Int J Speech Technol
DOI 10.1007/s10772-012-9182-0

Environment dependent noise tracking for speech enhancement

Nitish Krishnamurthy · John H.L. Hansen

Received: 21 July 2012 / Accepted: 10 November 2012

Abstract Numerous efforts have focused on the problem of reducing the impact of noise on the performance of various speech systems such as speech recognition, speaker recognition, and speech coding. These approaches consider alternative speech features, improved speech modeling, or alternative training for acoustic speech models. This study presents an alternative viewpoint by approaching the same problem from the noise perspective. Here, a framework is developed to analyze and use the noise information available for improving performance of speech systems. The proposed framework focuses on explicitly modeling the noise and its impact on speech system performance in the context of speech enhancement. The framework is then employed for development of a novel noise tracking algorithm for achieving better speech enhancement under highly evolving noise types. The first part of this study employs a noise update rate in conjunction with a target enhancement algorithm to evaluate the need for tracking in many enhancement algorithms. It is shown that noise tracking is more beneficial in some environments than others. This is evaluated using the Log-MMSE enhancement scheme for a corpus of four noise types consisting of Babble (BAB), White Gaussian (WGN), Aircraft Cockpit (ACN), and Highway Car (CAR) using the Itakura-Saito (IS) (Gray et al. in IEEE Trans. Acoust. Speech Signal Process. 28:367–376, 1980) quality measure. A test set of 200 speech utterances from the TIMIT corpus are used for evaluations. The new *Environmentally Aware Noise Tracking* (EA-NT) method is shown to be superior in comparison with the contemporary noise tracking algorithms. Evaluations are performed for speech degraded using a corpus of four noise types consisting of: Babble (BAB), Machine Gun (MGN), Large Crowd (LCR), and White Gaussian (WGN). Unlike existing approaches, this study provides an effective foundation for addressing noise in speech by emphasizing noise modeling so that available resources can be used to achieve more reliable overall performance in speech systems.

Keywords Noise · Speech · Speech enhancement · Noise tracking

1 Introduction

One of the main factors limiting performance of speech systems is acoustic/environmental noise. Efforts to alleviate this factor have been historically addressed under the area of speech enhancement. Specifically, estimating the changing noise parameters within speech environments with accuracy and speed have been studied in the domain of "Noise Tracking". The main focus of this study is to demonstrate the necessity and advantages of environment specific noise tracking solutions. For example, noise tracking requirements for highly time varying noise types like babble noise are different than a stationary noise scenario like car noise. In babble, statistical properties of noise change rapidly with time depending on the number of speakers constituting babble. In this scenario, the focus is to track the variation of the noise floor with time, minimizing the time lag. Conversely, noise in a car has fewer time dependent parameters and these variables vary slowly with time. In a car, the focus of noise tracking is to have an accurate representation of the frequency content of the noise with time. In this scenario, the speed of noise tracking becomes a secondary concern.

N. Krishnamurthy · J.H.L. Hansen (✉)
Center for Robust Speech Systems, University of Texas at Dallas, Dallas, TX, USA
e-mail: john.hansen@utdallas.edu

Published online: 13 December 2012

As evident from this discussion, the motivation for the proposed approach is that some noise environments are easier to predict and estimate than others due to their relatively slow time varying structures. As the time variability of noise increases, the focus of noise tracking changes from accuracy of the frequency content evaluation to more careful assessment of the time varying the noise floor. The focus of this study is to demonstrate the necessity for tracking solutions tailored to highly time varying noise.

There are two main parts of this study: first, characterization of the effect of time varying noise on speech enhancement performance. Using improvement in enhancement quality as a function of noise update rate, it is demonstrated that tracking can be extremely beneficial for highly time varying environments. Second, based on this observation, a heuristic environment tracking solution is developed for tracking in time varying environments.

This study proposes the use of noise update rate for speech enhancement to characterize the noise environment. The noise update rate is used to parametrize the time varying noise in terms of enhanced speech quality. The dependence of enhancement quality on update rates is performed using dual channel systems, where one channel is normally dedicated to noise estimates for enhancing the speech and the other channel is the noisy speech. We note that the tracking solutions focus on single channel conditions. This dual channel condition simulates the best possible tracking scenario where the estimated noise is the exactly the same as the degrading noise. This allows the parameterization of the effectiveness of tracking by varying the estimation rate. Under non-stationary conditions, frequent noise updates are required to achieve an effective estimation of the noise spectral structure. Conversely, stationary noise conditions require fewer estimates across time. This strategy was studied in Krishnamurthy and Hansen (2006) to predict the output enhancement quality for a given enhancement scheme for a given environment. Here, it is used to identify the effectiveness of tracking for individual environments.

After demonstrating the environment dependent benefits of tracking, a model based tracking scheme is proposed for superior performance in these environments. This is achieved by first parameterizing the impact of noise on speech for a given environment using statistical models, and then using these models to predict the noise in a particular frame during speech system deployment. This is different from previously proposed work on noise tracking (Sect. 2) as this strategy is the first to actively incorporate off-line environment information for noise tracking. This is especially useful in conditions where there is a sudden burst of background noise or there is rapid changes in background with respect to speech. This differs from most contemporary noise tracking schemes which are designed with the assumption that noise changes slowly as compared to speech (Sect. 2). Previous approaches do not work for environments which change at a rate that is either comparable (babble) or greater (machine-gun) than the time rate of speech. The basis of the proposed approach is that noise and noise-speech interaction in an environment can be statistically characterized over a period of time prior to noise tracking. During tracking, the pre-gathered information concerning the environment can be used. The proposed approach uses a pre-observed noise frame from earlier knowledge of the environment, or a noise reservoir of the signal as a noise estimate. Using the noise-only parts enables construction of a degraded speech model with an available clean speech side-corpus. When a noisy speech frame is observed, the closest matching degraded frame from the database is searched and the pre-observed noise used to degrade this frame is employed as the noise estimate.

