible speech ($\text{SNR} = -10\,\text{dB}$) to relatively noise-free speech ($\text{SNR} \geq +30\,\text{dB}$). Unlike our eigen-channel effort, three SNR estimates are required for the i-vector models and the best performance does not occur when the three SNR estimates are equal. Like our eigen-channel work, no modifications of the recognition processes are required.

For i-vector processes, extracted features and processes associated with LID, GID, and speaker diarization are similar to those used for SID, so noise affects all of these processes similarly. We examine SID, LID, GID, and diarization, demonstrating a significant performance improvement for these i-vector processes. Performance of our original eigen-channel work was measured as EER.[12] This same metric is used in the current work, with the exception of diarization, whose performance is measured as diarization error rate (DER).

The i-vector approach provides a means of reducing large dimensional input data to a small dimensional feature vector while retaining speaker relevant information and improving classification performance.[17–19] Improvements in throughput are realized because comparisons of the small i-vector generated voiceprints require only 150 floating-point numbers, compared to the numerically large (~512 to 2048) Gaussian mixture models (GMM), resulting in thousands of floating-point numbers required in the original Eigen-channel implementation. For large voiceprint libraries, i-vector speaker comparisons are many times faster. For 10 000 speakers, SID may be more than 30 times faster than real time.

a)Electronic mail: waveland@erols.com

Prior to performing speech recognition/classification functions in extremely noisy environments, time segments of active speech must be differentiated from periods of speech inactivity. We use two different speech activity detection (SAD) algorithms, depending on the signal quality. The first of these is based entirely on the speech envelope. Envelope-based SAD algorithms generally perform very well in most situations. Bach *et al*. demonstrated a 10 dB SNR advantage of their envelope-based SAD over mel-frequency cepstral coefficient (MFCC) based processes.[20] An early example of such SAD algorithms is the syllable rate speech activity detector (SRSAD) that was developed in the 1990s.[21] SRSAD is still in common use because it reliably distinguishes speech from noise and other signals in voice grade channels and because of its ability to process about 8000 voice grade channels in real time. A shortcoming of SRSAD is that it does not perform well in negative SNR conditions. The newer syllable rate voice activity detector (SRVAD), used in this current effort, is based on the same principles as SRSAD but performs much better in severe noise.[22]

One key issue in matching training sets to test data is reliably estimating the quality of the data. In our previous eigen-channel work, this was accomplished with an iterative procedure in which SNR and speech activity were iteratively re-estimated to obtain final SNR and speech activity estimates. For the present i-vector experiments, this process was too computationally intensive for large scale applications. In this work, SNR was estimated using SAD segmentation. The signal-plus-noise power was estimated from the envelope of the speech segments, and the noise power was estimated from periods of non-speech. This process was efficient and accurate enough for our purposes.

This paper is structured as follows. Section II describes the attributes of the speech activity algorithms used in the simulations, automatic estimation of speech activity, ground truth, and generation of speech at various quality levels. Section III introduces the i-vector processing approach, while Secs. IV, V, and VI specifically address these new SID, GID, and LID i-vector approaches with classification results, as a function of the inputted speech test SNR. Section VII covers noise-robust diarization. The paper closes with general conclusions and recommendations in Sec. VIII.

II. SPEECH ACTIVITY DETECTION AND QUALITY ESTIMATION

SRVAD is used for the lowest speech signal levels ($-10\,\text{dB} < \text{SNR} < 0\,\text{dB}$). It uses only the AM envelope waveform. The process is loosely based on the syllable rate since it attempts to exploit syllable rate features in the spectrum of the signal envelope. At the 8 kHz sample rate of the original speech signal, the signal envelope is significantly over sampled. In computing the signal envelope, the signal is normalized to have zero mean and unit variance by subtracting the signal mean and dividing by the square root of the signal variance. The analytic signal is computed using the Hilbert transform and the signal envelope is computed as the magnitude of the analytic signal. A lowpass filter with a cutoff frequency of 15 Hz is applied, and the filtered envelope is down sampled by a factor of 128 to a sample rate of 62.5 Hz.

A one-sided spectrogram representing the signal envelope was computed with 64-point (corresponding to 1.024 s) hamming windowed Fourier transforms with an overlap of 58 samples. The spectrogram was "over-normalized" by dividing the magnitudes of components of each spectrum by the sum of the squared spectral magnitudes. Three features were extracted from each spectrum: (1) the maximum spectral magnitude, (2) the absolute difference computed as the sum of absolute values of the difference of the spectral magnitudes and the mean spectral magnitude, and (3) the centroid of the over-normalized envelope spectrum.

