WIDE MATCHING - AN APPROACH TO IMPROVING NOISE ROBUSTNESS FOR SPEECH ENHANCEMENT

Ji Ming, Danny Crookes

School of Electronics, Electrical Engineering and Computer Science
Queen's University Belfast, Belfast BT7 1NN, UK

ABSTRACT

It is shown that under certain conditions it is possible to obtain a good speech estimate from noise without requiring noise estimation. We study an implementation of the theory, namely *wide matching*, for speech enhancement. The new approach performs sentence-wide joint speech segment estimation subject to maximum recognizability to gain noise robustness. Experiments have been conducted to evaluate the new approach with variable noises and SNRs from -5 dB to noise free. It is shown that the new approach, without any estimation of the noise, significantly outperformed conventional methods in the low SNR conditions while retaining comparable performance in the high SNR conditions. It is further suggested that the wide matching and deep learning approaches can be combined towards a highly robust and accurate speech estimator.

Index Terms— Wide matching, noise robustness, speech enhancement, speech recognition

1. INTRODUCTION

Most deep neural network (DNN) systems are based on discriminating relatively short speech segments (typically, of 9 to 31 frames) and hence have limited robustness to untrained noise. For example, the recent DNN-based systems for speech recognition [1–6], speech enhancement [7–10] as well as for image denoising [11] would normally require proper training for the noise types and SNR levels. In this paper, we propose a complementary approach to speech enhancement by modeling very long speech segments, i.e., going wide, with an aim of improving noise robustness without requiring noise training or estimation. We will point out that the new approach and the deep learning approach can be neatly combined towards a highly robust and accurate estimator for estimating speech from noise. Our idea can be best explained by using an oracle experiment.

We took a clean speech database (TIMIT) and expressed each training sentence as a short-time power spectrum (STPS) sequence $S = (s_1, s_2, ..., s_T)$, where s_t is the STPS vector at frame time t. Then we took each core test sentence, added different types of noise (airport, babble, car, restaurant, street and train station) at an SNR of 0 dB, and converted it to a STPS sequence $X = (x_1, x_2, ..., x_\mathcal{T})$, where each noisy STPS vector x_t can be approximately expressed as $x_t = s'_t + n_t$, where s'_t represents the underlying clean speech STPS vector and n_t represents the noise STPS vector. For each noisy frame x_t, we aimed at finding a best matching (clean) speech frame from the training data as an estimate for s'_t. We obtained the estimate, denoted by $\hat{s}'_t$, by maximizing the following *normalized sample correlation coefficient* over all the training data

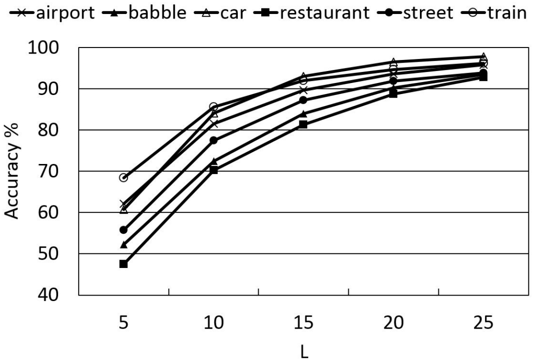

Fig. 1. An oracle experiment showing frame identification accuracy increases with the length of the segments being correlated, for variable test noises at SNR=0 dB, without noise estimation.

$$\hat{s}'_t = \arg\max_{s_\tau} R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L})$$
$$= \max_{s_\tau} \frac{\sum_{l=-L}^{L}(x_{t+l} - m_\mathbf{x})^\mathrm{T}(s_{\tau+l} - m_\mathbf{s})}{\sigma_\mathbf{x}\sigma_\mathbf{s}} \tag{1}$$