This method relies more on learning the process signature than statistically characterizing the noise, and hence it is not impacted by the non stationary nature of noise. Noisy speech frames are used to reconstruct a “speechy noise” frame, from which a noise estimate is constructed. This method works especially well in scenarios such as babble noise, where the implied assumption is that the speakers in the background do not change with time. Similarly, for impulse period noise types such as a jackhammer or machine gun noise, the signature of the device does not change with time allowing us to obtain an accurate estimate every frame. Conversely, stationary and slowly varying environments (e.g., white noise, pink noise) do not require a noise estimate every frame. By employing intelligent noise estimate/update rates, it is possible to conserve overall computational resources. This is extremely important for mobile devices requiring small footprint speech applications.

It should be noted that the focus of this paper is not to provide a better enhancement solution, but to provide methodologies for incorporating environment information into speech systems. The applications discussed in this study are examples where speech enhancement benefits from extracted environment specific information. These examples can easily be extended to other speech applications for robustness in varying environmental conditions since the estimation of background noise parameters and the rate of background update are relevant information for all practical speech systems (e.g., coding, speech recognition, speaker ID, etc.).

This paper is divided into three core phases; Sect. 2 deals with approaches utilizing environment information in various speech systems and considers previous research on noise tracking for enhancement. Section 4 describes the proposed algorithms for noise tracking and update rate has been elaborated in (Sect. 3). These algorithms are evaluated for enhancement along with comprehensive testing across different noise conditions in Sect. 5. The last part of this study considers future applications of the formulated noise engineering framework.

2 Previous research

2.1 Environmentally aware speech systems

The most recent approaches to incorporate environment information in speech systems include performing condition dependent model evaluations as proposed by Xu et al. (2007, 2006). Specific to the vehicle environment, the knowledge of vehicle-events was leveraged in noisy ASR, where vehicle event specific acoustic-models were dynamically chosen for decoding noisy speech (Environmental Sniffing by Akbacak and Hansen 2007). Other approaches based on classification of the environmental noise for better performance include Kates (1995), where environment information is used to improve hearing aid performance, as well as Ma et al. (2003) who performed acoustic background noise classification for generic context aware applications. In the car environment, El-Maleh et al. (1999) proposed frame level noise classification for mobile acoustic environment, with focus on speech coding.

A main area of application for environment aware speech systems is noise tracking. Here the power density spectrum of noise is updated in the absence of speech. The purpose of noise tracking is to estimate the noise in those parts of the input where the speech "corrupts" the noise signal. The best possible estimate of noise is needed to achieve an effective enhancement solution. Let y denote the received speech signal, n and s denote the noise and speech components of the signal respectively. Under an additive noise assumption, this can be written as,

$$y = s + n. \tag{1}$$

If we assume further that the noise and speech are statistically uncorrelated and orthogonal with either the speech or noise being a zero mean process, the autocorrelation can be written as,

$$R_{yy}(\tau) = R_{ss}(\tau) + R_{nn}(\tau), \tag{2}$$

where, R_{yy}, R_{ss}, and R_{nn} are the received signal, speech and noise autocorrelations respectively. If this is a function of time, the above equation can be written as,

$$R_{yy}(\tau, t) = R_{ss}(\tau, t) + R_{nn}(\tau, t). \tag{3}$$

Noise tracking requires that we estimate $R_{nn}(\tau, t)$ as a function of time. There have been many approaches to this problem. An overview of the prevalent algorithms has been described in Fukane and Sahare (2011). One of the simplest approximations being the approach where $R_{nn}(\tau, t)$ is a linear function in time, where $S_{begin}(k)$ and $S_{end}(k)$ denote the known power density spectra at the beginning and end of the utterance, and k is the FFT bin number. With this, the intermediate power-density spectra at time index i can be linearly estimated using the relation,

$$S_i(k) = \frac{S_{end}(k) - S_{begin}(k)}{N} \cdot i + S_{begin}(k). \tag{4}$$

Another approach to this problem utilizes the fact that the power of the degraded speech is always greater than the power of the noise only part of the signal. Since speech is an intermittent, time varying signal during voice communications, the noise can be tracked for short durations when the speaker pauses by tracking the minimum over a window of time,

$$S_i(k) = \min\{S_{i-L}(k), \dots, S_{i+L}(k)\}. \tag{5}$$

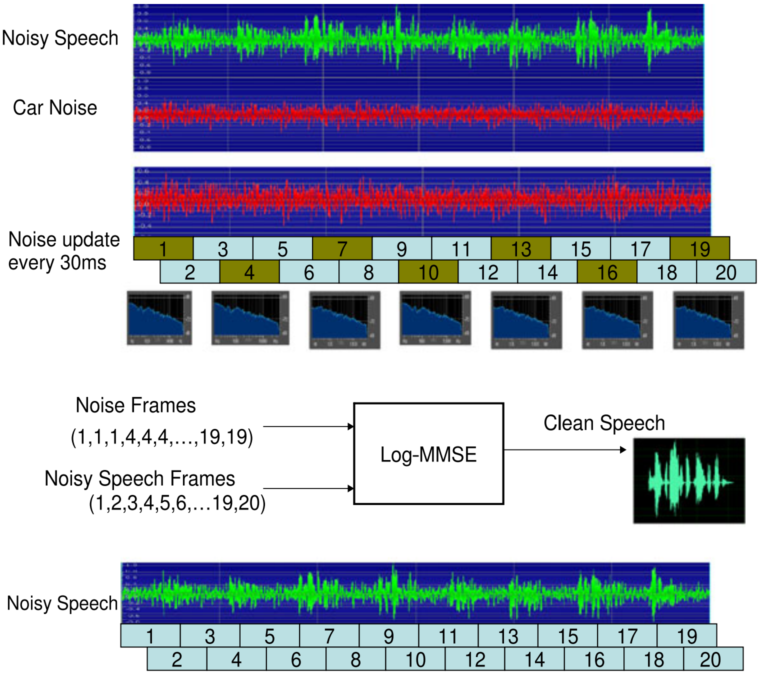
The minima of noise is tracked over a window length of $2L + 1$ about the target frame. This approach for noise tracking was first proposed by Martin (1994). Later, Cohen (2003), proposed approaches where the power spectral densities were weighted using speech presence probabilities before they were used to decide the minimum across the time frames. There has been much work towards obtaining an accurate estimate of the smoothing terms for the recursive estimation of noise and estimation of the signal presence probabilities, as noted by Cohen (2003). Rangachari and Loizou (2006), proposed advancements over the MCRA scheme that adapts faster to changing noise levels. This approach was further extended by Hendriks et al. (2008) where they performed minima tracking on an eigen decomposition subspace instead of the FFT bins. Examples of approaches that use decompositions other than FFT include Chatlan and Soraghan (2009).

