A Hybrid Approach for Speech Enhancement Using MoG Model and Neural Network Phoneme Classifier

Shlomo E. Chazan, Jacob Goldberger, and Sharon Gannot, *Senior Member, IEEE*

***Abstract*—In this paper, we present a single-microphone speech enhancement algorithm. A hybrid approach is proposed merging the generative mixture of Gaussians (MoG) model and the discriminative deep neural network (DNN). The proposed algorithm is executed in two phases, the training phase, which does not recur, and the test phase. First, the noise-free speech log-power spectral density is modeled as an MoG, representing the phoneme-based diversity in the speech signal. A DNN is then trained with phoneme labeled database of clean speech signals for phoneme classification with mel-frequency cepstral coefficients as the input features. In the test phase, a noisy utterance of an untrained speech is processed. Given the phoneme classification results of the noisy speech utterance, a speech presence probability (SPP) is obtained using both the generative and discriminative models. SPP-controlled attenuation is then applied to the noisy speech while simultaneously, the noise estimate is updated. The discriminative DNN maintains the continuity of the speech and the generative phoneme-based MoG preserves the speech spectral structure. Extensive experimental study using real speech and noise signals is provided. We also compare the proposed algorithm with alternative speech enhancement algorithms. We show that we obtain a significant improvement over previous methods in terms of speech quality measures. Finally, we analyze the contribution of all components of the proposed algorithm indicating their combined importance.**

***Index Terms*—MixMax model, neural-network, phoneme classification, speech enhancement.**

I. Introduction

ENHANCING noisy speech received by a single microphone is a widely-explored problem. A plethora of approaches can be found in the literature [1]. Although many current devices are equipped with multiple microphones, there are still many applications for which only a single microphone is available.

The celebrated short-time spectral amplitude estimator (STSA) and log-spectral amplitude estimator (LSAE) [2], [3] are widely-used model-based algorithms. The optimally modified log spectral amplitude (OMLSA) estimator and, in particular, the improved minima controlled recursive averaging (IMCRA) noise estimator are specifically tailored to nonstationary noise environments [4], [5]. However, fast changes in noise statistics may cause a severe performance degradation, often manifested as *musical noise* artifacts at the output of the enhancement algorithm.

Recently, DNN techniques gained a lot of popularity due to theoretical and algorithmic progress, and the availability of more data and more processing power. Unlike past learning algorithms for DNN, it is now possible to infer the parameters of the DNN with many layers, and hence the name *deep learning*. Deep learning methods were mainly applied to speech recognition and, more recently, to speech enhancement as well. DNN and a deep auto-encoder (DAE) were used as a nonlinear filters in [6] and [7], respectively. The networks are trained on stereo (noisy and clean) audio features, to infer the complex mapping from noisy to clean speech. An experimental study testing this approach is given in [8]. The DNN significantly reduces the noise level. However, the enhanced signals suffer from noticeable *speech distortion*.

Other methods attempt to train a DNN to find a mask, which classifies the time-frequency bins into speech/noise classes. Given the binary mask, the noisy bins are suppressed. In [9] for instance, a support vector machine (SVM) is used to estimate the ideal binary mask (IBM) for speech separation from non-speech background interference. A DNN is trained to find the input features for the SVM. A simpler approach is to train the DNN itself to find the IBM. Different targets for the DNN are presented in [10]. The IBM has shown advantageous in terms of intelligibility [11]. Yet, the binary mask is known to introduce artifacts such as musical noise. Although, for intelligibility tasks, this might not be problematic, for speech enhancement tasks the application of IBM may not sufficient. To circumvent this phenomenon, in [12] the DNN is trained to find the ideal ratio mask (IRM), which is a soft mask. A comparison between the IBM and the IRM is presented in [13], demonstrating that latter outperforms the former in terms of speech quality. Although not assuming any specific model for the enhanced signals, these method do not generalize well to noise types that are beyond the noisy training set, resulting in poor enhancement in an untrained noise environment. To circumvent this problem, in [14] the DNN was trained with more than 100 different types of noise. Nevertheless, in real-life where the number of noise types is unlimited, this approach may not be satisfactory.

Training-based algorithms, such as mixmax (MM) [15], were also developed. These algorithms are carried out in two phases, the training phase and the test phase. In the training phase the parameters of the model are found, usually with an unsupervised machine learning algorithms, such as the expectation-maximization (EM) algorithm in [15]. In the test phase, the enhancement is carried out using the learned model parameters. One weakness of the algorithm is that the speech

Manuscript received May 10, 2016; revised August 14, 2016 and October 7, 2016; accepted October 7, 2016. Date of publication October 18, 2016; date of current version November 4, 2016. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rongshan Yu.

The authors are with the Faculty of Engineering, Bar-Ilan University, Ramat-Gan 5290002, Israel (e-mail: Shlomi.Chazan@biu.ac.il; Jacob.Goldberger@biu.ac.il; Sharon.Gannot@biu.ac.il).

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

Digital Object Identifier 10.1109/TASLP.2016.2618007

parameters are found in an unsupervised manner that ignores the phoneme-based structure of speech. Another drawback of the MM algorithm is that the noise parameters are estimated once at the beginning of the utterance and then are kept fixed during the entire utterance. This enhancement approach is not always sufficient for real-life noises.

In this study, we propose a *hybrid* approach, which integrates two distinctive paradigm for speech enhancement, namely the generative model-based approach (using MoG model) and the discriminative DNN approach. As in [15], we use a two phase algorithm. In the training phase, the *clean* speech is modeled by a phoneme-based MoG that is built using phoneme-labeled database. A DNN is then trained to classify clean[1] time-frame features to one of the phonemes in the phoneme-based MoG. Once the training phase is over, it does not recur. In the test phase a noisy utterance is processed. The DNN estimates the phonemes, and use them to calculate an SPP utilizing the generative model. The SPP controls the amount of attenuation of the noisy time-frequency bin (the lower is the SPP the higher is the attenuation). Simultaneously, when the SPP is low, the noise estimate is updated. The continuity of the speech is maintained by feeding the DNN with context frames on top of the current frame. In addition, the DNN supports the calculation of the SPP. Furthermore, the phoneme-based MoG and the soft SPP preserve the spectral structure of the speech, thus alleviating the *musical noise* phenomenon.

The contribution of this paper is four-fold. First, we substitute the original generic MoG speech modelling in [15] with the more appropriate MoG modelling, inferred from its phoneme structure. Second, we derive an accurate bin-wise SPP detector. The proposed SPP detector is a combination of the phoneme-based generative MoG speech model and a discriminative DNN, utilized as a powerful phoneme classifier. Third, based on the proposed SPP detector, a fast adaption mechanism for the noise statistics is derived. Finally, a simplified reconstruction procedure is proposed, based on spectral attenuation, with the noise attenuation level controlled by the SPP, thus alleviating the annoying *musical noise* artifacts. To summarize, our proposed *hybrid* approach, harnesses the power of the modern DNN classifier to improve the well-established model-based speech enhancement method.

The rest of the paper is organized as follows. In Section II, a generative model is presented. Section III presents the proposed enhancement algorithm and describes its implementation in details. A comprehensive experimental results using speech databases in various noise types are presented in Section IV. In Section V the building blocks of the algorithm are analyzed. Finally, some conclusions are drawn and the paper is summarized in Section VI.

II. A Generative Noisy Speech Model

In this section, a generative model of the noisy speech signal is presented. We follow the model proposed by Nádas *et al.* [16] that was utilized in [15].

[1]The DNN is trained on clean signals in order to remain general and not to bias the network towards certain noise types.

A. Maximization Approximation

Let $x(t)$ and $y(t)\quad 0<t<T$ denote the speech and noise signals, respectively. The observed noisy signal $z(t)$ is given by

$$z(t)=x(t)+y(t). \tag{1}$$

Applying the short-time Fourier transform (STFT) (with frame-length set to L samples and overlap between successive frames set to $3L/4$ samples) to $z(t)$ yields $Z(n,k)$ with n the frame index and $k=0,1,\ldots,L-1$ the frequency index. The frame index n is henceforth omitted for brevity, whenever applicable.

Let $\mathbf{Z}$ denote the $L/2+1$ dimensional log-spectral vector, defined by

$$Z_k=\log|Z(k)|,\quad k=0,1,\ldots,L/2.$$

Note that the other frequency bins can be obtained by the symmetry of the discrete Fourier transform (DFT). Similarly, we define $\mathbf{X}$ and $\mathbf{Y}$ to be the log-spectral vectors of the speech and noise signals, respectively.

It is assumed that the noise is statistically independent of the speech signal. Furthermore, it is assumed that both the speech and noise are zero-mean stochastic processes. Due to these assumptions the following approximation can be justified:

$$|Z(k)|^2\approx|X(k))|^2+|Y(k)|^2$$

hence

$$2Z_k\approx\log(e^{2X_k}+e^{2Y_k}).$$

Following Nádas *et al.* [16], the noisy log-spectral can be further approximated:

$$\mathbf{Z}\approx\max(\mathbf{X},\mathbf{Y}) \tag{2}$$

where the maximization is component-wise over the elements of $\mathbf{X}$ and $\mathbf{Y}$. This approximation was found useful for speech recognition [16], speech enhancement [15], [17] and speech separation tasks [18], [19]. In a speech enhancement task, only the noisy signal $\mathbf{Z}$ is observed, and the aim is to estimate the clean speech $\mathbf{X}$.

B. Clean Speech Model - Phoneme Based MoG

It is well-known that a speech utterance can be described as a time-series of phonemes, i.e. speech is uttered by pronouncing a series of phonemes [20]. In our approach, we give this observation a probabilistic description, namely the log-spectral vector of the clean speech signal, $\mathbf{X}$, is modeled by a MoG distribution, where each mixture component is associated with a specific phoneme. Training a MoG for each phoneme is a common practice in speech recognition. Here our goal is not to extract the exact transcription but to improve the speech quality. Hence, to make our enhancement algorithm robust to many variants of noise types and signal to noise ratio (SNR) levels, a parsimonious speech model that provides a simple but yet effective parametric representation of the clean speech was built. Unlike [15], that uses unsupervised clustering of the speech frames, we use here a supervised clustering, explicitly utilizing the labels of the phonemes of the training speech signals. Based

on the MoG model, the probability density function $f_{\mathbf{X}}(\mathbf{x})$ of the clean speech $\mathbf{X}$, can be written as

$$f_{\mathbf{X}}(\mathbf{x}) = \sum_{i=1}^{m} c_i f_i(\mathbf{x}) = \sum_{i=1}^{m} c_i \prod_k f_{i,k}(x_k) \tag{3}$$

where m is the number of mixture components and

$$f_{i,k}(x_k) = \frac{1}{\sqrt{2\pi}\sigma_{i,k}} \exp\left\{-\frac{(x_k - \mu_{i,k})^2}{2\sigma_{i,k}^2}\right\}. \tag{4}$$

Let I be the phoneme indicator random variable (r.v.) associated with the MoG r.v. $\mathbf{X}$, i.e. $p(I = i) = c_i$. The term $f_i(\mathbf{x})$ is the Gaussian probability density function (p.d.f.) of $\mathbf{X}$ given that $I = i$. The scalar c_i is the probability of the i-th mixture and $\mu_{i,k}$ and $\sigma_{i,k}$ are the mean and the standard deviation of the k-th entry of the i-th mixture Gaussian, respectively.