A speech/silence determination is made every 6th envelope sample, or every 0.096 s using quadratic discriminant analysis (QDA).[22] Each QDA calculation produces two pattern recognition distances, corresponding to the speech and the noise, where typically one would choose the lower distance to be the correct estimate of speech activity or absence. To prevent the decision from rapidly oscillating between the two classes, a voting algorithm is employed. This is the case for SRSAD. In SRVAD, we have opted to simplify the process by median filtering the two distance vectors to remove any short impulsive artifacts. The seven-point median filter outputs for the two QDA distance vectors, are designed to smooth out class changes of less than 0.3 s.

Advantages of envelope activity detection include a significant sampling rate reduction, allowing simultaneous real-time monitoring of multiple channels. Because features are extracted from the speech envelope spectrum and not the speech spectrum itself, this process is insensitive to spectral color, which can be imparted to a voice channel by sensor hardware used in acquisition. Away from any speech activity, the false alarm rate is almost negligible ($P_{fa} < 0.1\%$). At higher SNR values, it typically declares speech activity 3 to 4 report samples prior to and after (0.288 to 0.384 s earlier and later) the actual beginning and ending of the speech segment. Additional edge correction processing now corrects for this. Unfortunately, that option was not available during the simulations.

SRVAD has been demonstrated to provide a 50% detection rate at the −10 dB SNR, where only the voiced speech segments can be detected. At this low SNR, the 50% detection rate was still sufficient to achieve SID performance with a 19% EER. At SNR levels of 0 and 30 dB, the SID EER was reduced to 7.5% and 3.23%, respectively. For SNR levels above +10 dB, we use a different SAD developed by Phonexia, Ltd. In this paper it is referred to as Phonexia VAD (or PVAD). PVAD is more complex and less efficient than SRVAD but is more accurate for higher quality signals.

For SNR levels greater than +10 dB, PVAD provided more accurate speech activity than SRVAD (operating without its edge correction option), primarily at the onset and termination of speech activity. PVAD invokes successive tests with decreasing throughput and increasing accuracy. The first of these is simply an energy-based test, followed by a technical signal removal function that removes telephony tones, pulses, flat spectra, etc. Next a fundamental speech (pitch) frequency is estimated and tracked, where longer

signals lacking a fundamental frequency are removed. Finally, a more computationally-intensive, but very precise, neural network SAD algorithm is applied.

To prepare data needed for testing these two SAD algorithms, a separate "reference" activity detection algorithm was used. This algorithm is based on a phoneme recognizer with some post-processing of the output speech segmentation and was applied to dual-channel clean recordings in order to automatically generate the SAD ground truth prior to degrading them with additive Gaussian white noise. When training/testing with other types of noise (car, babble), the original multiple quality model lattice approach to noise-robust SID also appeared to be valid. As a result only wide-band stationary white noise was used during this effort. The main reason for using wide-band stationary noise as the degradation medium was that it is most often encountered when monitoring communication channels that have very challenged, barely detectable link budgets. The noise spectrum itself need not be necessarily white, because both SAD algorithms use AM envelope features that are insensitive to spectral shape. Only a reasonable short term stationarity assumption is required. SRVAD does not apply technical signal removal of telephony tones, pulses, etc., as PVAD does, so these scenarios were not included in the simulations.

III. i-VECTOR THEORETICAL BACKGROUND

In the i-vector approach the speaker and channel is modeled by one GMM. The GMM is represented as a super-vector of GMM means (concatenation of mean vectors from all Gaussians). The speaker and channel dependent super-vector is adapted from a universal background model (UBM) super-vector using a shift that is given by a linear combination of basis vectors with the maximal variability. The number of basis vectors is much smaller (hundreds) in comparison to the super-vector size (tens of thousands). This can be described by the expression

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{s}$ is the speaker and channel dependent super-vector, $\mathbf{m}$ is the UBM GMM mean super-vector, $\mathbf{T}$ is a low-rank matrix representing M basis vectors spanning the subspace with maximum variability in the mean super-vector space, and $\mathbf{w}$ is a standard normal distributed vector of size M. For each observation X, the goal is to estimate the parameters of the posterior probability of $\mathbf{w}$,

$$\mathrm{p}(\mathbf{w}|\mathrm{X}) = \mathrm{N}(\mathbf{w}; \mathbf{w}x, \mathbf{L}^{-1}x). \tag{2}$$

