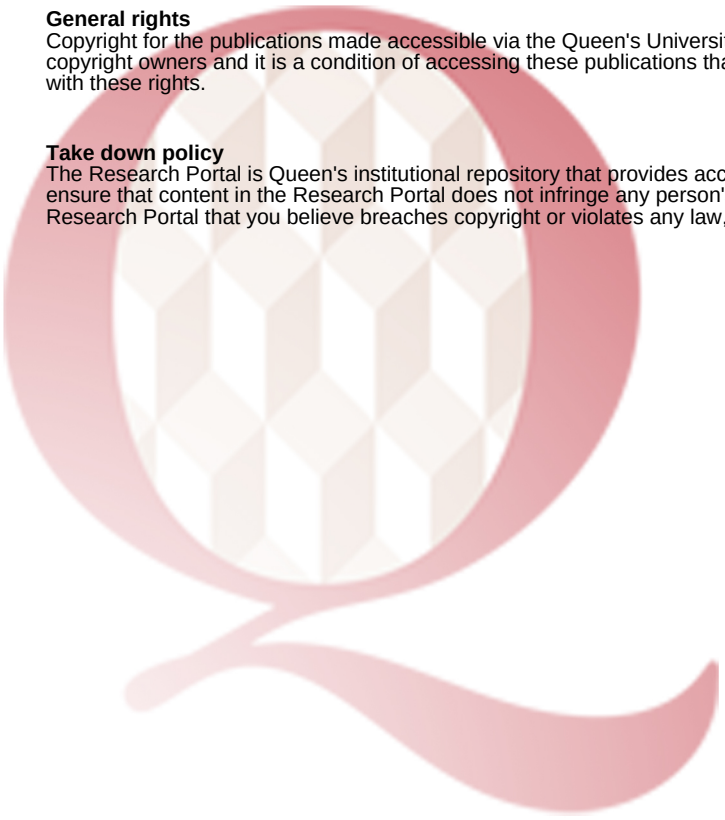
Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition

Ming, J., & Crookes, D. (2017). Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(3), 531-543. DOI: 10.1109/TASLP.2017.2651406

Published in:
IEEE/ACM Transactions on Audio, Speech, and Language Processing

Document Version:
Peer reviewed version



Speech Enhancement Based on Full-Sentence Correlation and Clean Speech Recognition

Ji Ming, *Member, IEEE*, Danny Crookes, *Senior Member, IEEE*

***Abstract*—Conventional speech enhancement methods, based on frame, multi-frame or segment estimation, require knowledge about the noise. This paper presents a new method which aims to reduce or effectively remove this requirement. It is shown that, by using the Zero-mean Normalized Correlation Coefficient (ZNCC) as the comparison measure, and by extending the effective length of speech segment matching to sentence-long speech utterances, it is possible to obtain an accurate speech estimate from noise without requiring specific knowledge about the noise. The new method, thus, could be used to deal with unpredictable noise or noise without proper training data. This paper is focused on realizing and evaluating this potential. We propose a novel realization that integrates full-sentence speech correlation with clean speech recognition, formulated as a constrained maximization problem, to overcome the data sparsity problem. Then we propose an efficient implementation algorithm to solve this constrained maximization problem, to produce speech sentence estimates. For evaluation, we build the new system on one training data set and test it on two different test data sets across two databases, for a range of different noises including highly nonstationary ones. It is shown that the new approach, without any estimation of the noise, is able to significantly outperform conventional methods which use optimized noise tracking, in terms of various objective measures including automatic speech recognition.**

***Index Terms*—Full-sentence correlation, recognizability constraint, wide matching, noise robustness, speech enhancement, speech recognition**

I. Introduction

In this research, we focus on the challenging problem of single-channel speech enhancement. We consider one of the worst-case scenarios: there is no specific knowledge about the noise, except its being additive and independent of the speech. To set our new approach in context, we group current techniques for single-channel speech enhancement into two main categories: *frame*-based methods and *segment*-based methods. Frame-based methods tend to treat each single speech frame, about 10-30 ms in length, one at a time. Examples include non-parametric spectral subtraction [1] and Wiener filtering [2], [3], parametric or statistical model based minimum mean-square error (MMSE) or maximum a posteriori (MAP) estimators (e.g., [4]–[6]), and data-driven (non-parametric or model-based) vector-quantization (VQ) codebook, Gaussian mixture model (GMM) and hidden Markov model (HMM) based estimators (e.g., [7]–[14]). However, because frames are so short, it can be difficult to distinguish speech from noise in them. Therefore, to recover speech, one must have knowledge about the noise, e.g., its power spectrum, probability distribution, or signal-to-noise ratio (SNR). When these noise statistics are not available, they are predicted by using neighbouring frames without significant speech event, based on voice activity detection, minimum statistics, time-recursive averaging, MMSE, or their combination (see, for example, [15]–[19]). Thus, these methods work for stationary or slowly-varying noise, but less well for fast-varying noise, because of its weak predictability.