Any residual correlation between the frequency bins is neglected. The log-spectral vector, which provides a full frequency-based description of the speech signal, is used here as the feature-set for the MoG model since it provides full description of the frequency content required for reconstruction. Using a MoG with diagonal covariance matrices is only a simplified modeling of a clean speech signal but has an advantage of a robust modeling that circumvents the need for large matrices inversion.

To set the MoG parameters we used the phoneme-labeled TIMIT database [21] as described in Section III-D.

C. Noisy Speech Model

Let $\mathbf{Y}$ define the log-spectral vector of the noise signal, and let $g_{\mathbf{Y}}(\mathbf{y})$ denote the p.d.f. of $\mathbf{Y}$. As with the log-spectral vector of the speech signal, it is assumed that the components of $\mathbf{Y}$ are statistically independent. For simplicity, $g_{\mathbf{Y}}(\mathbf{y})$ is modeled as a single Gaussian, with diagonal covariance i.e.,

$$g_{\mathbf{Y}}(\mathbf{y}) = \prod_k g_k(y_k) \tag{5}$$

where

$$g_k(y_k) = \frac{1}{\sqrt{2\pi}\sigma_{Y,k}} \exp\left\{-\frac{(y_k - \mu_{Y,k})^2}{2\sigma_{Y,k}^2}\right\}. \tag{6}$$

Initial estimation and adaptation of the noise parameters will be explained in Section III-E.

Using the maximum assumption in the log-spectral vector of the noisy speech $\mathbf{Z} = \max(\mathbf{X}, \mathbf{Y})$, as explained above, it can be verified [16] that the p.d.f. of $\mathbf{Z}$ is given by the following mixture model:

$$h(\mathbf{z}) = \sum_{i=1}^{m} c_i h_i(\mathbf{z}) = \sum_{i=1}^{m} c_i \prod_k h_{i,k}(z_k) \tag{7}$$

where

$$h_{i,k}(z_k) = f_{i,k}(z_k)G_k(z_k) + F_{i,k}(z_k)g_k(z_k) \tag{8}$$

such that $F_{i,k}(x)$ and $G_k(y)$ are the cumulative distribution functions of the Gaussian densities $f_{i,k}(x)$ and $g_k(y)$, respectively. The term $h_i(\mathbf{z})$ is the p.d.f. of $\mathbf{Z}$ given that $I = i$.

The generative modeling described above was nicknamed MM [15], [16], since it is based the modelling of the clean speech as a (Gaussian) mixture p.d.f. and the noisy speech is modeled as the maximum of the clean speech and the noise signal. Originally, the mixture components were not associated with phonemes, but rather learned in an unsupervised manner.

III. The Neural-Network MixMax Algorithm

In this section, we describe the proposed enhancement algorithm. In Section III-A we remind the minimum mean square error (MMSE) estimator based on the MM model [15], [16]. We then propose in Section III-B a new variant of the estimator that utilizes the same model but allows for better noise attenuation. In Section III-C a DNN approach is introduced as a tool for accurate phoneme classification. Issues regarding the training of the DNN are discussed in Section III-D. Finally, test-phase noise adaption is discussed in Section III-E.

A. The MMSE-Based Approach

An MMSE of the clean speech $\mathbf{X}$ from measurement $\mathbf{z}$ is obtained by the conditional expectation $\hat{\mathbf{x}} = E(\mathbf{X}|\mathbf{Z} = \mathbf{z})$. Note, that since the p.d.f. of both $\mathbf{X}$ and $\mathbf{Z}$ is non-Gaussian, this estimator is not expected to be linear. Utilizing the generative model described in the previous section we can obtain a closed-form solution for the MMSE estimator as follows.

$$\hat{\mathbf{x}} = \sum_{i=1}^{m} p(I = i|\mathbf{Z} = \mathbf{z}) E(\mathbf{X}|\mathbf{Z} = \mathbf{z}, I = i). \tag{9}$$

The posterior probability $p(I = i|\mathbf{Z} = \mathbf{z})$ can be easily obtained from (7) by applying the Bayes' rule:

$$p(I = i|\mathbf{Z} = \mathbf{z}) = \frac{c_i h_i(\mathbf{z})}{h(\mathbf{z})}. \tag{10}$$

Since the covariance matrices of both the speech and the noise models are diagonal, we can separately compute

$$\hat{\mathbf{x}}_i = E(\mathbf{X}|\mathbf{Z} = \mathbf{z}, I = i)$$

for each frequency bin. For the k-th frequency bin we obtain:

$$\begin{aligned} \hat{x}_{i,k} &= E(X_k|Z_k = z_k, I = i) \\ &= \rho_{i,k}^{\text{MM}} z_k + (1 - \rho_{i,k}^{\text{MM}}) E(X_k|X_k < z_k, I = i) \end{aligned} \tag{11}$$

such that

$$\rho_{i,k}^{\text{MM}} = p(X_k > Y_k|Z_k = z_k, I = i) = \frac{f_{i,k}(z_k)G_k(z_k)}{h_{i,k}(z_k)} \tag{12}$$

and for the second term in (11) is:

$$E(X_k|X_k < z_k, I = i) = \mu_{i,k} - \sigma_{i,k}^2 \frac{f_{i,k}(z_k)}{F_{i,k}(z_k)}. \tag{13}$$

The closed-form expression for the MMSE estimator of the clean speech $\hat{\mathbf{x}} = E(\mathbf{X}|\mathbf{Z} = \mathbf{z})$ [16] is obtained from (9), (11), (12), (13). These expressions are the core of the MM speech enhancement algorithm proposed by Burshtein and Gannot [15]. In their approach, the MoG parameters of the clean speech are inferred from a database of speech utterances utilizing the EM algorithm in an unsupervised manner.

B. Soft Mask Estimation of the Clean Speech

Assuming the maximization model in (2) is valid, $\rho_{i,k}^{\text{MM}}$ was obtained in (12). Summing over all the possible mixture components, we obtain:

$$\rho_k^{\text{MM}} = \sum_{i=1}^{m} p(I=i|\mathbf{Z}=\mathbf{z})\rho_{i,k}^{\text{MM}} = p(X_k > Y_k|\mathbf{Z}=\mathbf{z}). \quad (14)$$

The term ρ_k^{MM} can be interpreted as the probability that, given the noisy speech vector $\mathbf{z}$, the k-th frequency bin of the current log-spectral vector $\mathbf{z}$ is originated from the clean speech and not from the noise. The probability ρ_k^{MM} can thus be viewed as a training-based SPP detector, namely the probability that the designated time-frequency bin is dominated by speech. Consequently, $(1-\rho_k^{\text{MM}})$ can be interpreted as the posterior probability that the k-th bin is dominated by noise.

Using ρ_k^{MM} and some straightforward applications of the Bayes' rule, we can derive the k-th frequency bin of the MMSE estimator $\hat{\mathbf{x}} = E(\mathbf{X}|\mathbf{Z}=\mathbf{z})$ from (9), (11) :

$$\begin{aligned}\hat{x}_k &= \sum_{i=1}^{m} p(I=i|\mathbf{Z}=\mathbf{z})\hat{x}_{i,k} \\ &= \rho_k^{\text{MM}} z_k + (1-\rho_k^{\text{MM}})E(X_k|X_k < z_k, \mathbf{Z}=\mathbf{z}) \quad (15)\end{aligned}$$

Hence, given the generative model, the enhancement procedure in (16) substitutes the frequency bins identified as noise with the a priori value drawn from the MoG model given by (13). Note, that (11) differs from (16), as the former is the conditional expectation given the measurement *per class* $I=i$, while the latter is the conditional expectation given the measurement, i.e. using the entire set of classes.

The structure of a voiced speech power spectral density (PSD) consists of dominant spectral lines which recur at multiples of the fundamental frequency (known as *pitch*). The PSD of different speakers pronouncing the same phoneme share similar properties, but are never identical. Hence, the MoG parameters inferred from multiple speakers, is never individualized to the current speaker and therefore cannot represent the specific periodicity. The phoneme-based MoG parameters are only capable of preserving the general structure of an averaged phoneme. This might lead to *residual noise* even when the algorithm identifies the noise correctly.

To circumvent this problem, we propose to substitute the conditional expectation that uses the MoG parameters with a simpler estimator based on the soft spectral attenuation paradigm, namely:

$$E(X_k|X_k < z_k, \mathbf{Z}=\mathbf{z}) \quad (16)$$

is substituted by:

$$z_k - \beta \quad (17)$$

where β is a noise *attenuation* level (in the log domain). It is well-known that frequency-selective attenuation is prone to *musical noise* [22] [23]. In our proposed method, the estimator also incorporates the soft mask deduced from the SPP, thus potentially alleviating the musical noise phenomenon.

Substituting $(z_k - \beta)$ in (15) we obtain the following simplified expression for the estimated clean speech:

$$\hat{x}_k = \rho_k^{\text{MM}} \cdot z_k + (1-\rho_k^{\text{MM}}) \cdot (z_k - \beta) \quad (18)$$

or, equivalently

$$\hat{x}_k = z_k - (1-\rho_k^{\text{MM}}) \cdot \beta \quad (19)$$

which can be interpreted as an SPP-driven spectral attenuation. In speech enhancement there is always a tradeoff between speech distortion and noise suppression.

C. Neural Network for Phoneme Classification

The gist of our approach is the calculation of the SPP (14). This calculation necessitates two terms, $\rho_{i,k}^{\text{MM}}$ which is given by (12) and the posterior phoneme probability $p_i \triangleq p(I=i|\mathbf{Z}=\mathbf{z})$. Utilizing the generative model defined in Section II, p_i is obtained from (7) by applying the Bayes' rule (10). This approach exhibits some major shortcomings. Estimating the required noise statistics is a cumbersome task, especially in time-varying scenarios. Furthermore, as the calculation in (10) is carried out independently for each frame, continuous and smooth speech output cannot be guaranteed.

In our approach, (unlike [15]) we adopt a *supervised* learning approach in which each mixture component of the clean speech is associated with a specific phoneme. Hence the computation of the posterior probability of a specific mixture is merely a phoneme classification task (given the noisy speech). To implement this supervised classification task, we substitute (10) with a DNN that is known to be significantly better than MoG models for phoneme classification tasks (see e.g. [24]).