The i-vector is the maximum *a posteriori* (or MAP) point estimate of the variable $\mathbf{w}$, i.e., the mean $\mathbf{w}$ of the posterior distribution $\mathrm{p}(\mathbf{w}|\mathrm{X})$. It maps most of the relevant information from a variable-length observation X to a fixed- (small-) dimensional vector. $\mathbf{L}^{-1}x$ is the precision of the posterior distribution. Averaging over time is done through a collection of GMM statistics. The i-vector extraction does not remove any channel or non-speech effect. It preserves the total variability of GMM statistics in the MAP sense. The channel variability is subsequently removed by linear discriminant analysis (LDA), i-vector mean normalization, normalization of the i-vector to unit length, and later in a classifier based on probabilistic-LDA (PLDA).[23]

A relative comparison of SID throughput was conducted on a Quad-Core AMD Opteron Processor 8356, with 128 GB of RAM, as a function of speaker library size. Due to the huge reduction of voiceprint data that has to be stored, memory requirements for the i-vector technique were essentially constant while Eigen-channel memory usage grew exponentially, as did processing times. Table I provides typical processing times for the i-vector and eigen-channel methods as a function of speaker library size, when processing one minute of speech. Note that for a small number of speakers, the i-vector method is slower because it must perform additional processing associated with dimensionality reduction down to a 150-point voiceprint. However, for most practical speaker library sizes (>100) this approach is well worthwhile as is evident from Table I.

IV. NOISE-ROBUST SID

As noted earlier, eigen-channel SID systems yielded the best performance in terms of minimum EER when the SNR of the training and testing records were identical. This is *not* the case for the i-vector-based SID system, where optimum performance is defined in terms of *three* SNR levels, not two. Global system parameters (UBM GMM parameters and T matrix) were calculated at one level of noise, defined as SNR1. Speaker enrollment (analogous to training) and testing quality, were represented by SNR2 and SNR3, respectively. A lattice of ten SNR levels (−10, −7.5, −5, −2.5, 0, 5, 10, 15, 20, >30 dB) were used for all three SNR variables. An SNR = +30 dB or greater is considered clean or essentially noiseless. The two algorithm components that make up the i-vector SID operation are voiceprint extraction and voiceprint comparison.

First, in the voiceprint extraction algorithm, illustrated in Fig. 1, after features are extracted from the original speech, statistics are collected using the UBM (consisting of a Gaussian mixture model with 512 Gaussians). Next, these statistics are converted to an i-vector, which preserves the global variability of speech in 400 floating-point numbers. The speaker-only information is extracted using LDA, resulting in the final voiceprint vector, which has 150 floats. The quality level of the speech used to calculate the global system parameters (i.e., the universal background model and the projection parameters) is SNR1. Depending upon whether it is an enrollment or test record, the SNR of the input speech can either be SNR2 or SNR3.

The second i-vector SID system algorithm, illustrated in Fig. 2, compares two voiceprints and generates a log

TABLE I. Relative Processing times (in seconds) for 60 seconds of speech.

Method	10 speakers	100 speakers	1000 speakers	10 000 speakers
i-vector	1.3274	1.3333	1.3575	1.9108
Eigen-channel	0.99338	5.7143	54.545	600

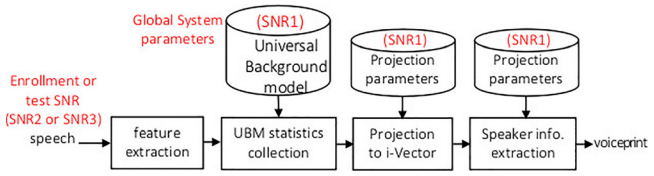

FIG. 1. (Color online) Voiceprint extraction module.

likelihood ratio (LLR) score, where the recording from which the voiceprints were generated may contain different noise levels (SNR2 and SNR3). The model and calibration parameters are extracted from speech with a noise level corresponding also to SNR1, as is the case with the first algorithm.

The voiceprint comparison is accomplished by probabilistic linear discriminant analysis (PLDA), followed by score calibration. The calibration is accomplished using logistic regression (LR) trained for specific combinations of enrollment and test record lengths. The output LLR score points to two hypotheses where the voiceprints are either spoken by the same or different speakers. Optionally, the LLR score can be converted to a percentage score using a sigmoidal function.

In order to limit the large number of simulations required, only female records were used on the NIST 2010 SRE set, condition 5 data,[24] because higher pitch female recordings tended to yield worst case EER SID results. In order to produce a noise-robust SID (NRSID) system, an optimum set of SNR results leading to a minimum EER had to be determined. A total of 10^3 simulations were executed, where the average speech duration was about 2 s. This resulted in ten 10×10 matrices of EER values that had to be searched for a minimum, as a combined function of SNR1, SNR2, and SNR3. To illustrate the effect of choosing equal and non-equal SNR levels, the reference activity detector was used to establish ground truth for speech activity. For the case of equal SNR levels (SNR1 = SNR2 = SNR3), all ten SNR diagonal matrices were compared to yield the NRSID performance shown in Table II.