Segment-based methods, on the other hand, aim to estimate a sequence of consecutive speech frames (called a segment), one segment at a time. Examples include inventory (or dictionary) based methods (e.g., [20]–[25]), longest matching segment (LMS) methods (e.g., [26]–[28] and some application examples [29]–[31]), and more recently, deep neural network (DNN) based methods (e.g., [32]–[40]). The inventory-based methods try to directly estimate some fixed-length speech segments or speech segments corresponding to phonemes. The LMS methods have gone one step further, by trying to identify the *longest* speech segments with matching training segments. These longest matching speech segments have variable lengths depending on the given training and test data, and were found, for example, to have an average length of 11-15 frames based on a number of speech databases (TIMIT, WSJ, Aurora 4) [26]–[28]. Most feedforward DNN systems (e.g., [32]–[38]) are trained to map fixed-length noisy speech segments, typically 9-15 frames long, to corresponding clean speech estimates. A system is described in [39] that combines several DNNs to model variable-length speech segments. Recurrent DNN systems (e.g., [41], [42]) may capture some longer speech context, depending on the available training data. Segments usually contain richer temporal dynamics than frames, exhibit greater distinction, and hence can better distinguish speech in noise.

It is no surprise that segment-based methods exhibit greater noise robustness than frame-based methods (see, for example, [21], [26], [37]). However, knowledge of noise remains essential for segment-based methods (an explanation will be provided below). For example, the state-of-the-art DNN systems, feedforward or recurrent structured, for speech enhancement [32]–[40], speech recognition [42]–[47] as well as for image denoising [48] all require proper training for noise, and a DNN system trained for one type of noise or distortion may not be applicable to significantly different types of noise or distortion. The ability to generalize to untrained noise conditions is one of the greatest challenges facing DNN studies. There are also other methods, for example, multi-frame methods [49], [50], which lie between the frame-based and the segment-based methods and try to capture the inter-

The authors are with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K. (e-mail: j.ming@qub.ac.uk, d.crookes@qub.ac.uk).

frame speech correlation for speech enhancement. It was reported that the multi-frame based Wiener or MVDR (minimum variance distortionless response) filtering improves the SNR or reduces the speech distortion over the corresponding independent-frame based filtering [49], [50].

In this paper, we present our work to extend the segment approach. We aim to identify much longer speech segments than can be modeled in current segment-based methods, effectively towards full-length speech sentence matching. We show that this could help further reduce or effectively remove the requirement for noise estimation or training. The new method, thus, could be used to deal with fast-varying or unpredictable noise for which we may not have a proper noise estimate or training data. Our studies are based on two novel approaches. First, we use a new measure for speech segment matching and thus extend the effective length of the matching to sentence-long speech utterances. This has the potential to maximize noise immunity without requiring specific information about the noise. Second, to realize this potential, we propose to integrate clean speech recognition into the enhancement process to regularize the formation of the potential matching estimates. To some degree, our new method emulates how humans sometimes separate speech in very noisy conditions - by trying to make semantic sense of the speech. Further, we propose an efficient implementation to put the new method into practical use. This paper is a substantial expansion of our preliminary study described in [51]. The expansions include an expanded theoretical study, an efficient implementation algorithm for speech enhancement, and extensive experimental investigations across two different databases. For convenience, we call the new method *wide matching*. This name particularly emphasizes our effort to try to directly match *sentence-wide* speech segments to improve noise robustness.

The remainder of the paper is organized as follows. Section II presents the new measure for speech segment matching and our hypothetical studies on the potential to maximize noise robustness without requiring noise information. Section III focuses on realizing this potential and details the wide matching approach. Section IV presents an efficient algorithm to implement the wide matching approach for speech enhancement. Experimental studies are described in Section V. Finally, conclusions are presented in Section VI.