The DNN is trained on a phoneme-labeled clean speech. For each log-spectral vector, $\mathbf{z}$ (identical to $\mathbf{x}$ for the clean speech), we calculate the corresponding mel-frequency cepstral coefficients (MFCC) features (and their respective delta and delta-delta features). To preserve the continuity of the speech signal, 17 MFCC vectors are concatenated (the current vector and 8 vectors in each side) to form the feature vector, denoted $\mathbf{v}$, which is a standard feature set for phoneme classification. This feature vector is used as the input to the DNN, and the phoneme label as the corresponding target. The phoneme-classification DNN is trained on clean signals. However, as part of the speech enhancement procedure, we apply it to noisy signals. To alleviate the mismatch problem between train and test conditions, we use a standard preprocessing stage for robust phoneme classification, namely cepstral mean and variance normalization (CMVN) [25].

The SPP is calculated using (14), which requires both $\rho_{i,k}^{\text{MM}}$ and p_i. While $\rho_{i,k}^{\text{MM}}$ is calculated from the generative model using (12), we propose to replace (10) for calculating p_i by a better phoneme-classification method.

It is therefore proposed to infer the posterior phoneme probability by utilizing the *discriminative* DNN, rather than resorting to the generative MoG model:

$$p_i^{\text{NN}} = p(I=i|\mathbf{v};\text{DNN}). \quad (20)$$

Note, that the compound feature vector $\mathbf{v}$ is used instead of the original log-spectrum $\mathbf{z}$. Finally, the SPP ρ_k is obtained using (12) and (20):

$$\rho_k \triangleq \rho_k^{\text{NN-MM}} = \sum_{i=1}^{m} p_i^{\text{NN}} \rho_{i,k}^{\text{MM}}. \tag{21}$$

The proposed SPP calculation is thus based on a *hybrid* method, utilizing both the generative MoG model and a discriminative approach to infer the posterior probability. For the latter we harness the known capabilities of the DNN.

D. Training the MoG Model and the DNN Classifier

We used the phoneme-labeled clean speech TIMIT database [21] to train a DNN phoneme classifier and the MoG phoneme-based generative model. We describe next the training procedure. We used 462 speakers from the training set of the database excluding all SA sentences, since they consist of identical sentences to all speakers in the database, and hence can bias the results.

In the training phase of the phoneme-based MoG we set the number of Gaussians to $m = 39$ (see [26]), where each Gaussian corresponds to one phoneme. All frames labeled by the i-th phoneme were grouped, and for each frequency bin the mean and variance were computed using (22). First, the log-spectrum of the segments of these clean speech utterances is calculated. Since the database is labeled, each segment is associated with a phoneme i. We can then calculate the following first- and second-moment with phoneme label i:

$$\begin{aligned} \mu_{i,k} &= \frac{1}{N_i} \sum_{n=1}^{N_i} x_{i,k}(n) \\ \sigma_{i,k}^2 &= \frac{1}{N_i - 1} \sum_{n=1}^{N_i} \left(x_{i,k}(n) - \mu_{i,k}\right)^2 \end{aligned} \tag{22}$$

where $x_{i,k}(n)$ is k-th bin of the n-th log-spectra vector with phoneme label i. The term, N_i is the total number of vectors associated with phoneme labeled i.

For training the DNN as a discriminative phoneme classifier, we used the MFCC feature vectors $\mathbf{v}$ powered by the delta and delta-delta coefficients. In total, 39 coefficients per time frame were used. Eight Context frames from each side were added to the current frame as proposed in [27]. Hence, each time frame was represented by 663 MFCC features. We used a DNN with 2 hidden layers. Each hidden layer comprising 500 neurons.

The network is constructed with rectified linear unit (ReLU) as the transfer function [27], [28] for the hidden layers:

$$\begin{aligned} h_{1,i} &= \max(0, \mathbf{w}_{1,i}^{\top}\mathbf{v} + b_{1,i}), \quad i = 1, \ldots, 500 \\ h_{2,i} &= \max(0, \mathbf{w}_{2,i}^{\top}\mathbf{h}_1 + b_{2,i}), \quad i = 1, \ldots, 500 \end{aligned}$$

and a softmax output layer to obtain m probabilities associated with the various phonemes:

$$p(I = i|\mathbf{v}) = \frac{\exp(\mathbf{w}_{3,i}^{\top}\mathbf{h}_2 + b_{3,i})}{\sum_{k=1}^{m} \exp(\mathbf{w}_{3,k}^{\top}\mathbf{h}_2 + b_{3,k})}, \quad i = 1, \ldots, m$$

where $\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{b}_1, \mathbf{b}_2$ and $\mathbf{b}_3$ are the weights matrices, and the bias vectors of the hidden layers and the output layer, respectively. Given a sequence of MFCC feature vectors $\mathbf{v}_1, .., \mathbf{v}_N$, where N is the total number of vectors in the training set, with the corresponding phoneme labels, $I_1, \ldots, I_N \in \{1, \ldots, m\}$, the DNN is trained to maximize the log-likelihood function:[2]

$$L(\mathbf{W}) = \sum_{t=1}^{N} \log p(I_t|\mathbf{v}_t; \mathbf{W}). \tag{23}$$

where $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \mathbf{w}_3, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3\}$ is the parameter set of the model. We can initialize the training procedure of $\mathbf{W}$ by random weights (or use pre-training methods (see [29])) and continue by applying back-propagation algorithm in conjunction with gradient ascent procedure. Additionally, we used the dropout method [30] to circumvent over-fitting of the DNN to the training set database. To avoid mismatch between train and test conditions, each utterance was normalized prior to the adaptation of the network, such that the sample-mean and sample-variance of the utterance are zero and one, respectively [25].

To verify the accuracy of the classifier, the trained DNN was applied to a clean *test set* (24-speaker core test set drawn from TIMIT database). Finally, during the test phase of the algorithm, the DNN is applied to speech signals contaminated by additive noise. We have therefore applied the CMVN procedure, prior to the application, of the classifier to circumvent the noisy test condition [25].

E. Noise parameter initialization and adaptation

To estimate the noise parameters it is assumed that the first part of the utterance (usually 0.25 Sec) the speech is inactive and it consists of noise-only segments. These first segments can therefore be used for initializing the parameters of the noise Gaussian distribution as follows:

$$\begin{aligned} \hat{\mu}_{Y,k} &= \frac{1}{N_Y} \sum_{n=1}^{N_Y} y_k(n) \\ \hat{\sigma}_{Y,k}^2 &= \frac{1}{N_Y - 1} \sum_{n=1}^{N_Y} \left(y_k(n) - \hat{\mu}_{Y,k}\right)^2 \end{aligned} \tag{24}$$

where N_Y is the number of vectors constructed form the noise-only samples, and $\hat{\mu}_{Y,k}$ and $\hat{\sigma}_{Y,k}^2$ are the initial estimate of the noise parameters. The term $y_k(n)$ denotes the k-th bin of the n-th noise vector.

In [15], the noise parameters remain fixed for the entire utterance, rendering this estimate incapable of processing nonstationary noise scenarios. The noise p.d.f. $g(\mathbf{y})$, is a vital element in (12). Therefore, updating the parameters of the noise Gaussian is crucial for calculating accurate SPP. Following an adaptation algorithm presented in [5] (see (3) in [5]) we use the SPP to update the noise parameters. We used the following adaptation

[2]Maximizing the log-likelihood function (which is the log-probability of the correct label given the MFCC feature vector) is equivalent to minimizing the cross entropy.

Algorithm 1: Training phase.

Input :
- Clean speech log-spectral vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N$
- Corresponding MFCC vectors $\mathbf{v}_1, \ldots, \mathbf{v}_N$
- Corresponding phoneme labels $i_1, \ldots, i_N$

Output: MoG parameters and trained DNN (22)

MoG construction:
for *i = 1:39* **do**
  for *k = 0:L/2* **do**
    $\hat{\mu}_{i,k} = \frac{1}{N_i} \sum_{n=1}^{N_i} x_{i,k}(n)$
    $\hat{\sigma}^2_{i,k} = \frac{1}{N_i - 1} \sum_{n=1}^{N_i} \left(x_{i,k}(n) - \mu_{i,k}\right)^2$
  end
end
DNN training:
Train a DNN for phoneme classification using $(\mathbf{v}_1, i_1), \ldots, (\mathbf{v}_N, i_N)$ by maximizing (23)

Algorithm 2: Test phase.

Input : Log-spectral vector of the noisy speech $\mathbf{z}$ and a corresponding MFCC vector $\mathbf{v}$.
Output: Estimated log-spectral vector of the clean speech $\hat{\mathbf{x}}$.

Compute the phoneme classification probabilities using the trained DNN (20):

$$p_i^{\mathrm{NN}} = p(I = i | \mathbf{v}; \mathrm{DNN}), \quad i = 1, \ldots, m$$

for *k = 0:L/2* **do**
  Compute (12):

$$\rho_{i,k}^{\mathrm{MM}} = p(Y_k < X_k | Z_k = z_k, I = i) = \frac{f_{i,k}(z_k) G_k(z_k)}{h_{i,k}(z_k)}, \qquad i = 1, \ldots, m$$

  Compute the speech presence probability (21):

$$\rho_k = \sum_{i=1}^{m} p_i^{\mathrm{NN}} \rho_{i,k}^{\mathrm{MM}}$$

  Estimate the clean speech by adapting (19) to the DNN-based SPP:

$$\hat{x}_k = z_k - (1 - \rho_k) \cdot \beta$$

  Adapt the noise parameters $\mu_{Y,k}$ and $\sigma_{Y,k}$ (25):

$$\hat{\mu}_{Y,k} \leftarrow \rho_k \hat{\mu}_{Y,k} + (1 - \rho_k) \left(\alpha z_k + (1 - \alpha) \hat{\mu}_{Y,k}\right)$$
$$\hat{\sigma}^2_{Y,k} \leftarrow \rho_k \hat{\sigma}^2_{Y,k} + (1 - \rho_k) \left(\alpha (z_k - \hat{\mu}_{Y,k})^2 + (1 - \alpha) \hat{\sigma}^2_{Y,k}\right)$$

end

scheme for the noise model parameters:

$$\hat{\mu}_{Y,k} \leftarrow \rho_k \hat{\mu}_{Y,k} + (1 - \rho_k) \left(\alpha z_k + (1 - \alpha) \hat{\mu}_{Y,k}\right) \tag{25}$$
$$\hat{\sigma}^2_{Y,k} \leftarrow \rho_k \hat{\sigma}^2_{Y,k} + (1 - \rho_k) \left(\alpha (z_k - \hat{\mu}_{Y,k})^2 + (1 - \alpha) \hat{\sigma}^2_{Y,k}\right)$$

where $0 < \alpha < 1$ is a smoothing parameter. Using this scheme, the noise statistics can be adapted during speech utterances, utilizing the frequency bins that are dominated by noise. This scheme is particularly useful in nonstationary noise scenarios. As a consequence, the first few segments, assumed to be dominated by noise, are only used for initializing the noise statistics and their influence is fading out as more data is collected. If speech is active during the first 0.25 Sec of the utterance, we might encounter some initial speech distortion problems, since speech components might influence the estimated noise statistics. The algorithm will recover very fast from this mismatched initialization, due to the capabilities of the DNN-based SPP to accurately distinguish between speech phonemes and noise.