All the above cited methods are based on the premise that noise changes slowly compared to the change in speech phoneme rate over a window length. This study analyzes the dependence of the noise type on speech enhancement systems and proposes a noise tracking solution for extremely time varying noise solution. The literature in noise and speech is varied and rich, yet, environment dependent processing solutions are not popular due to various reasons. The next section describes Update Rate and uses it as a measure of noise variability.

3 Environment evaluation using update rate

To analyze the time varying nature of noise in the context of a speech enhancement system, the noise update rate required to maintain a given speech quality is employed. To demonstrate its usefulness as a measure of noise variability, it should be noted that as variability noise increases, the frequency of updates required to achieve a given quality of enhanced speech increases. Another way of looking at noise

Fig. 1 Description of the Noise Modeling framework. Here, one noise update every three frames from the noise only channel is used to enhance speech in the primary channel



update rate is that, for an increase in the noise update frequency, the improvement in enhancement quality in babble noise would be greater than in the car noise scenario. The previous example demonstrates the utility of noise update rate required for a given quality of enhancement as a parameter to describe the noise environment.

To demonstrate the dependence of noise update rate on noise type, a model is constructed by varying the noise update rate against a resultant enhancement quality measure for different noise types. This procedure is carried out across different SNR values using speech degraded under 4 different environmental conditions. Figure 1 describes the noise update rate process. Here, the first/primary channel consists of noisy speech in a car environment at a given SNR. The second/secondary channel contains the car noise instance used to obtain the noisy speech in channel 1 by mixing speech at the required SNR. For the purpose of this evaluation, a dual channel system is assumed, with one channel containing only noise and the second containing the noise degraded speech. The noise estimates from the second channel are used to enhance the speech in the primary channel. This evaluation framework assumes a single channel environment with the second channel being the golden reference of the noise estimate to evaluate the impact of update rate on speech enhancement. The noise frames from channel 2 are used in a sample and hold mode for a given update frequency. This process ensures that the noise channel and the noisy speech are in sync. The noise frames and the speech frames are then used by the LogMMSE process to enhance the noisy speech. The enhanced speech is used to evaluate the impact of the noise update rate on the speech-quality for a given noise environment. An average objective measure of speech enhancement across a corpus is then plotted as a function of noise update rate. These plots are indicative of the relative stationarity of the noise. For a stationary noise type, it is expected that the increase in update rate does not correspond to a linear increase in the speech quality as opposed to non-stationary noise where frequent noise updates is expected to result in further enhanced speech.

3.1 Noise update rate based assessment of noise properties

In this section, the above methodology is used to evaluate a corpus of noise types for their time variability with respect to speech signals. These noise types are aircraft cockpit, multi-speaker babble, stationary car, and white Gaussian noise (AIR, BAB, HWY, WGN). The noise types under consideration are selected because of their varying degree of stationarity and their spectral properties. The degraded utterances are enhanced using the log-MMSE algorithm. The noise updates for the log-MMSE algorithm are performed using noise from the second channel (i.e., to ensure an even noise frame update process). The average IS (Itakura-Saito) (Rabiner and Schafer 1978) measure is calculated for a set 192 phonetically balanced utterances from the TIMIT corpus. This evaluation is performed for different noise update-rates. From these experiments an enhanced vs. noise type vs. noise update-rate model space is obtained. The same procedure is carried out for a range of SNR values. This results in the final update rate model for each of the given environmental conditions. These data points are interpolated to obtain an estimate of the model for the given SNR conditions.

