INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Noise Aware and Combined Noise Models for Speech Denoising in Unknown Noise Conditions

Pavlos Papadopoulos[1], *Colin Vaz*[1], *Shrikanth Narayanan*[1]

[1]University of Southern California
ppapadop@usc.edu, cvaz@usc.edu, shri@sipi.usc.edu

Abstract

Traditional denoising schemes require prior knowledge or statistics of the noise corrupting the signal, or estimate the noise from noise-only portions of the signal, which requires knowledge of speech boundaries. Extending denoising methods to perform well in unknown noise conditions can facilitate processing of data captured in different real life environments, and relax rigid data acquisition protocols. In this paper we propose two methods for denoising speech signals in unknown noise conditions. The first method has two stages. In the first stage we use Long Term Signal Variability features to decide which noise model to use from a pool of available models. Once we determine the noise type, we use Nonnegative Matrix Factorization with a dictionary trained on that noise to denoise the signal. In the second method, we create a combined noise dictionary from different types of noise, and use that dictionary in the denoising phase. Both of our systems improve signal quality, as measured by PESQ scores, for all the noise types we tested, and for different Signal to Noise Ratio levels.

Index Terms: Nonnegative Matrix Factorization, Speech Signal Processing, Denoising, Long-Term Signal Variability

1. Introduction and Relation To Prior Work

Real life speech processing can be challenged by different environment noise and channel conditions, degrading the performance of speech applications. In the last few years data availability, and the need of speech applications operating under a variety of noise conditions, has renewed the efforts on more sophisticated denoising schemes. For example, subspace methods with time and spectral constraints ([1], [2]) as well as Nonnegative Matrix Factorization methods ([3], [4]) are not restricted to specific noise types (e.g. stationary or quasi-stationary). However, all these methods require either prior information about the noise conditions that corrupt the speech signal or a robust estimate of the noise. This type of knowledge cannot always be obtained, especially if the data are collected from various sources and under varying noise conditions.

The motivation behind this work is to design a system that would be able to operate under unknown noise conditions at different Signal to Noise Ratio (SNR) levels without requiring prior information about the noise. To that end, we developed two methods. The first method has two stages. In the first stage, we choose an appropriate pre-trained noise model to "match" the noise of the corrupted signal. The work in [5] addresses the problem of noise classification in speech signals using Bark Scale features, and we followed a similar approach in [6]. This is a classic pattern classification task, where the tested signal is corrupted by one of the noises that the system was trained on. The problem with this approach arises when the signal is corrupted by a type of noise that the system was not trained on. The sensitivity of Bark Scale features and MFCCs to noise results in poor generalization properties when such systems encounter unknown types of noise. To overcome these issues we use a method that is based on Long-Term Signal Variability (LTSV), introduced in [7]. Once we compute LTSV features on the test signal we construct a histogram of its values and find the Kullback–Leibler (KL) distance of the signal LTSV histogram to other LTSV histrograms in our training set. In the second stage of our system, we employ Nonnegative Matrix Factorization (NMF) to denoise the signal, using the noise model we selected in the first phase. NMF has been succesfully used in speech denoising ([3], [4]), for a variety of noise types.

In the second method, we do not supply a chosen noise model to NMF. Instead we create a "combined" noise dictionary from a variety of noises, which is used by NMF for denoising. The intuition behind this approach is that if this combined dictionary contains noise types which are similar to the noise that corrupts the signal, then the atoms of the dictionary will adequately model the noise that corrupts the test signal.

The rest of the paper is organized as follows. In Section 2 we present the system based on model selection. In Section 3 we describe the system based on the combined noise dictionary. In Section 4, we present and discuss our results. Finally, in Section 5, we draw our conclusions and provide future research directions for a speech denoising framework in unknown noise conditions.