II. Segment Correlation and Noise Robustness

Given a noisy speech signal, we aim to extract the underlying clean speech signal. We achieve this through identifying a matching clean speech signal from a clean speech corpus. We aim for high noise immunity. Thus, we need a method to compare noisy speech and clean speech that is immune to *any* background noise. Such a method may not exist for comparing frames or short segments of speech but fortunately, exist for comparing long segments of speech. At the heart of our studies is such a measure, *Zero-mean Normalized Correlation Coefficient* (ZNCC), which has a simple expression for comparing

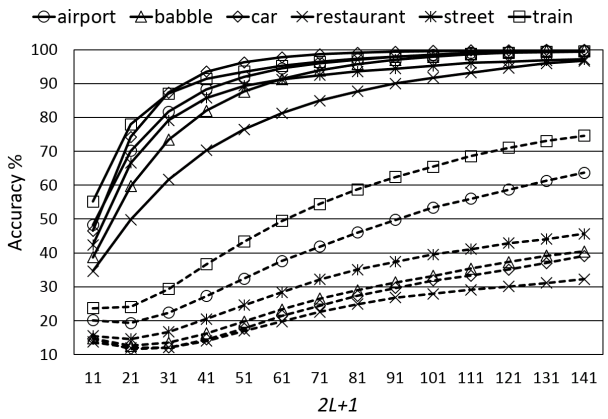


Fig. 1. An oracle experiment showing accuracy of finding best matching segments as a function of segment length L (in number of frames) without noise estimation, based on maximum ZNCC (solid lines) and minimum Euclidean distance (dashed lines), for six types of noise with SNR = -5 dB, for 57,919 noisy test speech segments for each type of noise and 1,124,863 clean training speech segments involving 486 speakers based on the TIMIT database.

noisy and clean speech segments:

$$R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L}) = \frac{\sum_{l=-L}^{L}[x_{t+l} - \mu(\mathbf{x}_{t\pm L})]^{\mathrm{T}}[s_{\tau+l} - \mu(\mathbf{s}_{\tau\pm L})]}{|\tilde{\mathbf{x}}_{t\pm L}||\tilde{\mathbf{s}}_{\tau\pm L}|} \quad (1)$$

where $\mathbf{x}_{t\pm L}$ represents a noisy speech segment centered at frame x_t and consisting of $2L+1$ consecutive frames from x_{t-L} to x_{t+L}; $\mathbf{s}_{\tau\pm L}$ represents a clean speech segment taken from a corpus speech sentence centered at some frame s_τ with $2L+1$ consecutive frames from $s_{\tau-L}$ to $s_{\tau+L}$. In our studies, we assume that the individual speech frames are represented by their corresponding short-time power spectra, and so x_t and s_τ are vectors of short-time power spectral coefficients, and T means vector transpose. In (1), $\mu(\mathbf{x}_{t\pm L})$ stands for the mean frame vector of segment $\mathbf{x}_{t\pm L}$, i.e., $\mu(\mathbf{x}_{t\pm L}) = \sum_{l=-L}^{L} x_{t+l}/(2L+1)$, and $|\tilde{\mathbf{x}}_{t\pm L}|$ stands for the zero-mean Euclidean norm of segment $\mathbf{x}_{t\pm L}$, i.e., $|\tilde{\mathbf{x}}_{t\pm L}|^2 = \sum_{l=-L}^{L}[x_{t+l} - \mu(\mathbf{x}_{t\pm L})]^{\mathrm{T}}[x_{t+l} - \mu(\mathbf{x}_{t\pm L})]$. The same definitions apply to the clean corpus speech segment $\mathbf{s}_{\tau\pm L}$, with mean frame vector $\mu(\mathbf{s}_{\tau\pm L})$ and zero-mean Euclidean norm $|\tilde{\mathbf{s}}_{\tau\pm L}|$. When $\mathbf{x}_{t\pm L}$ and $\mathbf{s}_{\tau\pm L}$ are short (L is small), $R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L})$ may or may not offer much advantage over other types of distances or likelihoods that have been used for speech comparison. However, when $\mathbf{x}_{t\pm L}$ and $\mathbf{s}_{\tau\pm L}$ are very long (L is very large), its advantage should become overwhelming – it should in theory become immune to *any* independent additive noise. In the following, we first use an oracle experiment to show the significance. Then we provide a theoretical explanation of the experimental results. The experiment was conducted on the TIMIT database with 3696 clean training sentences and 192 clean core test sentences. We took the core test sentences and added different types of noise for testing. Given a noisy speech segment of length $2L+1$, we seek to find a matching (clean) speech segment estimate from the corpus, which includes the training data and the clean version of the test segment as the best candidate (hence the 'oracle'). With different noises (airport, babble,