where $\mathbf{x}_{t\pm L}$ denotes a segment of noisy frames centered at frame x_t from x_{t-L} to x_{t+L}, $m_\mathbf{x}$ is the mean vector of $\mathbf{x}_{t\pm L}$, and $\sigma_\mathbf{x}$ is the mean-removed Euclidean norm of $\mathbf{x}_{t\pm L}$, i.e., $\sigma^2_\mathbf{x} = \sum_{l=-L}^{L}(x_{t+l} - m_x)^\mathrm{T}(x_{t+l} - m_x)$. The same definition applies to the clean training speech segment $\mathbf{s}_{\tau\pm L}$, with mean vector $m_\mathbf{s}$ and mean-removed Euclidean norm $\sigma_\mathbf{s}$. In this experiment, we included the clean version of the test sentence in the training data (hence the 'oracle'), to examine under what condition the *best* matching estimates would be chosen. Fig. 1 shows the accuracy rates of finding the best matching frames based on (1) for variable segment lengths $2L + 1$, averaged over all the frames of all the core test sentences. We see that as L increased, the best matching estimates using (1) were found with a rapidly increasing probability regardless of the noise. However, the same experiment using other types of distances or likelihoods failed to see a similar trend. This oracle experiment, and the theory below, suggest the potential of a new approach to accurate speech estimation without requiring estimation of the noise if it is independent of the speech. The remainder of the paper is aimed at generalizing the approach to more realistic test speech that is unseen in the training data. When the test speech is unseen and noisy, we concatenate a number of short training segments into full sentences (i.e., the longest possible speech segments for the given noisy sentences) with maximum normalized correlation coefficients, subject to the independence of the noise, to obtain noise-robust speech estimates.

Modeling long speech segments has been an active research topic. This is important because longer speech segments can be

978-1-4799-9988-0/16/$31.00 ©2016 IEEE

distinguished easier in noise. Current methods were able to model some short speech segments, for example, the speech segments corresponding to some phonetic classes [12–14], and the speech segments about 9-31 frames long in DNN systems [6–10]. Our longest matching segment (LMS) approach [15–17] was able to find the presumably longest individual speech segments between the training and test speech that match. However, these segment-based methods process the individual speech segments either independently or with limited correlation (limited by the available training data) as in some recurrent DNNs [1, 2, 4, 5], with limited effect in capturing the longer-distance, cross-segment dependence of speech for speech-to-speech separation in noise. Hence they all require some noise estimation or training. The new approach presented in this paper is radically different: it performs sentence-wide joint speech segment estimation to gain noise robustness, and thus, it has the potential to reduce or remove the need for noise estimation or training as will be shown in our experiments. For convenience, we call the new approach *wide matching*. In the following, we first describe the wide matching theory. Then, we present a method to implement wide matching for unseen test speech.

2. THE WIDE-MATCHING THEORY

Following the same notations as used above, for noisy frames $x_t = s'_t + n_t$, we can decompose the normalized sample correlation coefficient $R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L})$, defined in (1), into two terms

$$R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L}) = \frac{\sigma_{\mathbf{s}'}}{\sigma_{\mathbf{x}}} R(\mathbf{s}'_{t\pm L}, \mathbf{s}_{\tau\pm L}) + \frac{\sigma_{\mathbf{n}}}{\sigma_{\mathbf{x}}} R(\mathbf{n}_{t\pm L}, \mathbf{s}_{\tau\pm L}) \quad (2)$$

where $\mathbf{s}'_{t\pm L}$ represents the underlying clean speech segment in the noisy segment $\mathbf{x}_{t\pm L}$ from s'_{t-L} to s'_{t+L}, and $\mathbf{n}_{t\pm L}$ represents the corresponding noise segment from n_{t-L} to n_{t+L}, with $m_{\mathbf{s}'}$, $m_{\mathbf{n}}$ (implied) and $\sigma_{\mathbf{s}'}$, $\sigma_{\mathbf{n}}$ representing the mean vector and mean-removed Euclidean norm of $\mathbf{s}'_{t\pm L}$ and $\mathbf{n}_{t\pm L}$, respectively. The first term is the normalized sample correlation between the underlying speech segment $\mathbf{s}'_{t\pm L}$ and the training speech segment $\mathbf{s}_{t\pm L}$, weighted by $\sigma_{\mathbf{s}'}/\sigma_{\mathbf{x}}$ which is constant for all the training segments, subject only to the SNR in the observation. The second term is the normalized sample correlation between the noise segment and the training speech segment, weighted by $\sigma_{\mathbf{n}}/\sigma_{\mathbf{x}}$ which is again independent of the training speech segment, subject only to the SNR in the observation. For independent noise and large L, we may assume