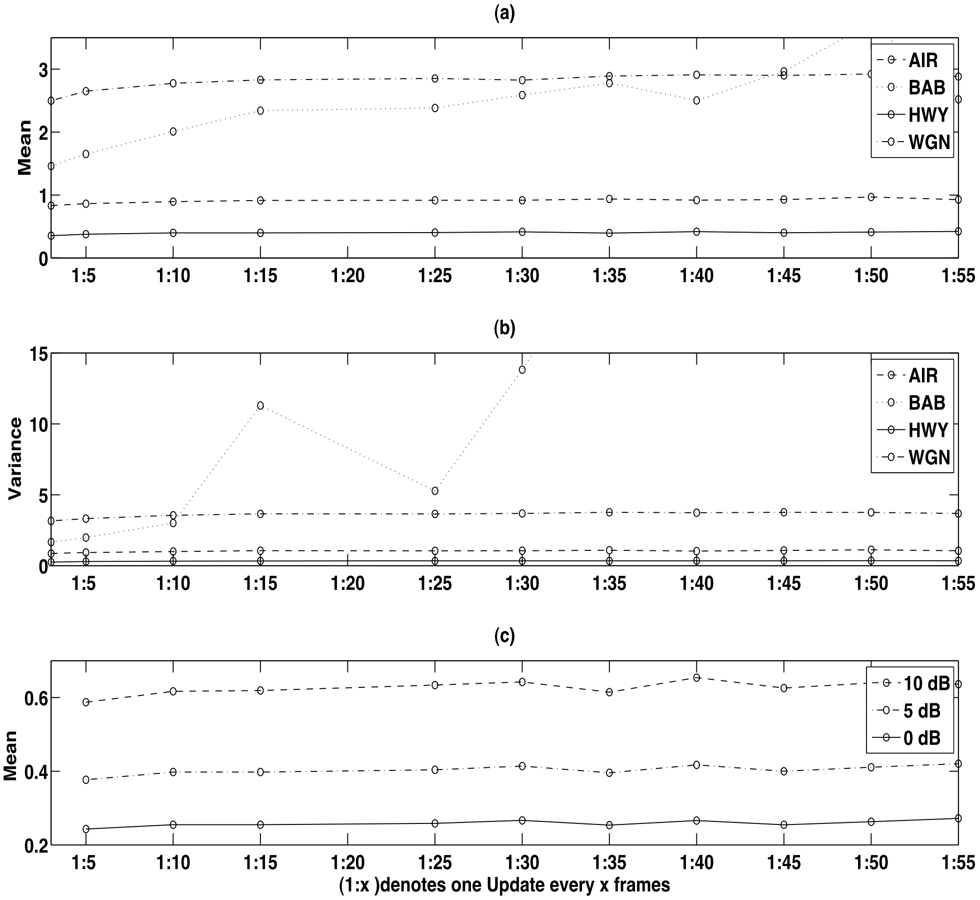


Fig. 2 (**a**) Update rate model showing the mean for white Gaussian noise, babble noise, aircraft cockpit noise and stationary car noise (**b**) the variance of the IS measures for different update rates, and (**c**) the update-rate Vs. IS measure model for aircraft cockpit noise for different SNR values 0 dB, 5 dB and 10 dB. The x-axis origin denotes 1:3 condition i.e. 1 update every 3 frames

Given the update rates, the corresponding IS measures can be estimated using this noise model. This section describes the procedure used to compute the appropriate frame rates.

Here, a 20 msec frame size was used along with a Hamming window, and the noise estimate was calculated as the magnitude square of the Fourier transform. The speech and noise signals were sampled at 16 kHz and PCM encoded. Figure 2(a) illustrates the mean IS values after enhancement across all frames in the 192 sentence set for different noise update rates with noise updates at a frequency of 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 frames respectively. This is represented as (1:5) which corresponds to one spectral frame spectral frame update every 5 frames. The variance in IS measure across all frames for different update rates is shown in Fig. 2(b) and the variation of mean is shown in Fig. 2(a). Figure 2(c) shows the mean IS values after enhancement for the aircraft cockpit noise model for update rates at SNR values of 0 dB, 5 dB, and 10 dB. A total of 24000 sentences across different SNR's and noise types were enhanced during the process of obtaining the update rate dependency in Fig. 2. The **Stationary Car Noise** is a low frequency restricted bandwidth noise from a Chevy SUV Blazer traveling on a highway at a speed of 65 mph. As observed from the noise model plots, an increase in noise update rate does not result in improved IS measure after enhancement. It should also be noted that update rates as low as one spectral frame response every 1000 ms (1:50 frames) give the same resulting quality of enhancement as one update every 100 ms (1:5 frames). This implies that an update rate of one update every 1000 ms is sufficient to characterize this noise. The **Aircraft Cockpit Noise** shows slightly more variability. This noise was recorded in a Lockheed C130 transport plane flying at 25,000 ft and is almost as stationary as the car noise. The major differences in the IS measure plots are due to the difference in noise bandwidths. The shape of the model plot is similar to that obtained for stationary car noise, however, the level is elevated. This is due to higher degradation of speech spectral structure. **White Gaussian Noise** is the most stationary of all the noise types considered. This noise type has the largest bandwidth of degradation in the structure of speech due to its full band spectrum. These observations are reflected in the IS measure plots, since they are flat but are shifted vertically, showing more degradation than AIR or HWY. Since the **Babble Noise** is the most time varying and has spectral properties similar to speech, it has the highest dependency on the update-rate parameter. This noise type shows a general rising trend (i.e., as the update rate is reduced the resulting enhancement suffers).

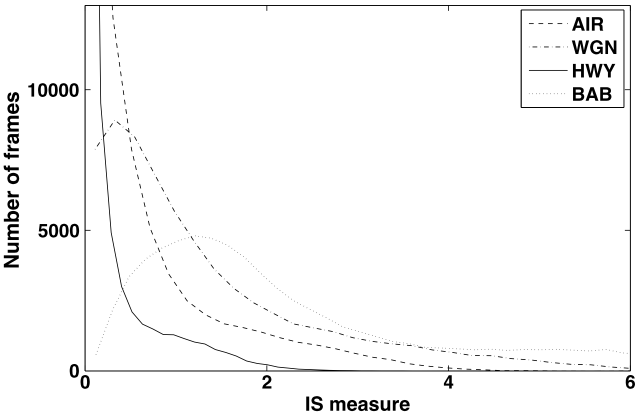


Fig. 3 Distribution of IS measures for speech files degraded with different noise types. The pdfs with heavy tailed distributions represent time varying noise types

Another interesting study is the variability within noise itself. This can be evaluated by observing the distribution of the IS measures within all frames for a noise type. For the four noise types under consideration this is illustrated in Fig. 3. The peakiness of the distribution is a direct indication of the amount of stationarity since, for stationary noise types the distance metrics would be clustered up together. These can also be used to evaluate the relative effect of noise on the speech utterance by comparison of the heaviness of the tail. As seen from the figures, stationary car noise and aircraft noise have shorter tails than the babble and White Gaussian noise types.