Using PVAD on the clean data to determine speech activity segmentation, the ten 10×10 matrices corresponding to the global system parameters (SNR1), the enrollment or training record parameters (SNR2) in the rows and the test record parameters (SNR3) in the columns of each 10×10 matrix, were searched to determine the optimal combination, which is shown in Table III. Differences in NRSID performance, when selecting equal SNR (Table II) or non-equal SNR levels (Table III) indicate an EER reduction at *all* SNR levels, with a more significant improvement at SNR levels of 0 dB and below.

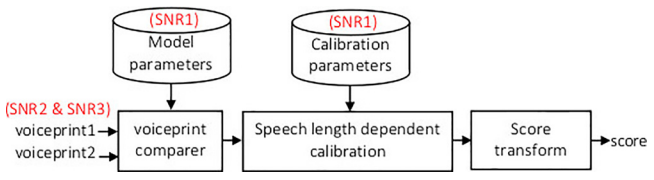

FIG. 2. (Color online) Voiceprint comparison module.

TABLE II. IOC EER SID performance for identical test, train and system SNRs.

SNR (dB)	−10.0	−7.5	−5.0	−2.5	0	5	10	15	20	>30
EER (%)	22.93	17.34	12.59	9.75	8.01	6.03	5.33	4.36	3.79	3.23

Figure 3 illustrates the original[12] unmodified eigen-channel-based SID performance, the resulting improvement to eigen-channel SID performance when the training and testing SNR levels are matched, the unmodified i-vector SID performance and the final operating capability (FOC) performance of the i-vector SID variant that is described above. The improvement between the unmodified eigen-channel (dashed circle) and SNR-matched (solid circle) SID versions is telling, as is the initial (dashed*) and FOC (solid*) performance of the i-vector SID variant. It is interesting to note that the EER of the unmodified eigen-channel performance at an SNR = 10 dB and FOC NRSID performance at an SNR = −10 dB, are essentially equal, allowing a 20 dB reduction in SNR, realizing nearly the same SID classification performance (EER ∼ 19%).

V. NOISE-ROBUST GID

The noise-robust gender identification (NRGID) system is also i-vector-based and can use the same voiceprints[17] generated by NRSID. If both NRSID and NRGID are always to be executed, the additional processing necessary to also realize gender classification processing is negligible. This is realized using the logistic regression classifier. In the second instance, if gender is to be used as a discriminator prior to any other speech triage processing to essentially "halve the search," then more processing would be required to extract the voiceprints. However; if it is then followed by NRSID as part of a larger speech triage scheme, the EER vs SNR performance for NRGID would be very similar. Figure 4 illustrates the processing blocks and the associated SNR definitions for this second "stand-alone" instance of NRGID.

The first step requiring voiceprint extraction, as shown in Fig. 1, is illustrated as a single module, where the system parameters are trained at the system SNR level or SNR1, as before. The second module in Fig. 4 classifies the voiceprints as to gender and assigns a score, where the model parameters were trained at a particular level, designated as SNR2. As before, the input test record quality can be designated as SNR3. As was the case with NRSID, the best configuration of system (SNR1), model (SNR2), and test quality (SNR3) is achieved by running simulations to determine the minimum EER from the ten 10×10 matrices. A total of 5000 records

TABLE III. SNR levels in (dB) used to obtain optimal (minimum EER) SID performance.

SNR1 = system	−2.5	−2.5	−2.5	0	0	5	15	20	>30	>30
SNR2 = train	−5.0	0	0	5	5	10	20	>30	>30	>30
SNR3 = test	−10.0	−7.5	−5.0	−2.5	0	5	10	15	20	>30
EER (%)	19.25	13.09	10.42	8.48	7.48	6.02	4.81	3.75	3.35	3.13

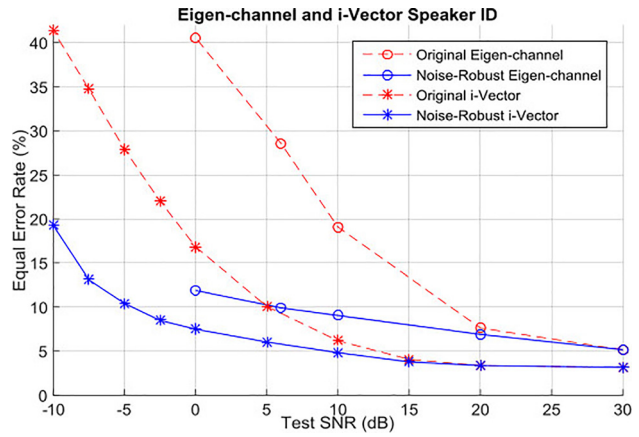

FIG. 3. (Color online) Eigen-channel and i-vector speaker ID performance.

from both genders of the aforementioned NIST SRE data[24] were used for experimentation purposes. Determining the optimum (non-equal) configuration of SNR1 and SNR2, as a function of the input test record at SNR3, a training set of 15 000 NIST SRE records were then used to yield the final EER vs SNR3 performance listed in Table IV below. Note that for the entire range of SNRs tested, the NRGID EER remained below 5%.