car, restaurant, street, and train station) with SNR = -5dB, we measured the retrieval accuracy using standard minimum Euclidean distance (equivalent to maximum Gaussian-based likelihood without noise information) and then using the new maximum ZNCC, for segment lengths $(2L+1)$ up to 141 frames ($\sim$1.4 s of speech). The results in Fig. 1 indicate that, as the segment length increases, the best matching estimates using maximum ZNCC were found with a rapidly increasing probability, approaching 100%, regardless of the noise. However, the same experiment using minimum distances failed to show this property.

The above oracle experimental results may be accounted for by a theory, described below. Consider additive noise and frames being represented in short-time power spectra. So each noisy speech frame can be approximately expressed as $x_t = s'_t + n_t$, where s'_t represents the underlying clean speech frame and n_t represents the noise frame. Thus, we can decompose the ZNCC $R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L})$, defined in (1), into two terms

$$\begin{aligned} R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L}) &= R(\mathbf{s}'_{t\pm L} + \mathbf{n}_{t\pm L}, \mathbf{s}_{\tau\pm L}) \\ &= \frac{\sum_{l=-L}^{L}[s'_{t+l} - \mu(\mathbf{s}'_{t\pm L})]^{\mathrm{T}}[s_{\tau+l} - \mu(\mathbf{s}_{\tau\pm L})]}{|\tilde{\mathbf{x}}_{t\pm L}||\tilde{\mathbf{s}}_{\tau\pm L}|} \\ &\quad + \frac{\sum_{l=-L}^{L}[n_{t+l} - \mu(\mathbf{n}_{t\pm L})]^{\mathrm{T}}[s_{\tau+l} - \mu(\mathbf{s}_{\tau\pm L})]}{|\tilde{\mathbf{x}}_{t\pm L}||\tilde{\mathbf{s}}_{\tau\pm L}|} \\ &= \frac{|\tilde{\mathbf{s}}'_{t\pm L}|}{|\tilde{\mathbf{x}}_{t\pm L}|} R(\mathbf{s}'_{t\pm L}, \mathbf{s}_{\tau\pm L}) + \frac{|\tilde{\mathbf{n}}_{t\pm L}|}{|\tilde{\mathbf{x}}_{t\pm L}|} R(\mathbf{n}_{t\pm L}, \mathbf{s}_{\tau\pm L}) \end{aligned} \quad (2)$$

where $\mathbf{s}'_{t\pm L}$ represents the underlying clean speech segment in the noisy segment $\mathbf{x}_{t\pm L}$ from frame s'_{t-L} to s'_{t+L}, and $\mathbf{n}_{t\pm L}$ represents the corresponding noise segment from frame n_{t-L} to n_{t+L}, with $\mu(\mathbf{s}'_{t\pm L})$, $\mu(\mathbf{n}_{t\pm L})$ and $|\tilde{\mathbf{s}}'_{t\pm L}|$, $|\tilde{\mathbf{n}}_{t\pm L}|$ representing the mean frame vector and zero-mean Euclidean norm of segment $\mathbf{s}'_{t\pm L}$ and $\mathbf{n}_{t\pm L}$, respectively. In (2), the first term is the ZNCC between the underlying speech segment $\mathbf{s}'_{t\pm L}$ and the corpus speech segment $\mathbf{s}_{\tau\pm L}$, weighted by $|\tilde{\mathbf{s}}'_{t\pm L}|/|\tilde{\mathbf{x}}_{t\pm L}|$ which is a constant for all the corpus segments, subject only to the SNR in the observation. The second term is the ZNCC between the noise segment and the corpus speech segment, weighted by $|\tilde{\mathbf{n}}_{t\pm L}|/|\tilde{\mathbf{x}}_{t\pm L}|$, which is again independent of the corpus segment, subject only to the SNR in the observation. For large L and noise independent of the corpus speech, it follows that the second term in (2) tends to zero:

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

where $\mu(n_t)$ and $\mu(s_\tau)$ represent the mean frame vector of the noise and speech processes, respectively. Eq. (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 [52]), and (4) [from (3)] is based on the assumption that the corpus speech and noise are statistically independent. Thus, with (2)–(4), for large L and independent noise we may have

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

That is, the maximum ZNCC (i.e., the matching accuracy) could become independent of the noise but depends only on the two speech segments being compared, one being underlying speech and the other a potential matching clean speech estimate. This hypothesis is in good agreement with the experimental results shown in Fig. 1.