$$\begin{aligned} R(\mathbf{n}_{t\pm L}, \mathbf{s}_{\tau\pm L}) &= \frac{\sum_{l=-L}^{L}(n_{t+l} - m_{\mathbf{n}})^{\mathrm{T}}(s_{\tau+l} - m_{\mathbf{s}})}{\sigma_{\mathbf{n}}\sigma_{\mathbf{s}}} \\ &\propto E[(n_t - m_{\mathbf{n}})^{\mathrm{T}}(s_\tau - m_{\mathbf{s}})] \quad (3) \\ &= E[n_t - m_{\mathbf{n}}]^{\mathrm{T}} E[s_\tau - m_{\mathbf{s}}] = 0 \quad (4) \end{aligned}$$

where (3) is based on the assumption that as the observation times (i.e., L) become large, the time average converges to the ensemble average (here we assume ergodicity for both the speech and noise processes [18]); (4) is based on the assumption that the training speech and noise are statistically independent. With (2)–(4), thus, for large L and independent noise, we may have

$$\max_{s_\tau} R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L}) \propto \max_{s_\tau} R(\mathbf{s}'_{t\pm L}, \mathbf{s}_{\tau\pm L}) \quad (5)$$

That is, the maximum correlation (i.e., the matching accuracy) could become independent of the noise but depends only on the two speech segments (one underlying and the other a potential estimate) being compared. This theory is in good agreement with the experimental results in Fig. 1. It suggests the potential to obtain estimates of speech without requiring noise estimation. For this, two conditions must be met: 1) the speech $\mathbf{s}'_{t\pm L}$ being correlated is long (i.e., L is large), and 2) the estimate $\mathbf{s}_{\tau\pm L}$ is independent of the noise.

3. WIDE MATCHING FOR SPEECH ENHANCEMENT

3.1. A constrained maximization problem

Let $X = (x_1, x_2, ..., x_T)$ be a noisy test sentence with the underlying speech sentence $S' = (s'_1, s'_2, ..., s'_T)$ unseen in the training data. We seek an approach to concatenating a number of short training speech segments into a full sentence as an estimate of S'. In the approach, the optimal element training segments are estimated jointly to maximize the sentence-wide correlation with the noisy sentence X. Given a noisy sentence, performing the sentence-wide correlation maximizes the length (i.e., L) of the speech signal to be correlated and hence the robustness to independent noise, i.e., to best fulfil Condition 1 as required in the above theory.

Suppose we can divide X into some K consecutive segments, denoted by $\mathbf{X} = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, ..., \mathbf{x}_{t_K})$, where each segment $\mathbf{x}_{t_k}$ is centered at some frame time t_k with frames from $x_{t_k-\gamma}$ to $x_{t_k+\gamma}$, where γ defines the length of the element segment. For simplicity, we assume a common γ for all the element segments and so γ can be implied in the expression. Adjacent element segments can have some overlap to improve the smoothness. In a similar way, denote by $\mathbf{S} = (g_{\tau_1}\mathbf{s}_{\tau_1}, g_{\tau_2}\mathbf{s}_{\tau_2}, ..., g_{\tau_K}\mathbf{s}_{\tau_K})$ a chain of K clean training segments as an estimate of the underlying speech sentence in $\mathbf{X}$, where each element training segment $\mathbf{s}_{\tau_k}$ consists of consecutive frames from $s_{\tau_k-\gamma}$ to $s_{\tau_k+\gamma}$, and g_{τ_k} is the gain of the element training segment in forming the sentence estimate. In $\mathbf{S}$, different training segments $\mathbf{s}_{\tau_k}$ can come from different training sentences/contexts to simulate unseen test speech. We estimate the optimal $\mathbf{S}$ based on the sentence-wide, normalized sample correlation coefficient between $\mathbf{X}$ and $\mathbf{S}$. After some manipulation, this can be written as

$$\begin{aligned} R(\mathbf{X}, \mathbf{S}) &= R(\mathbf{x}_{t_1}\mathbf{x}_{t_2}...\mathbf{x}_{t_K}, g_{\tau_1}\mathbf{s}_{\tau_1}g_{\tau_2}\mathbf{s}_{\tau_2}...g_{\tau_K}\mathbf{s}_{\tau_K}) \\ &= \frac{\sum_{k=1}^{K} g_{\tau_k} \sum_{l=-\gamma}^{\gamma} x_{t_k+l}^{\mathrm{T}} s_{\tau_k+l} - L m_{\mathbf{X}}^{\mathrm{T}} m_{\mathbf{S}}}{\sigma_{\mathbf{X}}\sigma_{\mathbf{S}}} \quad (6) \end{aligned}$$