The train and test algorithms are summarized in Algorithms 1 and 2, respectively. In Fig. 1 we present a block diagram summarizing the entire algorithm. The blue lines represent the generative path, the light blue lines represent the discriminative path and the orange lines represent the noise adaptation path. We dub the proposed algorithm neural network mixture-maximum (NN-MM) to emphasize its hybrid nature, as a combination of the generative MM model and the phoneme-classification DNN.

IV. Experimental Study

In this section we present a comparative experimental study. We first describe the experiment setup in Section IV-A. In Section IV-B we examine the influence of the attenuation factor. Objective quality measure results are then presented in Section IV-C. Finally, the algorithm is tested with an untrained database in Section IV-D.

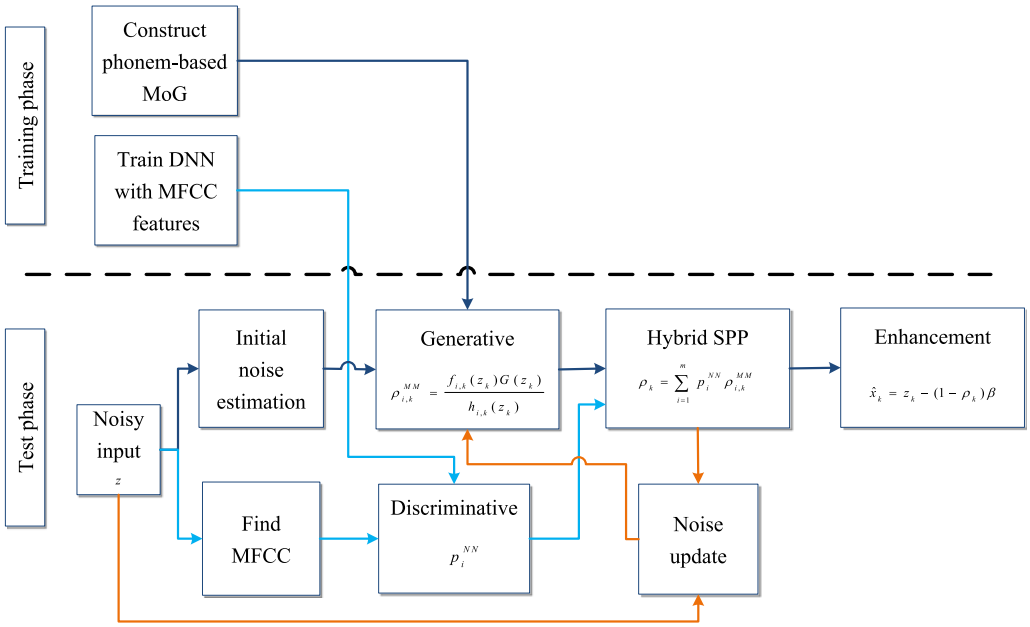

Fig. 1. Block diagram of the NN-MM algorithm.

TABLE I
EXPERIMENTAL SETUP

Train phase		
	Database	Details
Phoneme-based MoG construction	TIMIT (train set)	Log-spectrum vectors
DNN train	TIMIT (train set)	MFCC vectors
Test phase		
	Database	Details
Speech	TIMIT (test set), WSJ	
Noise	NOISEX-92	White, Speech-like, Room, Car, Babble, Factory
SNR	-	-5, 0, 5, 10, 15 dB
Objective measurements	-	PESQ, Composite measure

A. Experimental Setup and Quality Measures

To test the proposed algorithm we have contaminated speech signal with several types of noise from NOISEX-92 database [31], namely *Speech-like*, *Babble*, *Car*, *Room*, *AWGN* and *Factory*. The noise was added to the clean signal drawn from the test set of the TIMIT database (24-speaker core test set), with 5 levels of SNR at -5 dB, 0 dB, 5 dB, 10 dB and 15 dB in order to represent various real-life scenarios. Note, that the train and test sets of TIMIT do not overlap. The algorithm was tested similarly, with the untrained wall street journal (WSJ) database [32]. Table I summarizes the experimental setup. We compared the proposed NN-MM algorithm to the OMLSA algorithm [4] with IMCRA noise estimator [5], a state-of-the-art algorithm for single microphone speech enhancement. The default parameters of the OMLSA were set according to [33].

In order to evaluate the performance of the NN-MM speech enhancement algorithm, several objective and subjective measures were used. The common perceptual evaluation of speech quality (PESQ) measure, which is known to have high correlation with subjective score [34], was used. Additionally, the composite measure, suggested by Hu and Loizou [35], was used. The composite measure weights the log likelihood ratio (LLR), the PESQ and the weighted spectral slope (WSS) [36] to predict the *rating* of the background distortion (Cbak), the speech distortion (Csig) and the overall quality (Covl) performance. The rating is based on the 1-5 mean opinion score (MOS) scale, with clean speech signal achieving MOS value of 4.5.

Finally, we have also carried out informal listening tests with approximately thirty listeners.[3]

B. Setting the Maximum Attenuation Level

In all experiments in the paper, we set β to correspond to attenuation of 20 dB, a value which yielded high noise suppression while maintaining low speech distortion. Fig. 2 illustrates

[3]Audio samples comparing the OMLSA, the original MM and the proposed NN-MM algorithm can be found in www.eng.biu.ac.il/gannot/speech-enhancement/hybrid-dnn-based-single-microphone-speech-enhancement/

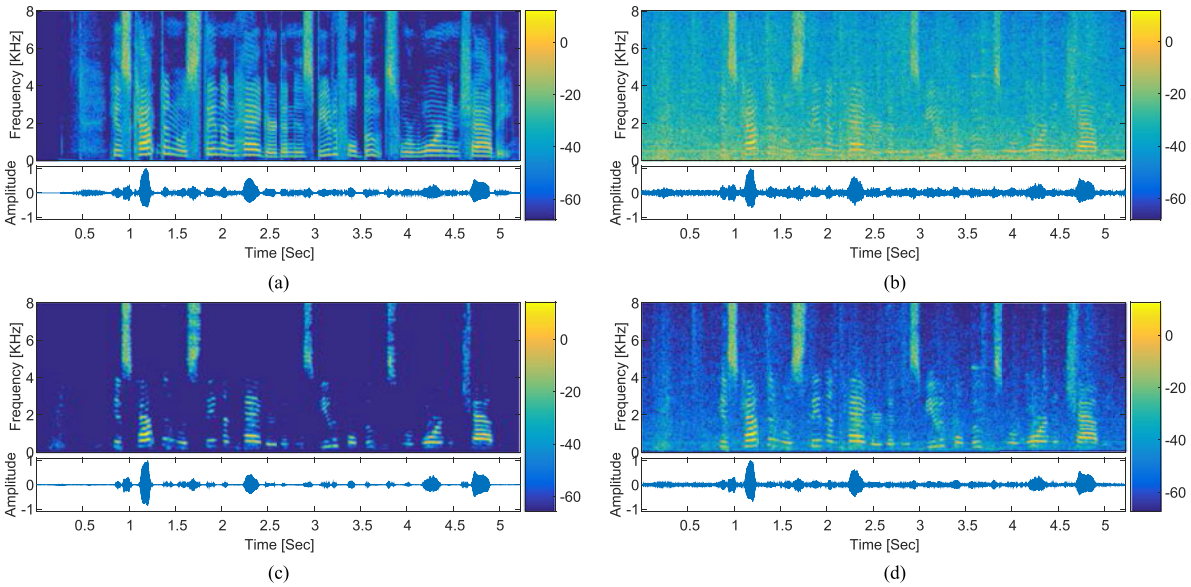

Fig. 2. The influence of excessively large β on the performance of the proposed algorithm. (a) Clean. (b) Factory noise (SNR 5 dB), (PESQ = 1.5). (c) $\beta \sim 60$ dB, (PESQ = 1.7). (d) $\beta \sim 20$ dB, (PESQ = 2.3).

the influence of excessively high attenuation which clearly leads to speech distortion. In Fig. 2(c), the value of β corresponds to an attenuation of 60 dB. It is evident that this excessively high attenuation results in severe speech distortion. Using the nominal β, corresponding with an attenuation of 20 dB, maintains low distortion while sacrificing noise reduction, as evident from Fig. 2(d). This conclusion is also supported by the PESQ measures. While for the high attenuation, the PESQ value is 1.7, for the nominal attenuation level, it increases to 2.3.

C. Objective Results - TIMIT Test Set

We first evaluate the objective results of the proposed NN-MM algorithm and compare it with the results obtained by the state-of-the-art OMLSA algorithm. To further examine the upper bound of the proposed method we also replaced the DNN classifier with an ideal classifier that always provides the correct phoneme, denoted *Oracle*. The test set was the core set of the TIMIT database.

Fig. 3 depicts the PESQ results for all examined algorithms for the Speech-like, Room, Factory and Babble noise types as a function of the input SNR. In Fig. 4 we show the Covl results for Factory and Room noises.

It is evident that the proposed NN-MM algorithm outperforms the OMLSA algorithm in the two designated objective measures. While in high SNR scenarios the difference between the algorithms is only marginal, it is more significant in low SNR. This can be explained as follows. In low SNR case, where the audio signal is very corrupted, using the phoneme classification adds significant information on the structure of the speech signal, consequently improving speech quality. Since the OMLSA algorithm does not use this information it may suffer from low speech quality in low SNR scenarios.

The Oracle naturally outperforms the NN-MM, but the performance difference is rather marginal in high SNR. In low SNR, although there is room for improvement, the performance degradation is not severe. This can be attributed to two factors. The first is the application of the CMVN [25]. Second, even if the algorithm misclassifies the correct phoneme by a phoneme with a similar spectral structure, it can still improve the quality of the enhanced speech.

To gain further insight, we have also compared the enhancement capabilities of the proposed algorithm and the state-of-the-art OMLSA algorithm in the challenging Babble noise environment. For this experiment we have set the SNR to 5 dB. We have encircled in the sonograms, depicted in Fig. 5, areas where the OMLSA exhibit a noticeable level of residual noise. On the contrary, the proposed NN-MM is less prone to this musical noise artifact, while maintaining comparable noise level at the output. The reader is also referred to the sound clips that can be found in our website.