2. Method Using Model Selection

In the first stage of our system, we choose an appropriate noise model to "match" the test signal. To achieve this goal, we first compute the LTSV values $L(m)$ for every frame m. LTSV is computed using the last R frames (analysis window) of the observed signal x with respect to the current frame of interest:

$$\begin{aligned} L(m) &\triangleq \frac{1}{K}\sum_{k=1}^{K}\left(\xi_k(m) - \overline{\xi(m)}\right)^2 \\ \overline{\xi(m)} &= \frac{1}{K}\sum_{k=1}^{K}\xi_k(m) \\ \xi_k(m) &\triangleq -\sum_{n=m-R+1}^{m}\frac{S(n,\omega_k)}{\sum_{l=m-R+1}^{m}S(l,\omega_k)} \\ &\quad \times \log\left(\frac{S(n,\omega_k)}{\sum_{l=m-R+1}^{m}S(l,\omega_k)}\right) \end{aligned} \tag{1}$$

where $S(n,\omega_k)$ is the short time spectrum computed for the n^{th} frame over $k = 1,\ldots,K$ frequencies. In our experiments the

http://dx.doi.org/10.21437/Interspeech.2016-501

length of the LTSV analysis window R is 30 frames, while for the short time spectrum we used a window of 25 ms with a shift of 10 ms. After calculating the LTSV values for each frame, we perform moving average smoothing (with a window span of 20 frames) to eliminate abrupt transitions of LTSV values. Then, we construct a normalized histogram P of the smoothed LTSV with 151 bins (the number of bins was chosen heuristically as the result of experimental procedure).

To build our training histogram set we used 300 utterances from the TIMIT database (including both male and female speakers). To each utterance we added fifteen different types of noise at five different SNR levels, from -5 dB to 15 dB with a step of 5 dB. Thus, we have a total of 1500 histograms for each of the fifteen different types of noise in the NOISEX-92 database [8], presented in Table 1.

NOISE TYPES	White
	Pink
	Speech Babble
	Tank
	Military Vehicle
	Car Interior
	Destroyer Engine Room
	Destroyer Operations Room
	F16 Cockpit
	Factory Floor 1
	Factory Floor 2
	High Frequency
	Machine Gun
	Jet Cockpit 1
	Jet Cockpit 2

Table 1: NOISEX-92 noise types used in our experiments

Given a test signal corrupted by noise, we determine which noise model matches it by calculating the KL distances between the test signal histogram P and those in our training histogram dataset Q:

$$D_{KL}(P||Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{2}$$

Then we do a majority voting on the 30 shortest distances. The result of the majority voting determines the appropriate noise model. In practice, this is a K-nearest neighbour algorithm where the data points are histograms, and the distance is the KL divergence.

We assume that the noise in the test signal did not appear in our training set. Thus, we simulate unknown noise conditions by excluding the noise that is in the test signal from the model selection process (e.g. if the signal is corrupted by pink noise we exclude it from the selection process). Notice that this model selection approach is different than a traditional pattern classification task, since the test signal is corrupted by a noise that does not exist in our noise dataset.

Once the model is selected we employ NMF to denoise the signal. In [9], different NMF objective functions are explored, which lead to different NMF variants. Generalized KL divergence (GKL), denoted as $D_{GKL}(\cdot||\cdot)$, has been successfully used for audio source separation in [10]. Hence in our experiments we will use the GKL variant of NMF, which leads to the minimization of the objective function:

$$D_{GKL}(V||WH) = \sum_{lr} \Big(V_{lr} \log \frac{V_{lr}}{(WH)_{lr}} - V_{lr} + (WH)_{lr} \Big) \tag{3}$$

In the training phase we used NMF on clean speech spectrograms[1] V_s to find a speech dictionary W_s and the time activations H_s. Similarly, we calculated spectrograms V_{n_i} with $i = 1 \ldots 15$ for the fifteen types of noise of Table 1, to find noise dictionaries W_{n_i}. Matrices W_s, W_{n_i} have dimensionality $n_f \times n_b$. In our experiments we used $n_f = 513$ and $n_b = 80$, which are values commonly used in literature (e.g. [4], [3]).

Each column of the dictionaries W_s, $\{W_{n_i}\}_{i=1}^{15}$ is a basis vector and represents a specific spectral "building block" of their respective signals. On the other hand, the matrices H_s, $\{H_{n_i}\}_{i=1}^{15}$ represent the time-varying activation levels of the basis vectors. To estimate W_s and H_s we used 1120 utterances of male speakers and 560 utterances of female speakers from the TIMIT database, while for W_{n_i} and H_{n_i} we used noise signals whose durations are approximately 3 minutes.