VI. NOISE-ROBUST LID

An i-vector extraction process was used to obtain language-prints (similar to voiceprints), enabling the comparison of these short vectors for rapid noise-robust language identification (NRLID). Multiclass logistic regression was used as a language classifier in this case. Figure 5 illustrates the processing blocks with the associated SNR definitions, similar in structure to i-vector gender classification in Fig. 4. System parameters were trained at SNR1 and were processed to extract a language-print, followed by the language classifier, using language model parameters trained at SNR2. A language and score were produced when the test record was input at SNR3.

Six languages from a pre-existing language pack, listed in Table V(a), were used for development. Using the same simulation software structure as for NRSID and NRGID, the optimum (minimum EER) combination of SNR1 and SNR2, as a function of the test SNR3 level, were determined from the ten 10×10 matrices. This decision criteria was then applied to the five target languages listed in Table V(b).

The only other difference in determining NRLID performance was that SRVAD was used for all activity

TABLE IV. Optimum EER NRGID performance for system and model SNRs vs test SNR.

Test SNR (dB)	−10.0	−7.5	−5.0	−2.5	0	5	10	15	20	>30
EER (%)	4.37	4.23	3.40	2.66	2.21	2.11	2.07	2.01	1.73	1.77

segmentation, where 20% of the lowest energy of the voice frame was included in the noise estimate needed to calculate the testing SNR. All of the speech segments were used for training. These assumptions resulted in the NRLID EER performance as a function of the test SNR, shown below in Table VI.

VII. NOISE-ROBUST DIARIZATION

The task of diarizing, or de-interleaving a noisy voice record with more than one speaker, without prior speaker training, was also considered a key component in the overall noise-robust speech triage capability. This process enables speaker separation into separate single-speaker voice records needed for speaker enrollment in NRSID. The diarization system used is based on a fully Bayesian approach that uses GMMs with eigen-voice priors. The approach[25] is very close to the vector based approach well known in speaker recognition.[17] There was no hard alignment of speaker to speech segment until the final step of the algorithm. A vector of speaker probabilities associated to each frame of speech, was estimated. The probabilities were initialized randomly and refined iteratively. The initial number of speakers was preset to be higher than the final number of expected speakers. Kenny's original work[26] was extended with a hidden Markov model layer that reduces rapid jumps between speakers.

Figure 6 illustrates this iterative structure, where (1) speaker GMM statistics are collected, (2) speaker factors (i-vectors) from these statistics are estimated, and finally (3) speaker factors are converted back to speaker GMMs, followed by a re-alignment of frames to speakers using the forward-backward algorithm. A Viterbi algorithm is used to obtain the best path through the speakers. The diarization algorithm then provides a speaker identity or no speaker decision every 10 ms. Even though the algorithm is iterative, it turned out to have a rather surprisingly fast throughput of approximately 20 × real-time, where 3 to 4 iterations were typically required for convergence.

The performance evaluation was applied to NIST SRE 2008 recordings, where the "reference" SAD yielded the

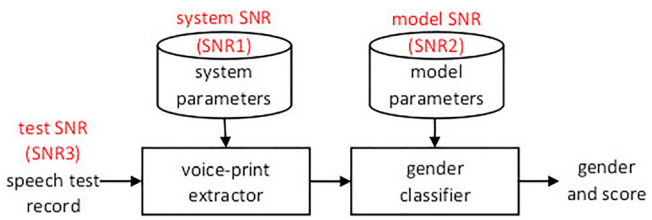

FIG. 4. (Color online) SNR definitions for NRGID.

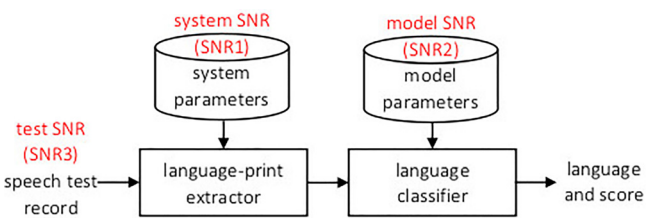

FIG. 5. (Color online) SNR definitions for NRLID.