D. Performance With a Different Database

Finally, we would like to demonstrate the capabilities of the proposed NN-MM algorithm when applied to speech signals from other databases. In this work, we have trained the phoneme-based MoG and the DNN using the TIMIT database. In this section, we apply the algorithm to 30 clean signals drawn from the WSJ database [32]. The signals were contaminated by Factory and Room noise, drawn from NOISEX-92 database, with several SNR levels. Note, that the algorithm is not trained with noisy signals, neither it uses any prior knowledge on the noise signal. Fig. 6 depicts the PESQ measure of the NN-MM algorithm in comparison with the OMLSA algorithm (since phoneme transcription of the WSJ database is not available, we do not have results for the oracle estimator). It is evident that the performance of proposed algorithm is maintained even for sentences drawn from a database other than the training database. The results for other noise types, not shown here due to space constraints, are comparable.

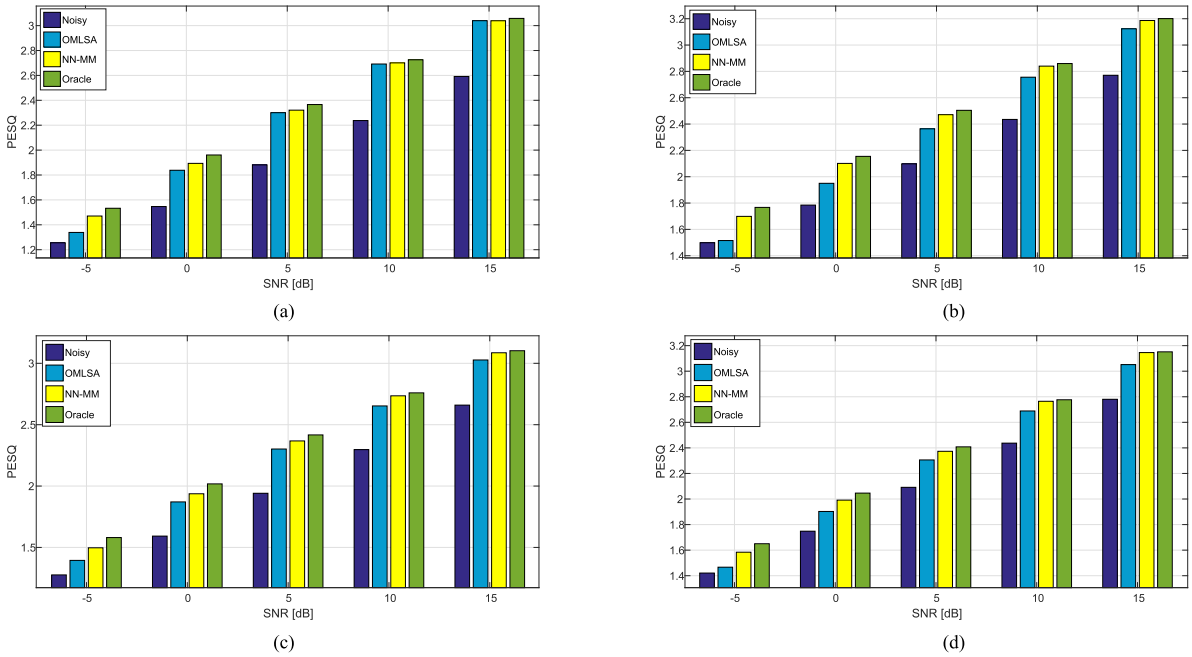

Fig. 3. Speech quality results (PESQ) for several noise types. (a) Speech noise. (b) Room noise. (c) Factory noise. (d) Babble noise.

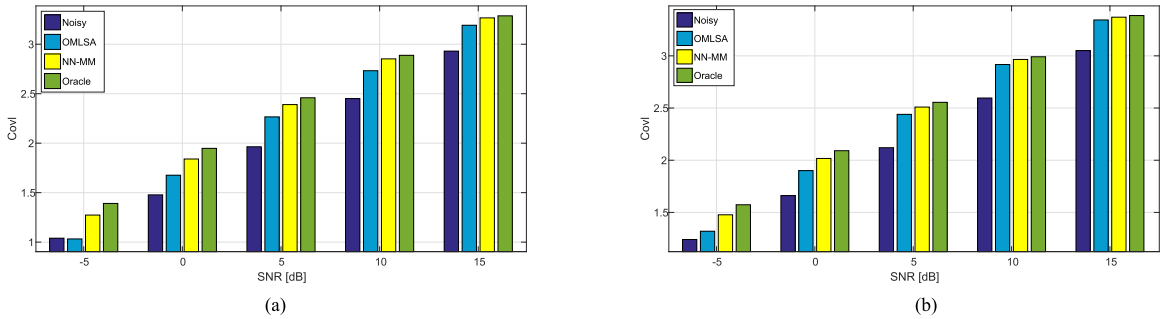

Fig. 4. Results of Covl in two noise types. (a) Factory noise. (b) Room noise.

V. Analysis of the Building Blocks of the Algorithm

In this section, we analyze the individual contributions of each component of the proposed algorithm to the overall performance. First, in Section V-A the two SPP estimators presented in this paper are compared. The advantage of the soft attenuation over the conditional expectation is presented in Section V-B. The importance of the phoneme-based MoG and the DNN phoneme classifier is discussed in Section V-C. Finally, in Section V-D the noise adaptation mechanism is tested in real-life scenario.

A. The Speech Presence Probability: ρ^{MM} Versus $\rho^{NN\text{-}MM}$

One of the major differences between the original MM algorithm [15] and the proposed NN-MM algorithm is the construction of the MoG model and the associated classification procedure. While the former uses unsupervised clustering procedure, based on the EM algorithm, and classifies speech segments using the generative model (10); the latter uses supervised clustering, utilizing the phonemes' transcription, and classifies speech segments using a discriminative approach via the application of DNN (20). Consequently, the clusters in the proposed algorithm consist of different utterances of the same phoneme, while each cluster obtained by the EM algorithm, may consist mixtures of different phonemes. We postulate that the supervised clustering and DNN classification is advantageous over the unsupervised clustering and the generative model. We will examine this claim in the current section, using clean speech signal contaminated by Room noise with SNR $= 5$ dB.

Define, $\mu_{i,k}^{\text{MM}}$ as the centroids of the clusters inferred from the EM algorithm and, respectively, $\mu_{i,k}^{\text{NN-MM}}$ as the centroids of the clusters obtained by the labeled phonemes, both inferred in the training phase. Now, define the *average* PSDs of the speech segment as the weighted average of the centroids of the clusters, namely $\overline{\mu}_k^{\text{MM}}$ and $\overline{\mu}_k^{\text{NN-MM}}$, respectively. The weights of the averaging are given by the associated posterior probabilities (either (10) or (20)), respectively. The average PSD obtained by the unsupervised clustering and the generative model is given by:

$$\overline{\mu}_k^{\text{MM}} = \sum_{i=1}^{m} p_i^{\text{MM}} \mu_{i,k}^{\text{MM}}. \tag{26}$$

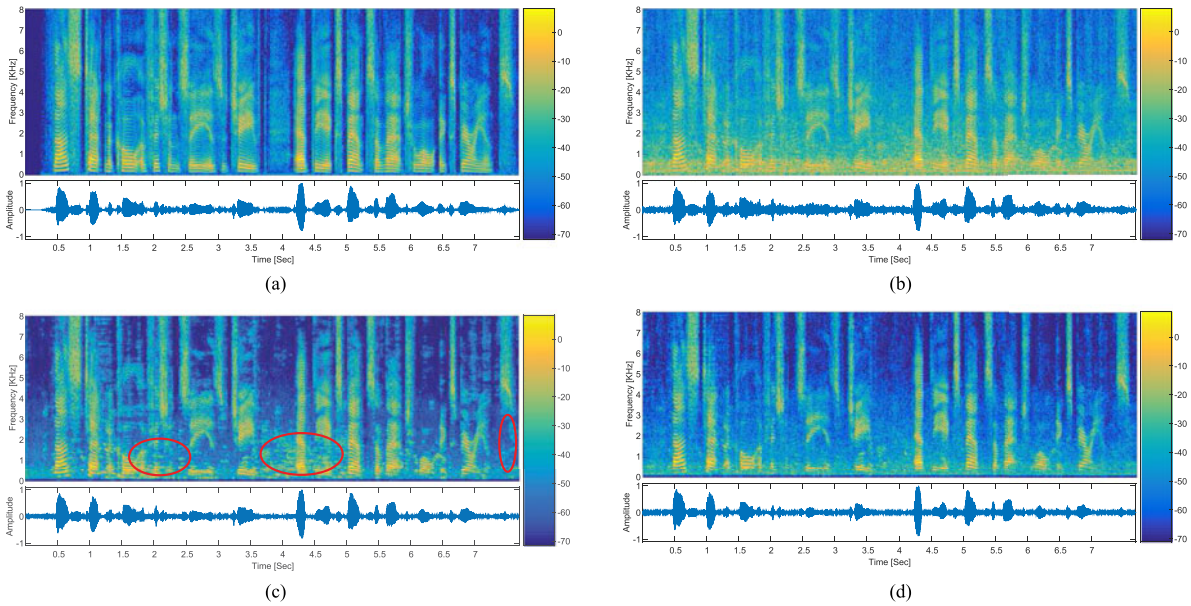

Fig. 5. STFT and time-domain plots of clean, noisy (Babble noise, SNR = 5 dB), and signals enhanced by the OMLSA and the NN-MM algorithms. (a) Clean signal. (b) Noisy signal (PESQ = 2.140). (c) Signal at the output of the OMLSA algorithm (PESQ = 2.251). (d) Signal at the output of the NN-MM algorithm (PESQ = 2.351).

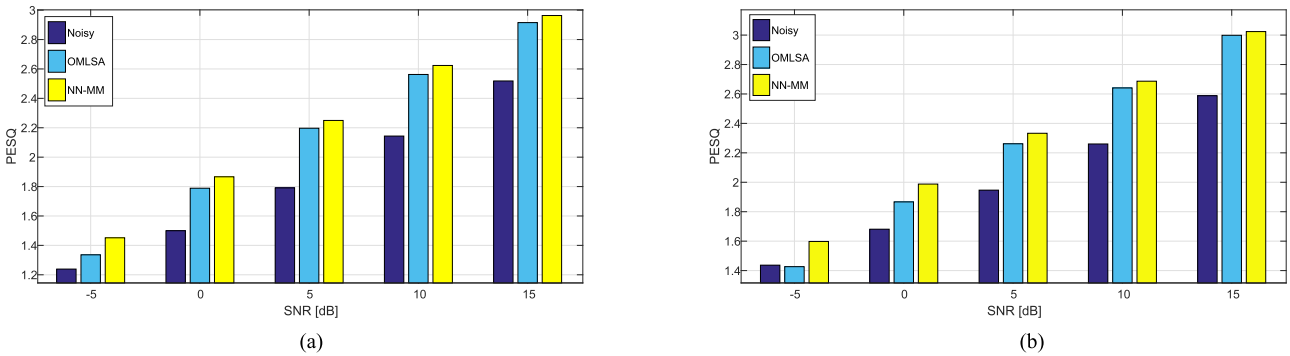

Fig. 6. PESQ results with WSJ database for various SNR levels. (a) Factory noise. (b) Room noise.