where $L = (2\gamma + 1)K$ is the sample length of the two full sentences being correlated, $m_{\mathbf{S}}$ and $\sigma_{\mathbf{S}}$ are the global mean vector and mean-removed Euclidean norm of the training segment chain $\mathbf{S}$,

$$m_{\mathbf{S}} = \frac{1}{L}\sum_{k=1}^{K} g_{\tau_k} \sum_{l=-\gamma}^{\gamma} s_{\tau_k+l} \quad (7)$$

$$\sigma_{\mathbf{S}}^2 = \sum_{k=1}^{K} g_{\tau_k}^2 \sum_{l=-\gamma}^{\gamma} s_{\tau_k+l}^{\mathrm{T}} s_{\tau_k+l} - L m_{\mathbf{S}}^{\mathrm{T}} m_{\mathbf{S}} \quad (8)$$

The above expressions apply to $m_{\mathbf{X}}$ and $\sigma_{\mathbf{X}}$, the global mean vector and mean-removed Euclidean norm of the noisy segment sequence $\mathbf{X}$ (without the gain terms). In the above sentence-wide correlation coefficient $R(\mathbf{X}, \mathbf{S})$, there is no assumption about the independence between the speech frames or spectral coefficients within the element segments, across the element segments or anywhere in the sentence.

Superficially, one may obtain an estimate of the optimal $\mathbf{S}$ by maximizing the normalized correlation coefficient $R(\mathbf{X}, \mathbf{S})$ over all possible chains of the training segments $g_{\tau_k}\mathbf{s}_{\tau_k}$. However, not all of the chains constitute realistic speech; some chains with larger $R(\mathbf{X}, \mathbf{S})$ may simulate the original noisy speech $\mathbf{X}$ well (as indicated in (2), the correlation coefficient $R(\mathbf{X}, \mathbf{S})$ for perfectly matching speech $\mathbf{S}'$ and $\mathbf{S}$ is confined around $\sigma_{\mathbf{S}'}/\sigma_{\mathbf{X}} < 1$ for noisy speech). These false positives can happen when the element segments are very short and hence some noisy speech may be simulated well by randomly chaining some very short speech segments. To make the estimate to be valid speech, which is independent of the noise and hence fulfils Condition 2 of the above theory, we use the estimate's *recognizability*, to a speech recognizer trained with *clean* speech data, to regularize the formation of the optimal estimate $\mathbf{S}$. Thus, we can express the problem to obtain an optimal speech estimate as the constrained maximization of the normalized correlation subject to the *maximum recognizability* of the estimate

$$\hat{\mathbf{S}} = \arg\max_{\mathbf{S}}[\log R(\mathbf{X},\mathbf{S}) + \lambda \log H(\mathbf{S})] \tag{9}$$

where $H(\mathbf{S})$ represents the confidence score of the estimate $\mathbf{S}$ to be valid speech, and λ is a Lagrange multiplier. For the proof of the concept, this paper uses an HMM-based phone recognizer to provide $H(\mathbf{S})$. The recognizer is trained with clean speech and learns the acoustic HMMs for context-independent phones, a bigram phone language model, and the duration probability distributions of each HMM state and monophone. The log likelihood score given by the recognizer for a given $\mathbf{S}$ can be stated as

$$\log H(\mathbf{S}) = [\log h(\mathbf{S}) + \sum_{i=1}^{I} \log p_i(d_i) + \sum_{u=1}^{U} \log p_u(d_u)]/L \tag{10}$$

where $h(\mathbf{S})$ denotes the likelihood score of $\mathbf{S}$ given by the Viterbi search, I and U are the numbers of HMM states and phones through which the best path traversed, p_i and p_u are the duration probability distributions of those states and phones, and d_i and d_u are the durations spent in each state and phone, respectively. It is assumed that among all possible training segment chains $\mathbf{S}$, the chains constituting valid clean speech are most recognizable to the recognizer, in terms of achieving large scores $H(\mathbf{S})$ (this is because valid clean speech is most likely to simultaneously fulfil the acoustic, phone language, state duration and phone duration constraints of clean speech learned by the recognizer). If such a sentence-long chain with a large noise-independent speech confidence score $H(\mathbf{S})$ *simultaneously* has a large correlation coefficient $R(\mathbf{X},\mathbf{S})$ with the noisy signal $\mathbf{X}$, or vice versa, then it can be assumed that this is an optimal estimate of the underlying speech in $\mathbf{X}$. Hence we have (9).