In the testing phase, we fix W_s assuming that its basis vectors accurately describe speech. However, since we do not have prior knowledge about the type of noise, n', that corrupts the signal, we cannot use $W_{n'}$ to describe its characteristics. Instead, we choose a noise n'' using the LTSV model selection process, and then use the corresponding noise dictionary $W_{n''}$. We expect that $W_{n''}$ will provide a good representation of the noise that corrupts the test signal. Once $W_{n''}$ is chosen we form a "complete" dictionary $W_c = [W_s \;\; W_{n''}]$ whose basis functions will be used to represent the test speech signal, which is corrupted by an unknown type of noise.

Afterwards, we compute the spectrogram V_t of the test signal. Having at our disposal W_c and V_t the goal is to find H_t by minimizing the objective function $D_{GKL}(V_t||W_c H_t)$ given by equation (3). The multiplicative update rule for H_t is given by:

$$H_{t_{ij}} \longleftarrow H_{t_{ij}} \frac{\sum_k W_{c_{ki}} V_{t_{kj}} / (W_c H_t)_{ij}}{\left[\sum_r W_{c_{ri}}\right]_\epsilon} \tag{4}$$

where $[\,]_\epsilon$ indicates that if the quantity within the brackets is less than $\epsilon > 0$ then it will be replaced with ϵ to prevent violations of the nonnegativity constraint and avoid divisions by zero.

Finally, we reconstruct the denoised spectrogram as $\hat{V}_s = W_s H_{t_{1:nb}}$, using the speech basis functions, along with the first n_b rows of H_t to approximate the target speech.

3. Method Using Combined Noise Dictionary

The second system we implemented uses a combined noise dictionary. In the training phase, we perform NMF on the speech and noise types separately. The goal is to minimize both $D_{GKL}(V_s||W_s H_s)$ and $D_{GKL}(V_{n_i}||W_{n_i} H_{n_i})$, where i is the noise type index.

[1] Speech dictionaries created through NMF are usually gender dependent ([3]) or even speaker dependent ([11]). We followed a gender dependent approach. We keep the notation V_s, W_s, H_s for notation convenience, but the reader should keep in mind that they are gender dependent.