Similarly, the average PSD obtained by the supervised clustering and the discriminative DNN are given by:

$$\overline{\mu}_k^{\text{NN-MM}} = \sum_{i=1}^{m} p_i^{\text{NN}} \mu_{i,k}^{\text{NN-MM}}. \tag{27}$$

In Figs. 7(a) and 7(b), the clean and noisy PSDs, respectively, are depicted. Figs. 7(c) and 7(d) illustrates the estimated averaged PSDs inferred from the training data, namely $\overline{\mu}^{\text{NN-MM}}$ and $\overline{\mu}^{\text{MM}}$. It is evident that $\overline{\mu}^{\text{NN-MM}}$ better maintains the envelope of clean speech PSD than $\overline{\mu}^{\text{MM}}$. It is important to stress that the clean speech utterance is drawn from the test set and does not appear in the training set.

Utilization of the phoneme-based MoG plays a central role in the estimation of the SPP, either ρ_k^{MM} or $\rho_k^{\text{NN-MM}}$, and hence in the enhancement attained by the algorithm.

To exemplify this, both SPPs are depicted in Figs. 7(e) and 7(f), respectively. It can be easily observed that $\rho_k^{\text{NN-MM}}$ has a better resemblance to the clean speech sonogram depicted in Fig. 7(a) and that it suffers from less artifacts. Additionally, it is also smoother than ρ_k^{MM}, both along the time and frequency axes. Conversely, vertical narrow spectral lines can be observed in ρ_k^{MM}. This spectral artifacts may be one of the causes for the differences in the enhancement capabilities of the two algorithms, as clearly depicted in Figs. 7(g) and 7(h).

We claim that the observed advantages of the proposed approach stem, at least partly, from the phoneme-based clusters and the better classification capabilities of the DNN. Moreover, the original MM algorithm is only utilizing the current frame for inferring the posterior probabilities, while the proposed algorithm takes into account the context of the phoneme by augmenting past and future frames to the current frame. This guarantees a smoother SPP and consequently less artifacts and *musical noise* at the output of the algorithm.

B. Advantage of the Soft Attenuation Over the Conditional Expectation

In Section V-A, the role of the proposed SPP $\rho^{\text{NN-MM}}$ was presented. It is indicated that the this SPP maintains the harmonic

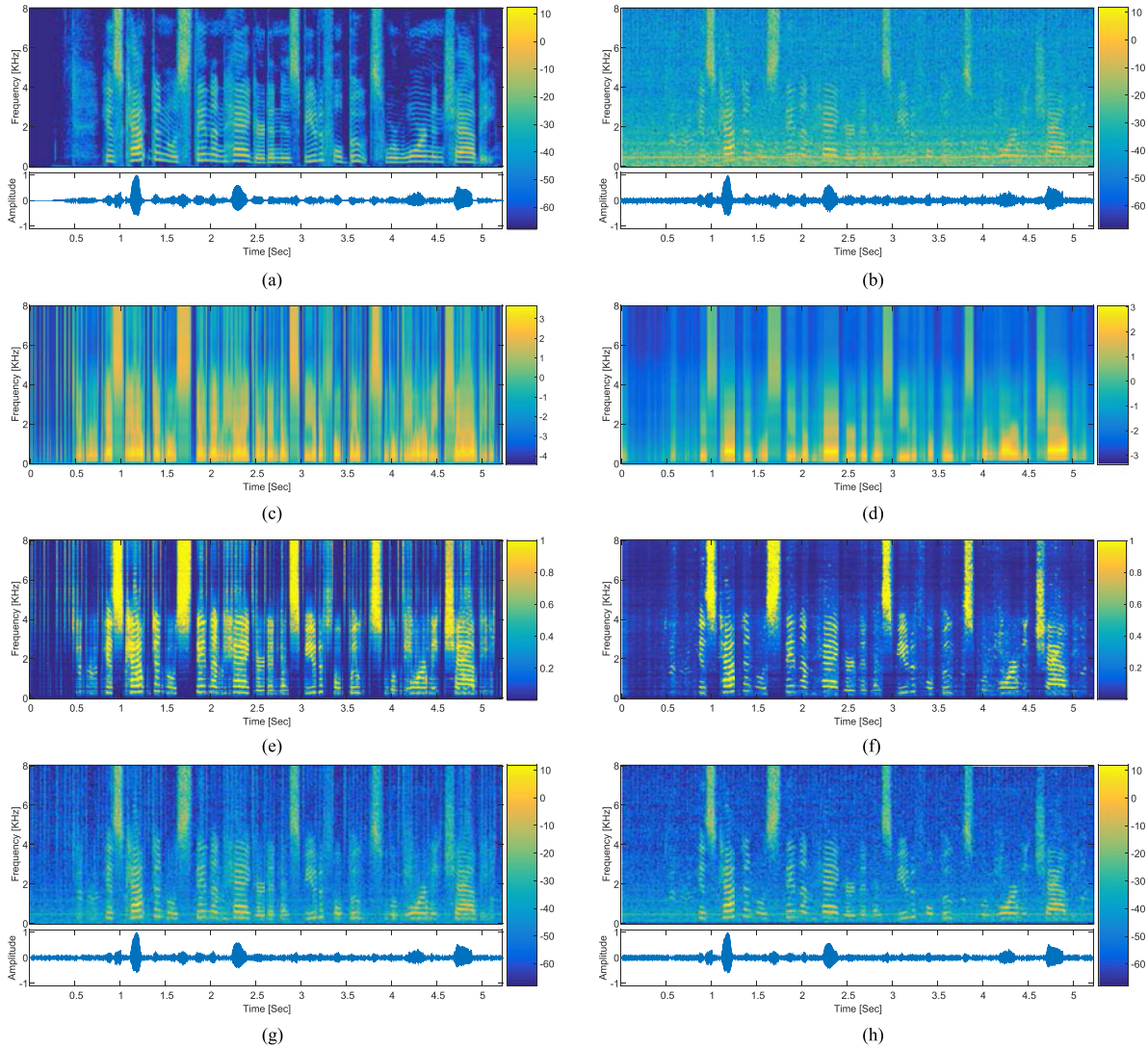

Fig. 7. Sonograms of the clean, noisy and enhanced signals together with the averaged PSD and the SPP using either the NN-MM model or the original MM model. (a) Clean signal. (b) Noisy signal (PESQ = 1.632). (c) $\overline{\mu}^{\text{MM}}$ (Log-spectrum). (d) $\overline{\mu}^{\text{NN-MM}}$ (Log-spectrum). (e) ρ^{MM} parameter. (f) $\rho^{\text{NN-MM}}$ parameter. (g) MixMax Enhanced (PESQ = 2.105). (h) NN-MM Enhanced (PESQ = 2.202).

structure of the voiced speech, i.e. the strong components at the harmonics of the pitch frequency. Consequently, the frequency content of the gaps between the harmonics were correctly classified as noise.

We will examine now the differences between the conditional expectation (16), and the proposed soft attenuation approach (17). We claim that the conditional expectation that substitutes the noisy utterance by an average of the respective cluster (phoneme), cannot accurately describe the gaps between the pitch harmonics of the current utterance, and therefore exhibits limited to noise reduction in these frequencies.

To support this claim we have enhanced a noisy utterance of speech and Factory noise at SNR = 5 dB.

As clearly indicated in Fig. 8, the amount of noise reduction between the harmonics is much higher for the proposed soft attenuator (17) as compared with the attenuation obtained by the conditional expectation (16). This is attributed to the averaging operation of different utterances with different pitch frequencies in (22). We therefore conclude that another important factor of the proposed algorithm is the utilization of the (simpler) soft mask.

C. *Phoneme Classification Task*

In Section V-A we have shown that the SPP obtained by the NN-MM algorithm better suits the clean speech signal than the respective SPP obtained by the original MM algorithm. We would like to have now a closer look at the role of the phoneme classifier in the proposed algorithm. For that, we have implemented two different variants of the MM algorithm. In these variants, we are using the enhancement procedure in (19) and the noise adaptation as explained in Section III-E. The first variant uses the unsupervised EM-based clustering and the generative model for classification (10). This variant is therefore denoted *MM EM-based*. Note, that this variant differs from the original MM algorithm in two major components, namely the noise

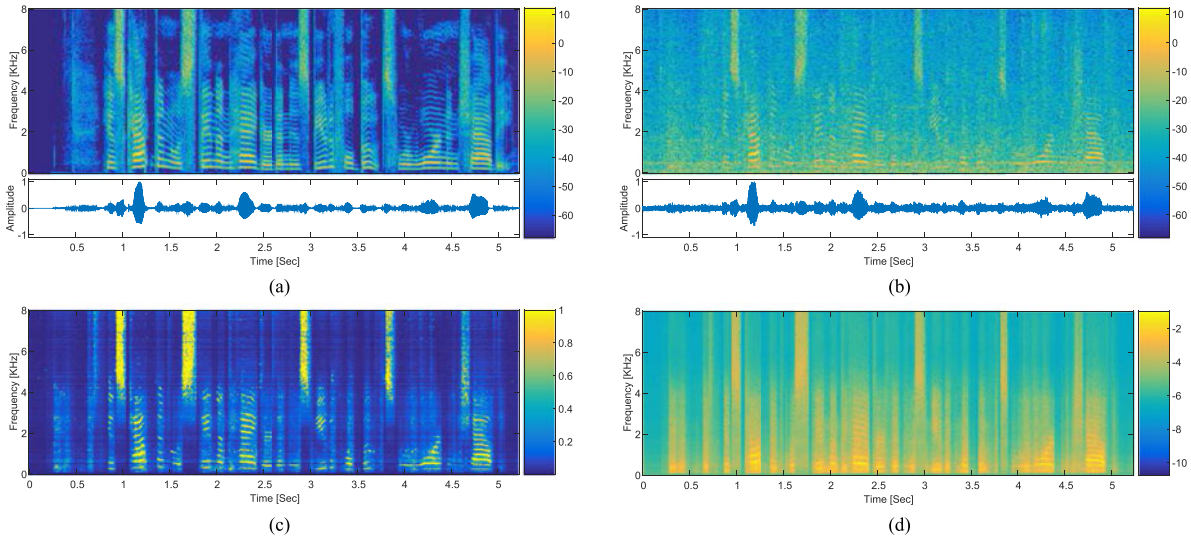

Fig. 8. Advantage of the soft attenuation over the conditional expectation. (a) Clean signal. (b) Noisy signal. (c) $\rho^{\text{NN-MM}}$. (d) $E(X|X < z, \mathbf{Z} = \mathbf{z})$