As evident from the above analysis, the improvement obtained from noise tracking increases with an increase in the non-stationary nature of noise. The above observations were used by Krishnamurthy and Hansen (2006) to estimate the expected speech quality for a given environment for a particular speech enhancement solution. Here, the authors also noted that the variance of errors increases with the increase in the non-stationary noise of the environment. As evident from the above analysis, it is evident that specific environment dependent processing can be leveraged for greater benefit in non-stationary environment types. This observation motivates the development of noise tracking solutions for highly non stationary noise environments.

4 Noise tracking

The previous section demonstrated the necessity of specific noise tracking solutions for highly time varying noise sources. As the rate of change of noise becomes closer to that of speech, the existing noise tracking solutions become ineffective as they are based on the assumption that the noise statistics vary slowly relative to the speech statistics. This leads to a very low noise floor for highly time varying noise types. In this section, a novel noise tracking solution is proposed that is based on statistically learning the noise patterns and then reusing the noise patterns during the noise tracking process. This approach leverages the fact that it is possible to obtain noise only section of any environment. Using these noise only sections, the non-stationary nature of the environment can be re-used for noise tracking. Previously available information from the environment, or noise available from a reservoir surrounding the speech utterance is used to create the noise statistical models. This noise reservoir is used to statistically model speech degraded with additive noise for a particular environment. The closest frame to the degraded speech at the current frame (which we call the target frame), is used to find the closest match between the training data to the test data. Next, the noise used to degrade this training frame is employed as a noise estimate for the current test frame. The noisy speech is given by,

$$y[n] = s[n] + d[n], \tag{6}$$

where, $y[n]$ is a frame of the received signal and $s[n]$, $d[n]$ are the speech and noise signal respectively. The power density spectrum is calculated using the assumption of a zero mean noise process that is independent of speech.

$$|S_y[\omega]|^2 = |S_s[\omega]|^2 + |S_d[\omega]|^2. \tag{7}$$

Here, let $\hat{d}[n]$ be an estimate of $d[n]$ such that we minimize,

$$\arg\min|\vec{\hat{y}} - \vec{y}|^2, \tag{8}$$

where,

$$\hat{y}[n] = \hat{s}[n] + \hat{d}[n], \tag{9}$$

where $\vec{\hat{y}}$ and $\vec{y}$ are extracted features from the test and target frames. The extracted features are used to reduce the dimensionality of the data. Furthermore, to increase efficiency of the process, the data is clustered into predetermined groups and after assigning the current test frame to a cluster, where the closest matching frame within the cluster is determined. In the proposed setup described in Fig. 4, 19-dimensional MFCCs are used as the feature vectors. MFCCs were chosen since they have been shown to perform well under most classification tasks for speech. Since no direct one-to-one mapping exists from the MFCC to the signal, the MFCCs are tagged along with the noise belonging to that frame and the noisy frame itself. The overall algorithm process is described using the following pseudo code:

Step 1: Extract the noise only parts from the noisy speech signal.

Step 2: Use this noise data to degrade a secondary clean speech data corpus. Save this degraded data corpus. This speech corpus could be from the same speaker or from a general pool of speakers.

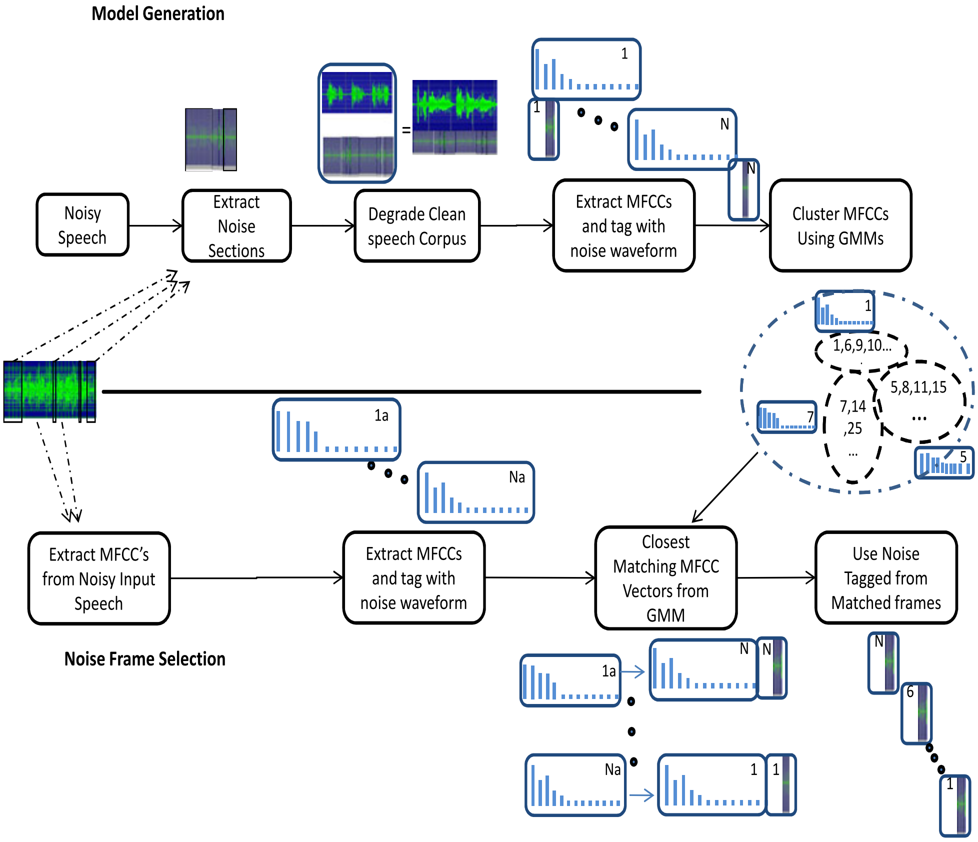

Model Generation
Noisy Speech
Extract Noise Sections
Degrade Clean speech Corpus
Extract MFCCs and tag with noise waveform
Cluster MFCCs Using GMMs
1
N
1,6,9,10...
5,8,11,15
7,14 ,25
7
5
1a
Na
Extract MFCC's from Noisy Input Speech
Extract MFCCs and tag with noise waveform
Closest Matching MFCC Vectors from GMM
Use Noise Tagged from Matched frames
Noise Frame Selection
6

Fig. 4 The Proposed Noise Tracking Algorithm tracks the noise by creating a "speechy noise" corpora and selecting noise from the closest matching noisy frame

Step 3: For each window frame, extract the feature vector and retain the noise signal used to degrade the secondary speech frame.