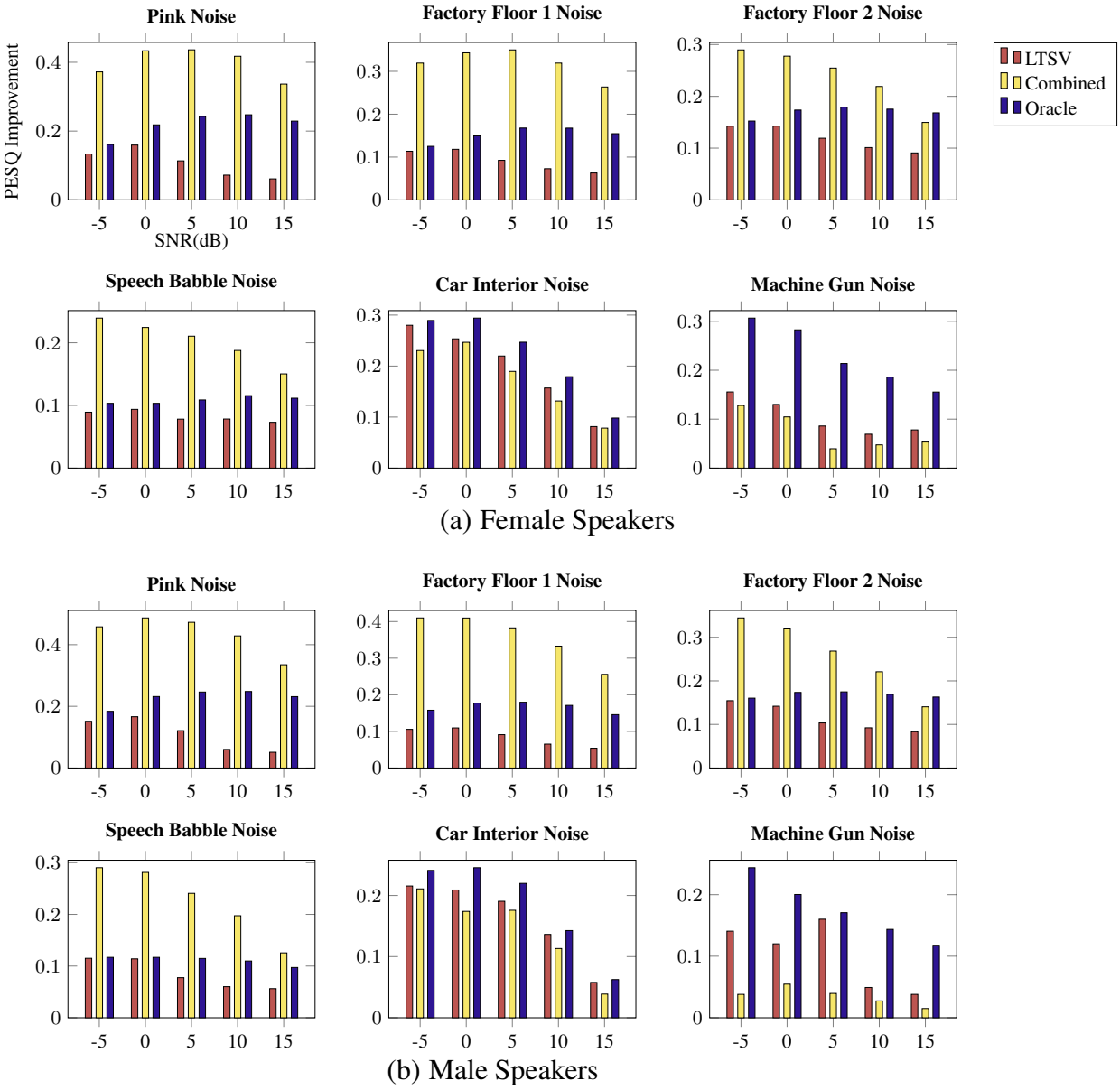

Figure 1: PESQ score improvements of the system for a variety of noise types under different SNR levels. Red depicts the performance of the LSTV system, yellow the performance of the combined dictionary system, while blue stands for the performance when we use the "true" noise model.

In the testing phase, W_s is again fixed. However, in this system we do not choose an noise dictionary. Instead we form a noise dictionary $W_a = [W_1 \;\; W_2 \;\; \ldots \;\; W_k]$ based on all the k available noise types. To simulate unknown noise conditions W_a does not contain the dictionary corresponding to the noise type corrupting the signal. For example, if it is pink noise that corrupts the signal, then W_a will not contain W_{pink}.

The intuition behind this approach is that the atoms of W_a will compensate for the missing atoms corresponding to the noise type that actually corrupts the signal. Of course for this approach to work, W_a should be created in a way that will contain similar types of noises to the one that corrupts the signal. Hence, when creating W_a one must include a variety of noise types so that W_a represents a wide range of noises.

To denoise the signal we follow the approach described in the previous section, we compute the spectrogram V_t of the test signal, we form W_c as $W_c = [W_s \;\; W_a]$, and we minimize the the objective function $D_{GKL}(V_t||W_c H_t)$ given by (3), in order to find to find H_t. Finally, we reconstruct the denoised spectrogram $\hat{V}_s = W_s H_{t_{1:nb}}$ to approximate the target speech.

4. Results and Discussion

To test the performance of our system we used 50 utterances of male speakers and 50 utterance of female speakers from the TIMIT database. Based on those utterances we created our test dataset by introducing different types of noise at five SNR levels, from -5 dB to 15 dB with a step of 5 dB. As a result, we have

250 male and 250 female noisy files for every type of noise (50 per SNR level).

We followed a leave-one-out cross-validation approach to simulate unknown noise conditions. For example, if the test signal was corrupted by white noise then we remove the LTSV histograms corresponding to white noise from the first system, while in the second system the combined noise dictionary, W_a, does not contain atoms from white noise.

To quantify our results we used the ITU Perceptual Evaluation of Speech Quality (PESQ) [12], a metric designed to match mean opinion scores of audio perceptual quality. Higher PESQ scores indicate better signal quality, and PESQ increments of the order of 0.5 offer noticeable improvements in terms of speech intelligibility [3].