3.2. An iterative estimation algorithm

We use a computationally efficient iterative algorithm to solve the above constrained maximization problem (9), which seeks a sentence-wide joint estimation of the element training segments to form the optimal speech sentence estimate $\hat{\mathbf{S}}$. Given a noisy sentence $\mathbf{X}$, we start with an initial estimate $\hat{\mathbf{S}}$ by separately estimating each element training segment $\mathbf{s}_{\tau_k}$ based on maximizing the segment-level correlation coefficient $R(\mathbf{x}_{t_k}, \mathbf{s}_{\tau_k})$ with a unit gain g_{τ_k}. Then we update this initial estimate by alternately re-estimating each element training segment with gain to maximize the sentence-wide constrained correlation coefficient (9); in re-estimating a specific element training segment, the other element training segments are fixed to their latest estimates. This alternate re-estimation process is iterated until convergence is achieved. For example, consider re-estimating the element training segments $g_{\tau_k}\mathbf{s}_{\tau_k}$ in the order from $k = 1$ to K. In the jth iteration, to obtain a new estimate of the optimal kth element training segment, denoted by $\hat{g}^j_{\tau_k}\hat{\mathbf{s}}^j_{\tau_k}$, we maximize (9) with respect to $g_{\tau_k}\mathbf{s}_{\tau_k}$, with the succeeding element training segments $g_{\tau_m}\mathbf{s}_{\tau_m}$ $(m > k)$ from the $(j-1)$th iteration, and the preceding element training segments $g_{\tau_m}\mathbf{s}_{\tau_m}$ $(m < k)$ from the jth iteration. Therefore in the jth iteration and kth stage, the optimal speech sentence estimate to be determined can be expressed as $\hat{\mathbf{S}}^j(g_{\tau_k}\mathbf{s}_{\tau_k}) = (\hat{g}^j_{\tau_1}\hat{\mathbf{s}}^j_{\tau_1}, ..., \hat{g}^j_{\tau_{k-1}}\hat{\mathbf{s}}^j_{\tau_{k-1}}, g_{\tau_k}\mathbf{s}_{\tau_k}, \hat{g}^{j-1}_{\tau_{k+1}}\hat{\mathbf{s}}^{j-1}_{\tau_{k+1}}, ..., \hat{g}^{j-1}_{\tau_K}\hat{\mathbf{s}}^{j-1}_{\tau_K})$, which is only a function of $g_{\tau_k}\mathbf{s}_{\tau_k}$, with the rest of the element training segments fixed to their latest optimal estimates from the appropriate iterations. The optimal sentence estimate can be obtained as follows

$$\begin{aligned}
&\hat{\mathbf{S}}^j(\hat{g}^j_{\tau_k}\hat{\mathbf{s}}^j_{\tau_k}) \\
&\quad = \arg\max_{g_{\tau_k}\mathbf{s}_{\tau_k}} [\log R(\mathbf{X}, \hat{\mathbf{S}}^j(g_{\tau_k}\mathbf{s}_{\tau_k})) + \lambda \log H(\hat{\mathbf{S}}^j(g_{\tau_k}\mathbf{s}_{\tau_k})] \\
&\qquad\qquad k = 1, 2, ..., K; j = 1, 2, ...
\end{aligned} \tag{11}$$

with $\hat{g}^0_{\tau_k}\hat{\mathbf{s}}^0_{\tau_k}$ corresponding to the initial estimates. Eq. (11) represents an iterative algorithm to implement the sentence-wide joint training segment estimation defined in (9). It manages to estimate the element training segments one at a time, subject to the constraints of all the other segments in the sentence, and hence can be calculated efficiently. It can be shown that this algorithm converges in terms of generating a speech sentence estimate that increases the constrained correlation coefficient with each iteration. Details are given below.