Step 4: Cluster all the extracted features from the secondary degraded speech into 128 clusters/mixtures using a GMM (Gaussian Mixture Model) (Reynolds and Rose 1995).

Step 5: Extract the feature vectors from the input noisy speech.

Step 6: For each feature vector, find the most likely GMM mixture component and within this mixture component, find the nearest degraded MFCC vector to this particular feature vector (target degraded speech) using the Euclidean distance.

Step 7: Employ the noise that was used to degrade the matching target degraded speech frame as the present noise estimate.

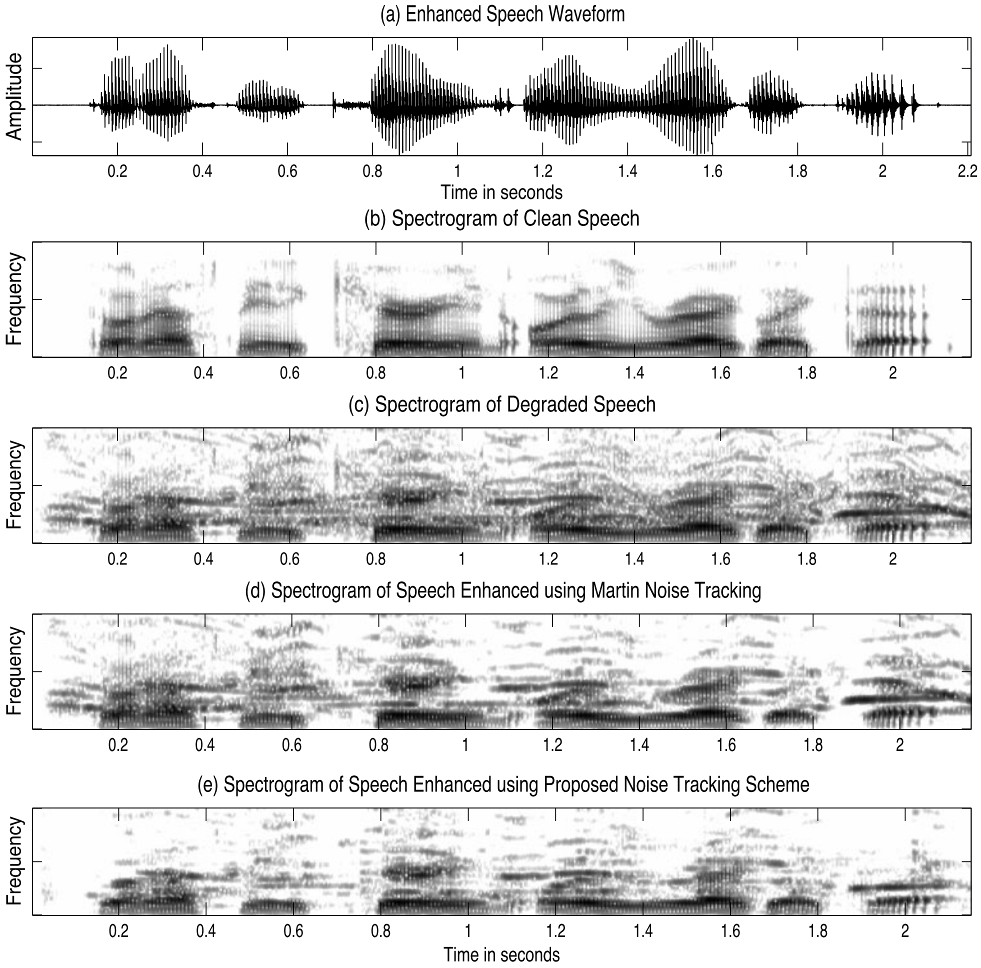
Here, a GMM was chosen over other clustering methods due to the availability of the second order statistics provided by the GMM structure which can be used in other applications. The next section describes the noise update rate evaluation algorithm and its applications to speech enhancement. An evaluation of the performance of the two schemes proposed and a comparison with available approaches is also presented.

5 Analysis and results

5.1 Noise tracking

Experimental setup To evaluate the proposed noise tracking algorithm, a single test file is degraded from the 192 TIMIT core test sentence set. Only male speakers are used for constructing the models. Babble noise at an SNR of 5 dB is used to degrade the speech file. To ensure that the test and train noise sequences are not the same, different noise observations are used. A sample from a large crowd "booing" has been employed for the experiment because of its non-stationary nature and potential impact on speech enhancement algorithms. The degraded speech frames are clustered into a 128 mixture GMM. A noise tracking algorithm is used to estimate the noise distortion in the speech degraded sections on a per frame basis. This noise estimate is used as the true noise to enhance the speech using the Log-MMSE algorithm (Ephraim and Malah 1985). The Log-MMSE scheme is chosen to emphasize the impact of the noise tracking problem on a traditional, well accepted method. In the enhancement process, the noisy frame of speech is enhanced using a noise estimate and an SNR estimate that is computed by the algorithm. We use the computed noise frame as the reference for SNR computation and speech enhancement. Figure 5 shows the re-

Fig. 5 (**a**) Waveform of clean speech and spectrograms of (**b**) clean (**c**) degraded, and (**d**) enhanced speech using Martin noise tracking, and (**e**) proposed noise tracking scheme



sulting waveforms and spectrograms of degraded and enhanced speech corrupted with babble noise. For comparison, the noise tracking methods from Martin (1994) and Cohen (2003), are compared in Fig. 5d, along with the new noise tracking method (Fig. 5e). It can be seen that the speech has been enhanced under extremely noisy conditions. After the enhancement process, a perceived level of music-like artifacts are present in the background, which are believed to be residual formants of the corrupting babble noise persisting after log-MMSE enhancement. These musical artifacts have been studied in detail in Cappe (1994), where it has been noticed that the perception of noise increases in lower SNR speech enhancement conditions. Again, our focus here is not to formulate a better enhancement algorithm but to formulate a better means of modeling and tracking noise across time by focusing on noise properties. It can be seen that the speech portions of the original signal have been preserved after enhancement. There are some artifacts in the beginning silence section of the utterance which are believed to be due to non-matching noise frames in the test and train section.