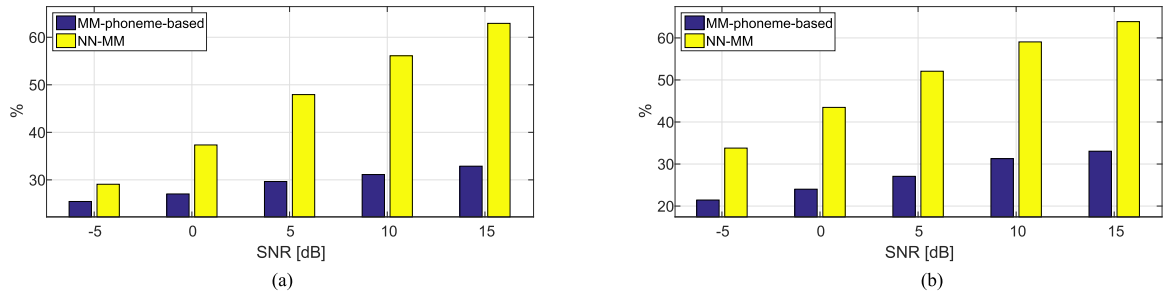

Fig. 9. Results of phoneme classification task performed on noisy data. (a) Factory noise. (b) Babble noise.

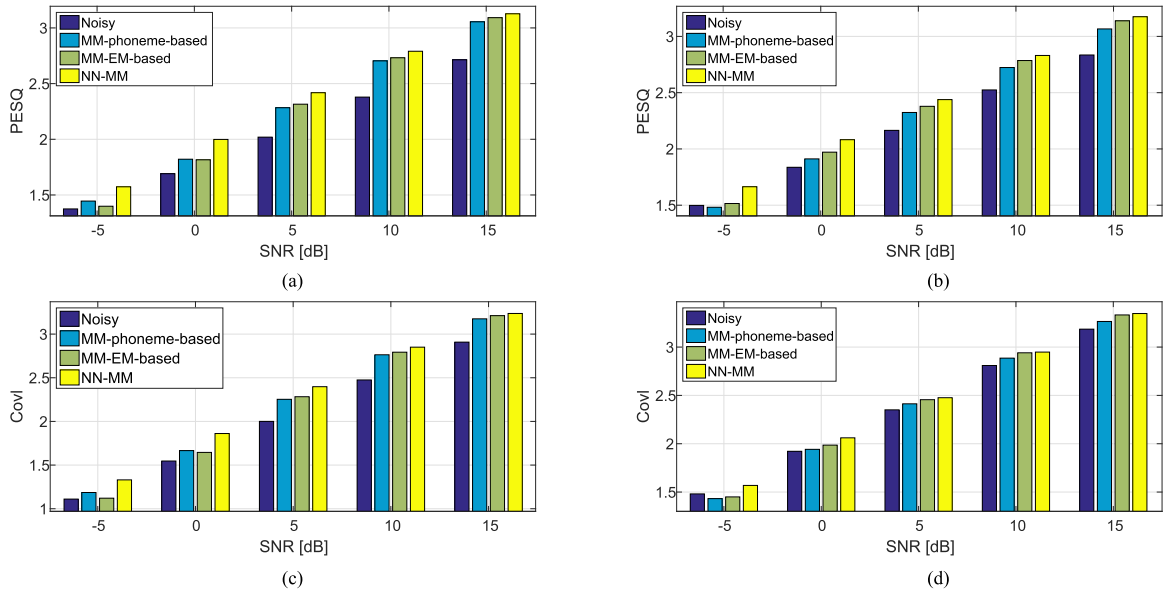

Fig. 10. PESQ and Covl results for several noise types. (a) PESQ results with Factory noise. (b) PESQ results with Babble noise. (c) Covl results with Factory noise. (d) Covl results with Babble noise.

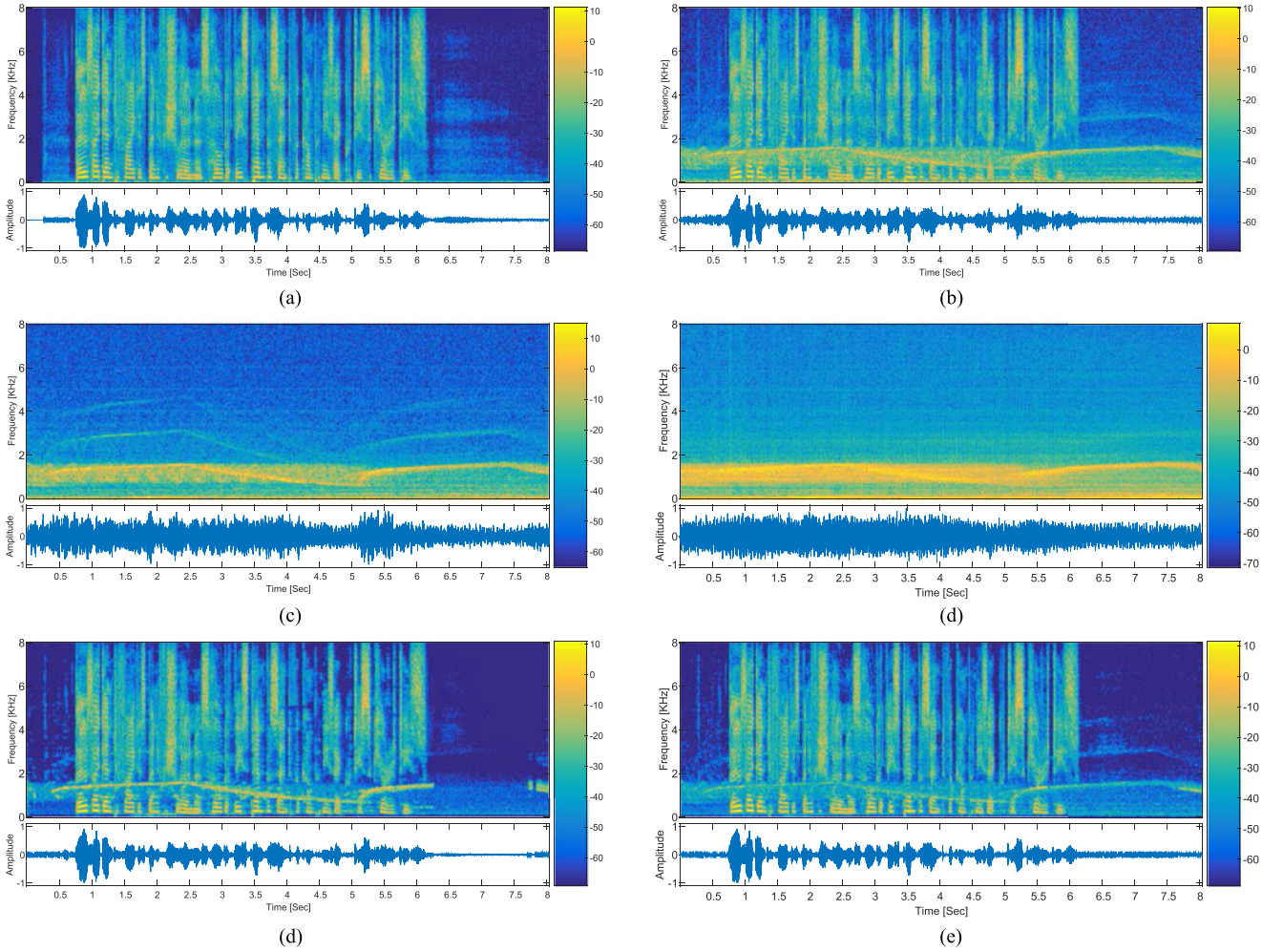

Fig. 11. Noise adaptation capabilities with highly nonstationary siren noise (SNR = 5 dB), and the outputs of the OMLSA and NN-MM algorithms. (a) Clean signal. (b) Noisy signal (SNR = 5 dB) (PESQ = 2.124). (c) Real noise. (d) Estimated noise. (e) OMLSA enhanced (PESQ = 1.847). (f) NN-MM enhanced (PESQ = 2.438).

adaptation and the enhancement procedure. The second variant uses supervised phoneme-based MoG and again (10). This variant is therefore denoted *MM phoneme-based*. These two variants will be compared with the proposed algorithm that uses supervised phoneme-based MoG and a discriminative classification (20).

Clean speech from the test set of the TIMIT database was contaminated by Babble and Factory noise with various SNR levels.

We first compare the success rate in the task of phoneme classification of the MM phoneme-based algorithm and the proposed DNN phoneme classifier. Fig. 9 depicts the percentage of correct classification results obtained on the test data. The results clearly indicate that the DNN-based classifier significantly outperforms the classifier based on the generative model, and is hence better suited for the task at hand.

We turn now to the assessment of the overall performance of the two variants and the proposed NN-MM algorithm as depicted in Fig. 10. It is evident that the proposed algorithm outperforms the two MM variants.

It can therefore be deduced that the phoneme-based model and the associated DNN-based classifier are significant contributors to the performance of the proposed algorithm.

D. The Noise adaptation

In this section we examine the contribution of another important component of the proposed algorithm, namely the noise adaptation scheme described in (25). We chose $\alpha = 0.06$ which approximately refers to memory length of 16 frames. A *city ambiance* noise [37] that consists of a siren and passing cars was chosen, as it is a highly nonstationary noise source with fast PSD changes during the speech utterance. The clean and noisy signals are depicted in Figs. 11(a) and 11(b). The input SNR was set to 5 dB (resulting in an input PESQ = 2.124).