TABLE V. Language ID data sets.

(a) Development			(b) Target		
Language	Train (h)	Test (h)	Language	Train (h)	Test (h)
Levantine Arabic	422.30	6.48	Gulf Arabic	52.23	50.20
Czech	39.50	8.24	Dari	108.67	4.65
Farsi	148.75	6.81	Hausa	56.82	56.03
Panjabi	10.47	4.88	Pashto	134.76	4.72
Thai	20.39	4.67	Somali	47.55	51.52
Urdu	95.36	5.03			

training data with the ground truth activity segmentation generated from clean data. The evaluation recordings were created by adding together two records into one channel, and then adding white noise. Care was taken so that there were no overlapping speakers in the performance evaluation records. SRVAD was then used to estimate the speech activity segmentation from these records. There are only two SNR levels at play in the diarization speaker architecture, displayed above: (1) the noise level at which the global GMM statistics are trained, and (2) the SNR of the input speech record containing two or more speakers.

The diarization error rate (DER), defined by NIST,[27] was used as the evaluation metric. During the diarization process, three types of errors can occur. The first is a speaker error in terms of the percentage of time that a speaker's identity is incorrectly assigned to another speaker, known as speaker error rate (SER). In the second instance, a miss error happens when speech is present in the reference record, but no speaker is identified. Finally, a false alarm error occurs when a speaker is identified but there is no speech activity present. The DER is the sum of all three of these error rates. Since there were only two SNR levels that could be varied, simulations to determine the optimum performance were run on a single two-dimensional (10×10) DER matrix.

Selecting the optimum system SNR settings as a function of input SNR and the minimum DER, resulted in the DER and SER vs SNR dependence, listed in Table VII and plotted in Fig. 7. It is evident that decreasing SNR increases DER monotonically. The decrease in SER below 0 dB, however, is due to worsening SRVAD performance, which is detecting less and less of the voiced speech, resulting in more correct speaker choices. This also results in increasing missed speaker and false alarm errors (both are not shown), increasing the overall DER.

When possible, normally speaker and language enrollment and training are performed with the "cleanest" speech records available. Unfortunately, these are not always available. The high DER percent levels at SNR levels below 0 and 5 dB, are not surprising, since diarization is one of the most challenging speech triage functions. Noise-robust diarization affords us the opportunity of enrolling and training the other triage functions from speech records that are less than ideal, down to SNR levels of about +5 dB, as shown in Fig. 7 and Table VII.

TABLE VI. Optimum EER NRLID performance vs test SNR.

Test SNR (dB)	−10.0	−7.5	−5.0	−2.5	0	5	10	15	20	>30
EER (%)	25.59	13.87	10.48	8.44	7.28	6.25	5.23	4.94	4.80	4.94

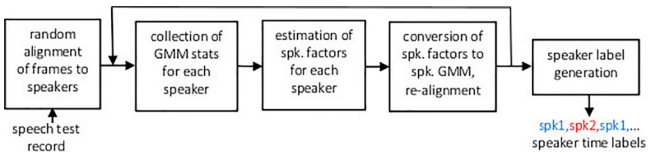

FIG. 6. (Color online) Diarization system architecture.

VIII. CONCLUSIONS AND RECOMMENDATIONS

We have presented a unified approach to classification of speech in noise. In our approach, we have demonstrated that i-vector-based SID, LID, and GID as well as diarization may be significantly improved by testing on models trained at appropriate SNR levels. Our i-vector work is based on our previous eigen-channel work in which it was found that the SID performance was greatly improved when the SNR level of the training data was nearly matched to the SNR of the test data.[12] In the present i-vector study, classification was found to be dependent on the SNR levels of the global models, the enrollment models and the test data. These SNR levels are independent, and the best performance does not occur when all three signal qualities are equal. Global models and enrollment models were generated for a lattice of SNR levels between −10 dB and +30 dB. For test signal quality at any given SNR, the best performance was achieved using the appropriate look-up tables that select simulation-based optimal combinations of global, testing, and enrollment SNR levels. Unlike other proposed robust SID, LID, GID, and diarization methods,[1–11] the methods we propose require no modification of the actual classification method itself. Our method only requires that global and enrollment models be generated for a lattice of SNR levels.

A critical component is speech activity detection which is used to separate segments of speech from segments of non-speech and to estimate SNR. Of two SAD algorithms used, the SRVAD algorithm, based on the signal envelope, does not attempt to remove or mitigate specific (telephony, non-white car, babble, etc.) types of noise or interference.[22] It may be possible to improve the performance of both noise-robust speech activity detection and speech classification functions in other environments by

TABLE VII. Noise-robust diarization performance (in terms of DER and SER) vs test SNR.