Results Having illustrated the performance of the proposed noise tracking algorithm for a single sentence, we now turn to a more comprehensive evaluation over a larger corpus to illustrate that the method scales up to general speech applications. Furthermore, the noise tracking algorithm was evaluated under three noise conditions including: LCR (large crowd noise), BAB (Babble noise), and MGN (Machine Gun Noise). These noise types have different levels of stationarity. BAB and MGN are non-stationary noise types whereas LCR is more stationary.The time varying noise characteristics can be visualized in Fig. 6. These plots describe the time evolution of the noise power spectral density. White noise has the least amount of time varying characteristics and machine gun noise is the most time varying.

These were used to degrade TIMIT sentences at SNR levels of -5 dB, 0 dB, and 5 dB. For these noise types, different noise samples were used for training and test phases. A set of 192 sentences were randomly chosen from the TIMIT corpus that were different from those used for training. A total of 6912 sentences were used to obtain the results (192 sentences × 3 SNRs × 3 noise types × 4 algorithms). The Itakura-Saito (IS) distance measure was used to assess objective speech quality performance. As seen in Table 1, the proposed noise tracking scheme either measurably outper-

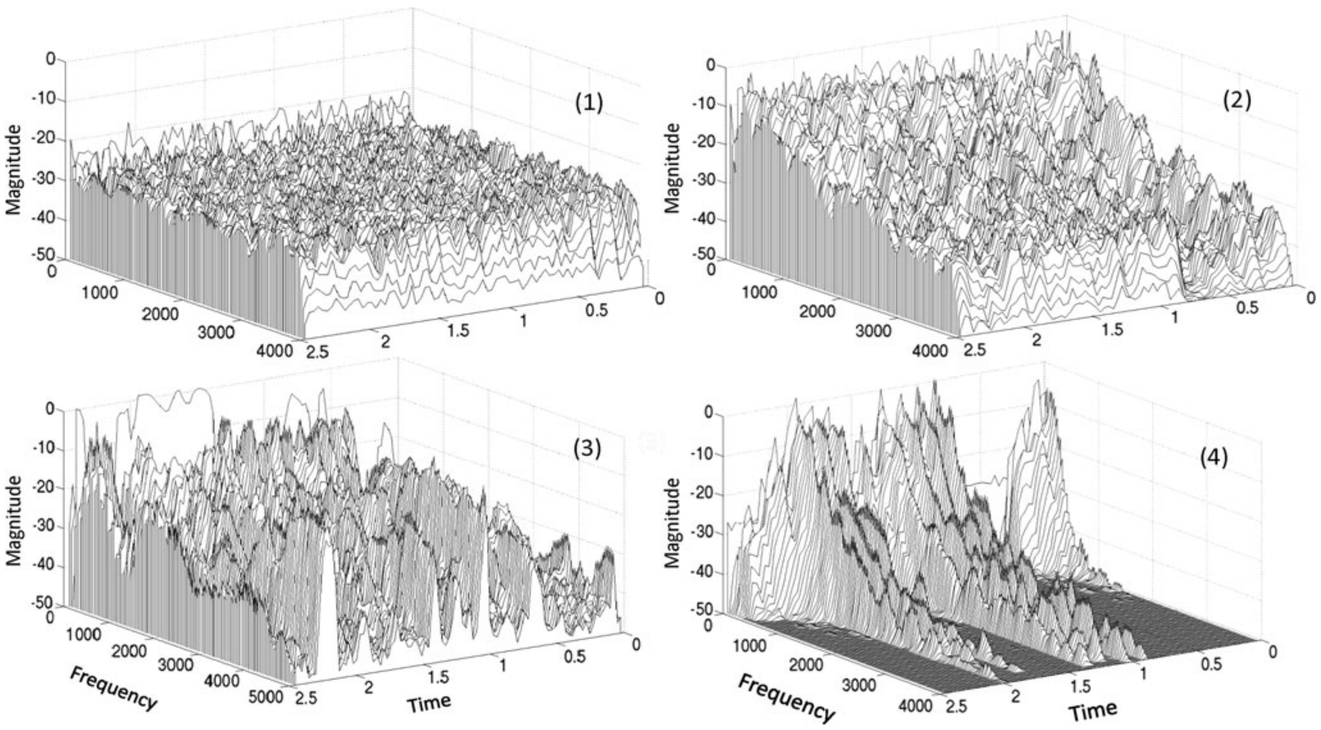


Fig. 6 Waterfall plots of 4 noise types used for evaluations in decreasing order of stationarity (**1**) white (**2**) large crowd (**3**) babble, and (**4**) machine gun noise

Table 1 Comparison of enhancement performance in different environments (BAB—babble, MGN—machine gun, LCR—large crowd). (a) Original degraded quality and quality of enhanced speech using (b) Martin's, (c) Cohen's and, (d) proposed new noise tracking schemes