In Fig. 11(c) the real noise STFT is depicted and in Fig. 11(d) its estimate using the proposed adaptation scheme and the SPP inferred by the NN-MM algorithm. It can be observed that the estimate is quite accurate even when the noise PSD changes very fast. Note that during speech dominant time-frequency bins, the noise estimate cannot adapt. These adaptation capabilities are

TABLE II
PESQ RESULTS TO TEST THE SIGNIFICANCE OF THE NOISE ADAPTATION

	PESQ				
SNR [dB]	−5	0	5	10	15
Noisy	1.89	2.03	2.26	2.51	2.81
Without noise adaptation	1.85	3.12	2.45	3.74	3.12
With noise adaptation	**1.95**	**2.21**	**2.64**	**2.97**	**3.28**

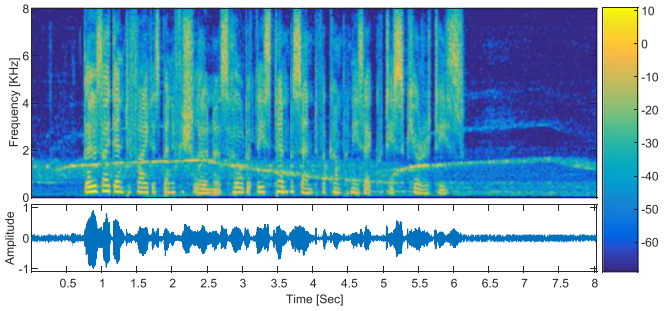

Fig. 12. NN-MM enhanced without noise adaptation, (PESQ = 2.122).

also reflected at the output of the algorithms, especially in comparison with the OMLSA algorithm, as depicted in Figs. 11(e) and 11(f). We observe that the NN-MM algorithm outperforms the OMLSA in reducing this challenging noise. This is also indicated by the PESQ measure. While the OMLSA degrade the speech quality (PESQ = 1.847), the proposed hybrid algorithm slightly improves it (PESQ = 2.438). The reader is also referred to our website where these sound clips can be heard.

To further evaluate the contribution of the noise adaptation to the overall performance of the algorithm, a city ambiance noise was added to clean speech utterances drawn from the TIMIT database with various SNR levels. Table II indicates the PESQ results for two variants of our proposed algorithm, namely with noise adaptation and without noise adaptation. The PESQ results were obtained by averaging over the entire TIMIT test set (approximately 120 sentences). It is clear that the noise adaptation significantly contributes to the performance of the algorithm and to the quality of the output signal.

To further assess the contribution of the noise adaptation we continue with the same experimental setup and examine the output of the NN-MM with the noise adaptation disabled. Fig. 12 illustrates the quality degradation of the enhanced signal (PESQ = 2.122), as compared with Fig. 11(f) depicting the output of the same algorithm with noise adaptation active.

It can be deduced that the noise estimation is another crucial contributor to the overall performance of the proposed algorithm.

VI. CONCLUSION

In this paper a novel speech enhancement scheme, denoted NN-MM, is presented. The proposed algorithm is based on a hybrid scheme which combines phoneme-based generative model for the clean speech signal with a discriminative, DNN-based SPP estimator. In the proposed algorithm, we adopt the advantages of model-based approaches and DNN approaches. While the former usually trades-off noise attenuation capabilities with residual musical noise, the latter often suffer from speech distortion artifacts.

In the proposed algorithm we take advantage of the *discriminative* nature of the DNN that preserves speech smoothness by using context frames. Moreover, the phoneme-based MoG model, where each Gaussian corresponds to a specific phoneme, preserves the general phoneme structure and reduces musical noise.

The proposed algorithm requires neither noise samples nor noisy speech utterances to train. Alternatively, using the embedded DNN-based SPP, allows for fast adaptation to fast-changing noise PSD.

A comprehensive set of experiments demonstrate the capabilities of the proposed algorithm in improving objective quality measures (as can also be verified by various sound examples). The NN-MM algorithm is shown to outperform state-of-the-art algorithm (OMLSA) for both stationary and nonstationary environmental noise and a variety of SNR levels. An elaborated analysis verifies their significant contribution of all the components of the proposed algorithm to the overall performance enhancement.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL, USA: CRC 2013.
[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.
[3] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-33, no. 2, pp. 443–445, Apr. 1985.
[4] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Process.*, vol. 81, no. 11, pp. 2403–2418, 2001.
[5] I. Cohen and B. Berdugo, "Noise estimation by minima controlled recursive averaging for robust speech enhancement," *IEEE Signal Process. Lett.*, vol. 9, no. 1, pp. 12–15, Jan. 2002.
[6] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 65–68, Jan. 2014.
[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2013, pp. 436–440.
[8] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2014.
[9] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 7, pp. 1381–1390, Jul. 2013.
[10] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.
[11] D. Wang, U. Kjems, M. S. Pedersen, J. B. Boldt, and T. Lunner, "Speech intelligibility in background noise with ideal binary time-frequency masking," *J. Acoust. Soc. Am.*, vol. 125, no. 4, pp. 2336–2347, 2009. [Online]. Available: http://scitation.aip.org/content/asa/journal/jasa/125/4/10.1121/1.3083233
[12] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501, 2006. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639306001129
[13] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*, ser. Signals and Communication Technology, G. R. Naik and W. Wang, Eds. Berlin, Germany: Springer, 2014, pp. 349–368. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-55016-4_12

[14] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 7–19, Jan. 2015.
[15] D. Burshtein and S. Gannot, "Speech enhancement using a mixture-maximum model," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 6, pp. 341–351, Sep. 2002.
[16] A. Nádas, D. Nahamoo, and M. Picheny, "Speech recognition using noise-adaptive prototypes," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 10, pp. 1495–1503, Oct. 1989.
[17] Y. Yeminy, S. Gannot, and Y. Keller, "Speech enhancement using a multidimensional mixture-maximum model," in *Proc. Int. Workshop Acoust. Echo Noise Control*, 2010.
[18] S. T. Roweis, "One microphone source separation," in *Proc. Neural Inf. Process. Syst.*, 2000, vol. 13, pp. 793–799.
[19] M. Radfar and R. Dansereau, "Single-channel speech separation using soft mask filtering," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 8, pp. 2299–2310, Nov. 2007.
[20] J. H. Hansen and M. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Process.*, vol. 39, no. 4, pp. 795–805, Apr. 1991.
[21] J. S. Garofolo *et al.*, "TIMIT: Acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, Univ. of Pennsylvania, Philadelphia, PA, USA, 1993.
[22] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.
[23] O. Cappé, "Elimination of the musical noise phenomenon with the ephraim and malah noise suppressor," *IEEE Trans. Speech Audio Process.*, vol. 2, no. 2, pp. 345–349, Apr. 1994.
[24] A. Mohamed, G. E. Dahl, and G. E. Hinton, "Deep belief networks for phone recognition," in *Proc. NIPS Workshop Deep Learning Speech Recog. Related Appl.*, 2009.
[25] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 22, no. 4, pp. 745–777, Apr. 2014.
[26] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 37, no. 11, pp. 1641–1648, Nov. 1989.
[27] L. Tóth, "Phone recognition with deep sparse rectifier neural networks," in *Proc. 2013 IEEE Int. Conf., Acoust., Speech, Signal Process.*, 2013, pp. 6985–6989.
[28] G. Dahl, T. Sainath, and G. Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *Proc. 2013 IEEE Int. Conf., Acoust., Speech, Signal Process.*, May 2013, pp. 8609–8613.
[29] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, Jan. 2009.
[30] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv:1207.0580, 2012.
[31] A. Varga and H. J. 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, 1993.
[32] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proc. Workshop Speech Natural Lang.*, 1992, pp. 357–362. [Online]. Available: http://dx.doi.org/10.3115/1075527.1075614
[33] "MATLAB software for speech enhancement based on optimally modified LSA (OM-LSA) speech estimator and improved minima controlled recursive averaging (IMCRA) noise estimation approach for robust speech enhancement," [Online]. Available: http://webee.technion.ac.il/people/IsraelCohen/
[34] Recommendation P.862, "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," vol. 14, 2001.
[35] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.
[36] D. Klatt, "Prediction of perceived phonetic distance from critical-band spectra: A first step," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, May 1982, vol. 7, pp. 1278–1281.
[37] "City sounds," [Online]. Available: http://soundbible.com/tags-city.html

Shlomo E. Chazan received the B.Sc. and M.Sc. degrees (Cum Laude) in electrical engineering from Bar-Ilan University, Ramat-Gan, Israel, in 2013 and 2015, respectively. He is currently working toward the Ph.D. degree in the Speech and Signal Processing Laboratory, Faculty of Engineering, Bar-Ilan University. His research interests include signal processing and deep learning for speech enhancement and speech processing. He received the Best Student Paper Award at the International Workshop on Acoustic Signal Enhancement (IWAENC) 2016, Xi'an, China.

Jacob Goldberger received the B.Sc. degree in mathematics and computer science from the Bar-Ilan University, Ramat Gan, Israel, the M.Sc. degree (cum laude) in mathematics, and the Ph.D. degree in engineering, both from the Tel-Aviv University, Tel-Aviv, Israel. In 2004, he joined the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel, where he is currently an Associate Professor. His main research interests include machine learning and deep learning methods with applications to information theory, computer vision, medical imaging, speech analysis, and natural language processing.

Sharon Gannot (S'92–M'01–SM'06) received the B.Sc. degree (summa cum laude) from the Technion Israel Institute of Technology, Haifa, Israel, in 1986 and the M.Sc. (cum laude) and Ph.D. degrees from Tel-Aviv University, Haifa, Israel, in 1995 and 2000, respectively, all in electrical engineering. In 2001, he held a Postdoctoral position at the Department of Electrical Engineering, K.U. Leuven, Leuven, Belgium. From 2002 to 2003, he held research and teaching position in the Faculty of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, Israel. He is currently a Full Professor in the Faculty of Engineering, Bar-Ilan University, Ramat-Gan, Israel, where he is heading the Speech and Signal Processing Laboratory and the Signal Processing Track. His research interests include multimicrophone speech processing and specifically distributed algorithms for ad hoc microphone arrays for noise reduction and speaker separation, dereverberation, single microphone speech enhancement, and speaker localization and tracking.

Prof. Gannot received the Bar-Ilan University Outstanding Lecturer Award for 2010 and 2014. He is also a corecipient of seven best paper awards.

He has served as an Associate Editor of the *EURASIP Journal of Advances in Signal Processing* during 2003–2012, and as an Editor of several special issues on multimicrophone speech processing of the same journal. He has also served as a Guest Editor of *ELSEVIER Speech Communication and Signal Processing journals*. He has served as an Associate Editor of IEEE TRANSACTIONS ON SPEECH, AUDIO, AND LANGUAGE PROCESSING during 2009–2013. He is currently a Senior Area Chair of the same journal. He also serves as a Reviewer of many IEEE journals and conferences.

He is a Member of the Audio and Acoustic Signal Processing Technical Committee of the IEEE since January 2010. He currently serves as the Committee Vice-Chair. He is also a Member of the Technical and Steering Committee of the International Workshop on Acoustic Signal Enhancement (IWAENC) since 2005 and was the General Co-Chair of IWAENC held at Tel-Aviv, Israel, in August 2010. He has served as the General Co-Chair of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics in October 2013. He was selected (with colleagues) to present a tutorial sessions in ICASSP 2012, EUSIPCO 2012, ICASSP 2013, and EUSIPCO 2013.