4. EXPERIMENTAL STUDIES

Experiments have been conducted to evaluate the proposed wide-matching approach for noisy speech enhancement, with a focus on its performance without any estimation of the noise. The TIMIT database was used in the experiments, which contains a training set with 3696 speech sentences from 462 speakers (326 male, 136 female), and a core test set with 192 speech sentences from 24 speakers (16 male, 8 female). There are no common speakers and sentence texts between the training set and test set. The test set was added with variable noises to form the unseen noisy test data.

Six different types of noise: airport, babble, car, restaurant, street and train station, taken from Aurora 4 [19], were added to each test sentence at four different SNRs: 10, 5, 0 and -5 dB, respectively, measured on each sentence basis. The signals were sampled at 16 kHz and divided into frames of 25 ms with a frame rate of 10 ms. Each frame was represented by a 40-coefficient short-time power spectral (STPS) vector, taken from the output of a 40-channel Mel-frequency filterbank. We formed the element training speech segments used to perform the sentence-wide correlation and speech estimation by taking each training frame in each training sentence and forming a segment around the frame with a fixed length of 11 frames (i.e., $\gamma = 5$ in (6), a figure borrowed from the previous DNN-based studies [20]). The noisy test sentences were each divided into a sequence of consecutive segments each with the same length of 11 frames and with 8-frame overlap between adjacent segments. As indicated in (9) or (11), the underlying speech is estimated by performing sentence-wide correlation with the noisy sentences subject to maximum recognizability. Table 1 shows the statistics of the length $L = (2\gamma + 1)K$ of the test signals $\mathbf{X}$ that have been correlated to derive the speech estimates $\hat{\mathbf{S}}$, for the 192 test sentences. We take the overlapping frames between successive element segments as effective signals as we found that some overlap did help improve

Table 1. Minimum, maximum and average sample length $L = (2\gamma + 1)K$ of the test sentences being correlated (unit: frame).

Min	Max	Average
440	2233	1023

Table 2. Comparing the new wide-matching method with conventional methods on the Segmental SNR, PESQ and STOI measures.

	Method/SNR(dB)	-5	0	5	10	clean
S	Unprocessed	-6.79	-4.26	-1.11	2.48	
e	LogMMSE	-2.59	-0.31	2.23	4.96	19.28
g	LogMMSE-SPU	-1.31	0.43	2.71	5.35	19.00
	Wiener filtering	-3.94	-1.42	1.40	4.38	19.69
S	KLT	-1.62	0.36	2.64	5.05	15.28
N	PKLT	-1.52	0.46	2.85	5.34	15.29
R	*Wide matching*	0.51	2.33	3.96	5.16	18.11
	Unprocessed	1.39	1.74	2.09	2.44	
P	LogMMSE	1.60	2.01	2.38	2.72	4.39
E	LogMMSE-SPU	1.26	1.73	2.18	2.58	4.37
S	Wiener filtering	1.53	1.93	2.31	2.67	4.41
Q	KLT	1.25	1.75	2.21	2.63	4.32
	PKLT	0.97	1.50	2.00	2.48	4.31
	Wide matching	1.76	2.19	2.54	2.79	4.22
	Unprocessed	0.58	0.69	0.79	0.88	
S	LogMMSE	0.54	0.67	0.78	0.86	0.99
T	LogMMSE-SPU	0.51	0.63	0.74	0.84	0.99
O	Wiener filtering	0.56	0.68	0.79	0.88	0.99
I	KLT	0.56	0.70	0.81	0.89	0.99
	PKLT	0.50	0.67	0.79	0.88	0.99
	Wide matching	0.68	0.79	0.86	0.90	0.98

the estimation accuracy, as in the DNN-based speech recognition. The large L contributed importantly to improving noise robustness without requiring noise estimation, to be shown below.