	(a)	(b)	(c)	(d)
BAB				
−5 dB	4.13	3.94	3.89	**3.44**
0 dB	3.46	3.27	3.28	**2.87**
5 dB	2.71	2.55	2.70	**2.15**
MGN				
−5 dB	3.71	4.13	4.96	**3.11**
0 dB	3.22	3.60	4.45	**2.55**
5 dB	2.08	6.84	4.66	**2.33**
LCR				
−5 dB	4.69	4.55	3.99	**3.46**
0 dB	4.01	4.87	3.97	**3.06**
5 dB	2.83	2.87	3.60	**2.50**
avg rel improv		−16.9 %	−14.06 %	13.71 %

forms other schemes (7 of 9 conditions), or produces comparable enhancement output for most cases. The relative improvements are calculated using,

$$R_i = \frac{IS_{degraded} - IS_{enhanced}}{IS_{degraded}} \times 100. \tag{10}$$

Table 1 shows the IS values of the degraded speech enhanced using noise estimated with previously established schemes and speech enhanced using the proposed noise estimation scheme. As seen from these evaluations, the quality of enhancement depends heavily on the stationarity of the noise. The proposed new noise environment tracking framework is seen to outperform both the existing methods for all noise types and levels. An average 13.71 % improvement in IS measure is obtained using the new tracking algorithm. The stationarity of the noise signal can also be used to decide how many noise updates to use per frame when noise-only frames are available. The noise estimates can be further improved by incorporating information about the current noise state or adapting the Train model set to the existing noisy speech file. Another advantage of the current proposed method is the consistency in performance as noise type/level charge, whereas other methods do not provide such consistent performance under changing conditions.

6 Conclusion

In this study, environment aware speech processing solutions were proposed with specific emphasis on noise tracking and noise update rate estimation. It was shown that by utilizing a framework where the noise properties are extracted and used for noise tracking, superior tracking performances can be obtained. The environmental properties were also used for determining the update rate of noise required for a given level of enhancement quality. The proposed framework explicitly modeled the pre-observed environmental noise and its impact on speech system performance. This framework was employed for developing a novel noise tracking algorithm to achieve better speech enhancement under highly evolving noise types. The enhancement was performed used the Log-MMSE algorithm. The new *Environmentally Aware Noise Tracking* (EA-NT) method was shown to have superior performance compared to the traditional noise tracking algorithms. Evaluations were performed for speech degraded using a corpus of four noise types consisting of: Babble (BAB), Machine Gun (MGN), Large Crowd (LCR), and White Gaussian (WGN). A test set of 200 speech utterances from the TIMIT corpus were used for evaluations and an average enhancement improvement of 13 % was obtained as opposed to other schemes that degrade the speech in similar environments. The second part of this study proposed an algorithm to predict the output quality of the enhanced speech for a given enhancement scheme by focusing on analysis of the noise environment. This framework was evaluated using the Log-MMSE enhancement scheme for a corpus of four noise types consisting of Babble (BAB), White Gaussian (WGN), Aircraft Cockpit (ACN), and Highway Car (CAR) using the Itakura-Saito (IS) quality measure. An average performance mismatch of 0.13 IS was obtained using the proposed algorithm to estimate the quality of the enhanced speech. The mismatch between the predicted and observed quality is on the order of slight coding distortions for the noise types considered. These advancements provide an effective foundation for addressing noise in speech by placing emphasis on noise modeling, so that available resources can be used more efficiently to achieve superior overall performance in speech systems.

References

Akbacak, M., & Hansen, J. H. L. (2007). Environmental sniffing: noise knowledge estimation for robust speech systems. *IEEE Trans on Audio, Speech and Language Processing*, *15*(2), 465–477.

Cappe, O. (1994). Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. *IEEE Transactions on Speech and Audio Processing*, *2*(2), 345–349.

Chatlan, N., & Soraghan, J. J. (2009). Emd-based noise estimation and tracking (enet) with application to speech enhancement. In *EUSIPCO*.

Cohen, I. (2003). Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. *IEEE Transactions on Speech and Audio Processing*, *11*, 466–475.

El-Maleh, K., Samouelian, A., & Kabal, P. (1999). Frame level noise classification in mobile environments. In *ICASSP-99*, Phoenix, USA (pp. 237–240).

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum meansquare logspectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*, 433–445.

Fukane, A. R., & Sahare, S. L. (2011). Noise estimation algorithms for speech enhancement in highly non-stationary environments. *International Journal of Computer Science Issues*, *8*(2), 39.

Gray, R., Buzo, A., Gray, A., & Matsuyama, Y. (1980). Distortion measures for speech processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*, 367–376.

Hendriks, R. C., Jensen, J., & Heusdens, R. (2008). Noise tracking using dft domain subspace decompositions. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*(3), 541–553.

Kates, J. M. (1995). Classification of background noises for hearing aid applications. *The Journal of the Acoustical Society of America*, *97*, 461–470.

Krishnamurthy, N., & Hansen, J. (2006). Noise update modeling for speech enhancement: when do we do enough? In *Interspeech*, Pittsburgh, PA.

Ma, L., Smith, D., & Milner, B. (2003). Environmental noise classification for context-aware applications. In *Lecture notes in computer science*. *Database and expert systems applications* (pp. 360–370).

Martin, R. (1994). Spectral subtraction based on minimum statistics. In *Proceedings European signal processing conf* (pp. 1182–1185).

Rabiner, L., & Schafer, R. (1978). *Digital processing of speech signals*. Englewood Cliffs: Prentice-Hall.

Rangachari, S., & Loizou, P. C. (2006). A noise-estimation algorithm for highly non-stationary environments. *Speech Communication*, *8*, 220–231.

Reynolds, D., & Rose, R. (1995). Robust text independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing*, *3*, 72–83.

Xu, H., Dalsgaard, P., Tan, Z., & Lindberg, B. (2006). Robust speech recognition from noise-type based feature compensation and model interpolation in a multiple model framework. In *ICASSP-06*, Toulouse, France (Vol. 1, pp. 1141–1144).

Xu, H., Dalsgaard, P., Tan, Z., & Lindberg, B. (2007). Noise condition-dependent training based on noise classification and SNR estimation. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(8), 2431–2443.