Test SNR (dB)	−7.5	−5.0	−2.5	0	5	10	15	20	>30
SER (%)	4.12	10.78	15.78	16.36	13.23	9.81	7.35	5.89	3.69
DER (%)	78.89	60.59	43.73	33.47	22.25	15.02	10.55	7.87	4.85

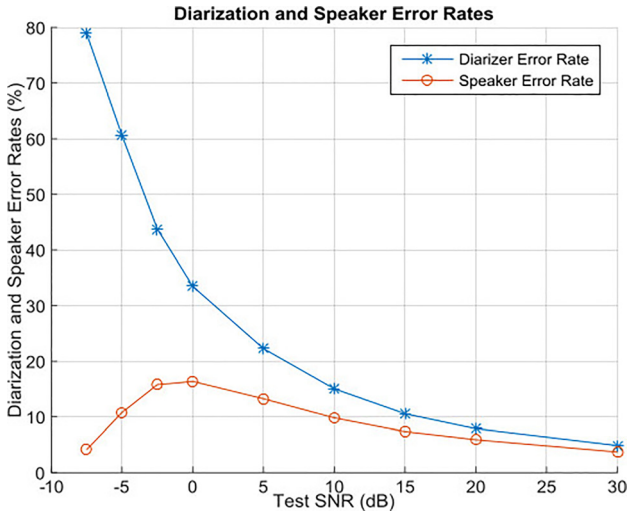

FIG. 7. (Color online) Noise-robust diarizer performance.

suppression or mitigation of reverberation, car, babble, and other types of noise. This feature is partially implemented in the PVAD and SRSAD[21] algorithms.

Figure 3 illustrates how SID noise-robustness has evolved from Eigen-channel unmodified performance (dashed circles) prior to 2008,[15] to the current SNR-lattice of i-vector SID models (solid*), where a 19% EER performance of the original and the evolved SID operations is maintained for a 20 dB reduction in SNR (from +10 dB to −10 dB). As was the case with SID, the EER of the other i-vector classification tools, dropped significantly when applying the SNR-lattice-based approach. At the lowest SNR (−10 dB), the EER exceeded 40% for all of the unmodified i-vector GID, SID, and LID models. This suggests that any MFCC-based speech classification function could realize increased classification performance in noise, if an SNR-lattice approach were used. Figure 8 summarizes the FOC Noise-robust SID, GID and LID performance. Diarization performance is shown in Fig. 7.

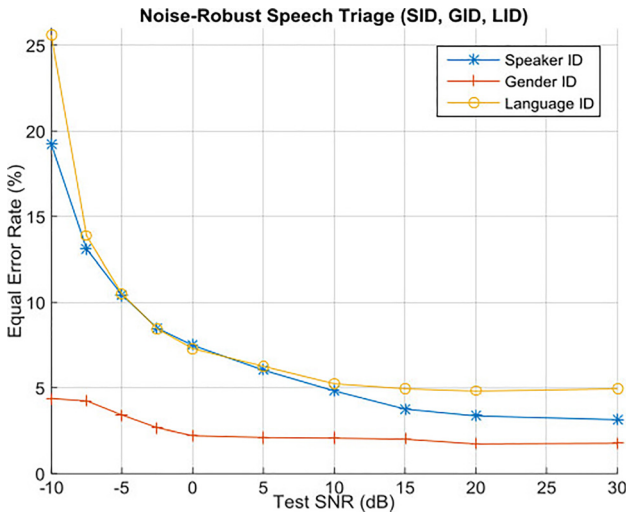

FIG. 8. (Color online) Noise-robust speech triage.

[1]P. Jancovic and M. Köküer, "Employment of voicing information of speech spectra for noise-robust speaker identification," in *15th European Signal Processing Conference (EUSIPCO)* (2007), pp. 2399–2403.

[2]K. Kumar, Q. Wu, Y. Wang, and M. Savvides, "Noise robust speaker identification using Bhattacharyya distance in adapted Gaussian models space," in *16th European Signal Processing Conference (EUSIPCO)* (2008), pp. 1–4.

[3]K. Matsumoto, N. Hayasaka, and Y. Iiguni, "Noise robust speaker identification by dividing MFCC," in *6th International Symposium on Communication, Control and Signal Processing (ISCCSP)* (2014), pp. 652–655.

[4]C. Tzagkarakis and A. Mouchtaris, "Reconstruction of missing features based on a low-rank assumption for robust speaker identification," in *IISA, The 5th International Conference on Information, Intelligence, Systems and Applications* (2014), pp. 432–437.