To form the recognizability constraint (10), we trained a simple HMM-based recognizer using the 3696 training sentences. The recognizer contains 61 3-state HMMs for the TIMIT monophones, and a bigram phone language model trained with the phonetic transcripts of the training sentences. It also contains a state-duration probability distribution for each of the 183 states and a phone-duration probability distribution for each of the 61 monophones, expressed as the appropriate histograms. In using the iterative algorithm (11) to derive the optimal element segment estimates, we assumed that each local segment gain (i.e., g_{τ_k}) could only change within the range [0.5, 2]. For each possible element training segment, we used a fast algorithm (the golden section method) to search its optimal gain within the range to maximize the sentence-wide constrained correlation. The extra computation was found to be minimal. Unless otherwise indicated, the following experiments were conducted with the constraining Lagrange multiplier $\lambda = 0.1$ in (11). We stopped the iteration when no change was found between successive estimates. When an optimal speech sentence estimate $\hat{\mathbf{S}}$ was obtained, the corresponding training speech frames were used to form optimal frequency-domain filters applied to the appropriate noisy speech frames to reconstruct the speech waveform, with the phase spectra taken from the noisy speech. The same reconstruction procedures were used in [15] [17].

Table 2 summarizes the comparisons of the wide-matching method against five conventional speech enhancement methods on the Segmental SNR, PESQ and STOI [21] measures, respectively, as a function of the input test sentence SNR averaged over 1152 test sentences (i.e., 192 test sentences per noise type × 6 noise types) under each SNR condition. The wide-matching method did not use any noise estimation while the conventional methods, LogMMSE [22], LogMMSE-SPU [23], Wiener filtering [24], KLT [25] and Perceptual KLT [26], each used an algorithm to estimate the noise. The proposed wide-matching method significantly outperformed all the conventional methods on all the three measures for each noise type in all the low SNR conditions, with only a slight drop in performance in the high-SNR (e.g., 10 dB or clean) conditions compared to the conventional methods. Importantly, wide matching improved the PESQ and STOI scores over those of the unprocessed noisy speech in all the noisy conditions.

Table 3. The importance of the correlation length and recognizability constraint (unconstrained when $\lambda = 0$), with SNR=0 dB.

Method/Measure	Segmental SNR	PESQ	STOI
Segment matching	1.01	1.82	0.69
Wide matching, $\lambda = 0$	1.47	1.94	0.73
Wide matching, $\lambda = 0.1$	2.33	2.19	0.79
Wide matching, $\lambda = 0.2$	2.34	2.18	0.79
Wide matching, $\lambda = 0.3$	2.34	2.17	0.79

Table 3 uses an example (SNR=0 dB) to show the importance of the length of the signals being correlated in improving noise robustness without noise estimation. It shows a comparison between the sentence-wide correlation (i.e., (11)) with a sentence-dependent length as shown in Table 1, and the segment-level correlation with a fixed length of 11 frames, averaged over 1152 test sentences from the six noise types. The segment-level correlation assumes independence between the element segments and was used to provide the initial estimates for iteration in (11). Table 3 also shows the importance of the recognizability constraint in wide matching to help obtain noise-independent speech estimates. Poorer-quality enhanced speech was obtained without this constraint (i.e., the Lagrange multiplier $\lambda = 0$ in (11)). Finally, Table 3 shows the stability of the proposed wide matching algorithm for a range of λ values.

Finally, Fig. 2 summarizes the convergence of the iterative algorithm (11), showing the average numbers of iteration used in the estimation, and the end-to-end values of the iteration of the constrained correlation and the corresponding sample correlation $R(\mathbf{X}, \mathbf{S})$, respectively, averaged overall all the test sentences and noise types.

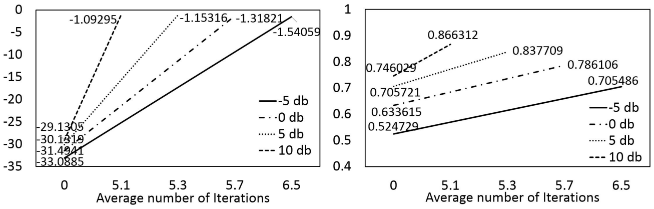

Fig. 2. Increases in the constrained correlation (left, log scale) and the corresponding sample correlation (right, linear scale) with iteration, as a function of the test sentence SNR.

5. CONCLUDING REMARKS

A new method, namely wide matching, was presented that performs sentence-wide joint speech segment estimation to improve noise robustness. Experimental results indicate that the new method has the potential to significantly outperform conventional methods without requiring noise estimation. The new method can be neatly combined with the deep learning method with mutual benefits. For example, a DNN-based speech recognizer can be used to replace the HMM-based one to provide more accurate constraint on the speech estimates, while the DNN-based recognizer can also benefit from the combination by having improved robustness to untrained noise.