In Fig. 1, we present average PESQ improvements for six different types of noise. We notice that for all the noise types and across all SNR levels, both our systems improve the signal quality. We observe that in many cases the LTSV system gives comparable results with the "oracle" system that uses the true noise type, especially in low SNR levels. However, in the case of Machine Gun noise the LTSV system falls behind compared to the oracle model because the noise pool does not contain similar noise types. Limited availability of diverse noise types can be a severe drawback to our system, since it relies on a large pool of noise models to be able to select one that closely resembles the noise that corrupts the signal. On the other hand, the system that uses a combined noise dictionary is able to outperform the oracle system in most cases. This reinforces our assumption that the atoms of the combined dictionary compensate for the missing atoms of the noise that corrupts the signal. This is clearly demonstrated in the factory floor cases, as well as in babble speech case. However, this system fails when the combined dictionary does not contain similar noise types with those corrupting the signal, as is showcased in the Machine Gun case, in which case the LTSV system performs slightly better. We observe a similar pattern in Car Interior noise, with the difference that the systems do not fail since there are similar noise types in the noise pool. This phenomenon warrants further investigation for a system that would use a method to guide the creation of the combined dictionary.

5. Conclusions and Future Work

In this paper we presented two systems that perform speech denoising under unknown noise conditions. In contrast to other state-of-the-art denoising algorithms, our systems does not require prior estimates or knowledge about the noise conditions like those in [1], [2]. Moreover, we do not make assumptions about the type of noise that corrupts the signal and generalizes to unknown types of noise.

We tested both the systems for various noise types with different statistical properties across different SNR levels, and demonstrated that it improves signal quality (as measured by PESQ scores) consistently. The performance of the system warrants further theoretical and empirical investigation that can help formalize the design.

To improve upon this work, we need to add additional features that will be able to capture noise characteristics, add an SNR estimation step to improve model selection at different SNR levels, and investigate other schemes that will match the test signal with the appropriate model (e.g. cross-entropy techniques, Deep Neural Networks, etc). At the denoising stage, we can employ more sophisticated NMF variations, incorporate temporal dynamics, add wavelet packet analysis [13], and incorporate additional knowledge, such as gender-specific models. Furthermore, we should investigate which atoms in the combined dictionary boost the performance, and combine our first system with the second to select those atoms. Finally, we should investigate how the noise model selection process can provide information to other denoising algorithms to operate in unknown noise conditions.

6. References

[1] Y. Hu and P. C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 4, pp. 334–341, 2003.

[2] S. Kamath and P. Loizou, "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002.

[3] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 4029–4032.

[4] K. W. Wilson, B. Raj, and P. Smaragdis, "Regularized nonnegative matrix factorization with temporal dependencies for speech denoising," in *INTERSPEECH , Conference of the International Speech Communication Association, Brisbane, Australia, September 22-26, 2008*, 2008, pp. 411–414.

[5] C. Eamdeelerd and K. Songwatana, "Audio noise classification using bark scale features and k-nn technique," in *International Symposium on Communications and Information Technologies*, 2008, pp. 131–134.

[6] P. Papadopoulos, A. Tsiartas, J. Gibson, and S. Narayanan, "A Supervised Signal-to-Noise Ratio Estimation of Speech Signals," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2014, pp. 8237–8241.

[7] P. Ghosh, A. Tsiartas, and S. Narayanan, "Robust Voice Activity Detection Using Long-Term Signal Variability," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 3, pp. 600–613, 2011.

[8] A. Varga and H. J. M. Steeneken, "Assessment for Automatic Speech Recognition II: NOISEX-92: A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, July 1993.

[9] A. Cichocki, R. Zdunek, and S.-I. Amari, "New algorithms for non-negative matrix factorization in applications to blind source separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings.*, 2006.

[10] P. Smaragdis, "From learning music to learning to separate," Forum Acusticum, 2005.

[11] F. Weninger, J. Feliu, and B. Schuller, "Supervised and semi-supervised suppression of background music in monaural speech recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[12] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings.*, 2001.

[13] C. Vaz, V. Ramanarayanan, and S. Narayanan, "A two-step technique for mri audio enhancement using dictionary learning and wavelet packet analysis." in *INTERSPEECH*, 2013, pp. 1312–1315.