[5]Z. Tan and M. Mak, "Bottleneck features from SNR-adaptive denoising deep classifier for speaker identification," in *Proceedings of APSIPA Annual Summit and Conference* (December, 2015), pp. 1035–1040.

[6]A. K. Dutta and K. S. Rao, "Robust language identification using power normalized cepstral coefficients," in *Eighth International Conference on Contemporary Computing (IC3)* (2015), pp. 253–256.

[7]S. Ganapathy, M. Omar, and J. Pelecanos, "Noisy channel adaptation in language identification," in *IEEE Spoken Language Technology Workshop* (2012), pp. 307–312.

[8]M. K. Rai, N. Fahad, M. S. Fahad, J. Yadav, and K. S. Rao, "Language identification using PLDA based on I-vector in noisy environment," in *International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Jaipur, India (September 2016), pp. 21–24.

[9]S. Ranjan, G. Liu, and J. H. L. Hansen, "An I-vector PLDA based gender identification approach for severely distorted and multilingual DARPA RATS data," in *ASRU* (2015), pp. 331–337.

[10]S. M. Mirrezaie, S. M. Ahadi, and A. Kashi, "Robust speaker diarization in a multi-speaker environment using autocorrelation-based noise subtraction," in *IEEE International Symposium on Signal Processing and Information Technology* (2007), pp. 291–296.

[11]Q. Li, Q. Fan, Y. Xiao, and W. Ye, "A comparable study on PNCC in speaker diarization for meetings," in *First ACIS International Symposium on Cryptography and Network Security, Data Mining and Knowledge Discovery, E-Commerce & Its Applications and Embedded Systems (CDEE)* (2010), pp. 157–160.

[12]A. L. Bartos and D. J. Nelson, "Enabling improved speaker recognition by voice quality estimation," in *IEEE 45th Asilomar Conference on Signals, Systems and Computers* (2011), pp. 595–599.

[13]D. A. Reynolds, "A Gaussian mixture modeling approach to text-independent speaker identification," Ph.D. thesis, Georgia Institute of Technology, Atlanta, GA (September, 1992).

[14]D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," in *IEEE Transactions on Speech and Audio Processes* (January, 1995), Vol. 3, pp. 72–83.

[15]http://www.clsp.jhu.edu/vfsrv/workshops/ws08/groups/rsrovc/index.html (Last viewed April 15, 2015).

[16]P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigen-channels in speaker recognition," in *IEEE Transactions on Audio, Speech, and Language Processes* (May 2007), Vol. 15, pp. 1435–1447.

[17]N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Oellet, "Front-end factor analysis for speaker verification," in *IEEE Transactions on Audio, Speech, and Language Processes* (May 2011), Vol. 19, pp. 788–798.

[18]O. Glembek, L. Burget, P. Matějka, M. Karafiát, and P. Kenny, "Simplification and optimization of I-vector extraction," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (2011), pp. 4516–4519.

[19]P. Matějka, O. Glembek, F. Castaldo, J. Alam, O. Plchot, P. Kenny, L. Burget, and J. Černocký, "Fullcovariance UBM and heavy-tailed PLDA in I-vector speaker verification," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, Prague (2011), pp. 4828–4831.

[20]J.-H. Bach, B. Kollmeier, and J. Anemuller, "Modulation-based detection of speech in real background noise: Generalization to novel background classes," in *Proceedings of IEEE International*

Conference on Acoustics, Speech, and Signal Processing, ICASSP (2010), pp. 41–44.
[21]D. C. Smith, J. Townsend, D. J. Nelson, and D. Richman, "A multivariate speech activity detector based on syllable rate," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP* (1999), pp. 73–76.
[22]A. L. Bartos, "Syllable rate voice activity detection (SRVAD) algorithm documentation," version 2, prepared for Naval Research Laboratory (NRL), October 22, 2012, Contract No. N00173-05-C-2049, available from NRL John.Banowetz@nrl.navy.mil.
[23]J. D. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE 11th International Conference on Computer Vision* (2007), pp. 1–8.
[24]http://www.nist.gov/itl/iad/mig/sre10results.cfm (Last viewed April 15, 2015).
[25]P. D. Kenny and F. Castaldo, "Diarization of telephone conversations using factor analysis," IEEE J. Select. Top. Sign. Process. **4**(6), 1059–1070 (2010).
[26]P. Kenny, "Bayesian analysis of speaker diarization with eigenvoice priors," technical report, CRIM, Montreal (May 2008).
[27]http://www.xavieranguera.com/phdthesis/node108.html (Last viewed April 15, 2015).