6. REFERENCES

[1] A.L. Maas, Q.V. Le, T.M. ONeil, O. Vinyals, P. Nguyen, and A.Y. Ng, "Recurrent neural networks for noise reduction in robust asr," in *Proceedings Interspeech*. ISCA, 2012, pp. 22–25.

[2] P. Brakel, D. Stroobandt, and B. Schrauwen, "Bidirectional truncated recurrent neural networks for efficient speech denoising," in *Proceedings Interspeech*. ISCA, 2013, pp. 2973–2977.

[3] M.L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2013, pp. 7398–7402.

[4] C. Weng, D. Yu, S. Watanabe, and B.H. Juang, "Recurrent deep neural networks for robust speech recognition," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 5569–5573.

[5] J.T. Geiger, Z. Zhang, F. Weninger, B. Schuller, and G. Rigoll, "Robust speech recognition using long short-term memory recurrent neural networks for hybrid acoustic modelling," in *Proceedings Interspeech*. ISCA, 2014, pp. 631–635.

[6] F. Li, P.S. Nidadavolu, and H. Hermansky, "A long, deep and wide artificial neural net for robust speech recognition in unknown noise," in *Proceedings Interspeech*. ISCA, 2014, pp. 358–362.

[7] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in *Proceedings Interspeech*. ISCA, 2013, pp. 436–440.

[8] D. Liu, P. Smaragdis, and M. Kim, "Experiments on deep learning for speech denoising," in *Proceedings Interspeech*. ISCA, 2014, pp. 2685–2689.

[9] B.Y. Xia and C.C. Bao, "Wiener filtering based speech enhancement with weighted denoising auto-encoder and noise classification," *Speech Communication*, vol. 60, pp. 13–29, 2014.

[10] Y. Xu, J. Du, L.R. Dai, and C.H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 7–19, 2015.

[11] Y.Q. Wang and J.M. Morel, "Can a single image denoising neural network handle all levels of gaussian noise?," *IEEE Signal Process. Lett.*, vol. 21, pp. 1150–1153, 2014.

[12] X. Xiao, P. Lee, and R.M. Nickel, "Inventory based speech enhancement for speaker dedicated speech communication systems," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2009, pp. 3877–3880.

[13] B. Raj, R. Singh, and T. Virtanen, "Phoneme-dependent nmf for speech enhancement in monaural mixtures," in *Proceedings Interspeech*. ISCA, 2011, pp. 1217–1220.

[14] R.M. Nickel, R.F. Astudillo, D. Kolossa, and R. Martin, "Corpus-based speech enhancement with uncertainty modeling and cepstral smoothing," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 983–997, 2013.

[15] J. Ming, R. Srinivasan, and D. Crookes, "A corpus-based approach to speech enhancement from nonstationary noise," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 822–836, 2011.

[16] J. Ming, R. Srinivasan, D. Crookes, and A. Jafari, "Close - a data-driven approach to speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1355–1368, 2013.

[17] J. Ming and D. Crookes, "An iterative longest matching segment approach to speech enhancement with additive noise and channel distortion," *Computer Speech and Language*, vol. 28, pp. 1269–1286, 2014.

[18] B.V. Gnedenko, *The Theory of Probability (translated from the Russian by B.D. Seckler)*, Chelsea, N.Y., 1962.

[19] G. Hirsch, "Experimental framework for the performance evaluation of speech recognition front-ends on a large vocabulary task," in *Version 2.0, STQ-Aurora DSR Working Group*. ETSI, November 2002.

[20] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 14–22, 2012.

[21] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, pp. 2125–2136, 2011.

[22] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 23, pp. 443–445, 1985.

[23] I. Cohen, "Optimal speech enhancement under signal presence uncertainty using log-spectra amplitude estimator," *IEEE Signal Process. Lett.*, vol. 9, pp. 113–116, 2002.

[24] P. Scalart and J. Filho, "Speech enhancement based on a priori signal to noise estimation," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 1996, pp. 629–632.

[25] Y. Hu and P. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 334–341, 2003.

[26] F. Jabloun and B. Champagne, "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 700–708, 2003.