However, if Euclidean distance is used for comparison, then lengthening the speech segments for comparison may not necessarily lead to comparable noise robustness. To link these two methods, consider the Euclidean distance between two zero-mean speech segments (i.e., the mean frame vector is subtracted from all frames in the segment): the noisy speech segment $\tilde{\mathbf{x}}_{t\pm L}$ (where $\tilde{}$ indicates mean removed), and the corpus speech segment $\tilde{\mathbf{s}}_{\tau\pm L}$, with the underlying speech segment represented by $\tilde{\mathbf{s}}'_{t\pm L}$. The Euclidean distance $|\tilde{\mathbf{x}}_{t\pm L} - \tilde{\mathbf{s}}_{\tau\pm L}|$ can be written as

$$\begin{aligned} &|\tilde{\mathbf{x}}_{t\pm L} - \tilde{\mathbf{s}}_{\tau\pm L}|^2 \\ &= |\tilde{\mathbf{x}}_{t\pm L}|^2 + |\tilde{\mathbf{s}}_{\tau\pm L}|^2 - 2\tilde{\mathbf{x}}_{t\pm L}^{\mathrm{T}}\tilde{\mathbf{s}}_{\tau\pm L} \\ &= |\tilde{\mathbf{x}}_{t\pm L}|^2 + |\tilde{\mathbf{s}}_{\tau\pm L}|^2 - 2|\tilde{\mathbf{x}}_{t\pm L}||\tilde{\mathbf{s}}_{\tau\pm L}| \frac{\tilde{\mathbf{x}}_{t\pm L}^{\mathrm{T}}\tilde{\mathbf{s}}_{\tau\pm L}}{|\tilde{\mathbf{x}}_{t\pm L}||\tilde{\mathbf{s}}_{\tau\pm L}|} \\ &= |\tilde{\mathbf{x}}_{t\pm L}|^2 + |\tilde{\mathbf{s}}_{\tau\pm L}|^2 - 2|\tilde{\mathbf{x}}_{t\pm L}||\tilde{\mathbf{s}}_{\tau\pm L}| R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L}) \end{aligned} \quad (6)$$

Assume large L and independent speech and noise [so $R(\mathbf{x}_{t\pm L}, \mathbf{s}_{\tau\pm L}) \approx |\tilde{\mathbf{s}}'_{t\pm L}|/|\tilde{\mathbf{x}}_{t\pm L}| R(\mathbf{s}'_{t\pm L}, \mathbf{s}_{\tau\pm L})$ based on (2) and (4)], and further assume perfectly correlated corpus speech segment $\mathbf{s}_{\tau\pm L}$ and underlying speech segment $\mathbf{s}'_{t\pm L}$ [so $R(\mathbf{s}'_{t\pm L}, \mathbf{s}_{\tau\pm L}) \approx 1$]. We will have

$$\begin{aligned} &|\tilde{\mathbf{x}}_{t\pm L} - \tilde{\mathbf{s}}_{\tau\pm L}|^2 \\ &\quad \approx |\tilde{\mathbf{x}}_{t\pm L}|^2 + |\tilde{\mathbf{s}}_{\tau\pm L}|^2 - 2|\tilde{\mathbf{s}}'_{t\pm L}||\tilde{\mathbf{s}}_{\tau\pm L}| \\ &\quad = (|\tilde{\mathbf{s}}_{\tau\pm L}| - |\tilde{\mathbf{s}}'_{t\pm L}|)^2 + |\tilde{\mathbf{x}}_{t\pm L}|^2 - |\tilde{\mathbf{s}}'_{t\pm L}|^2 \end{aligned} \quad (7)$$

As indicated in (7), even having all the assumptions made for ZNCC, and having the presumable noise robustness of ZNCC, there still remains uncertainty in the Euclidean distance for identifying the best matching corpus estimate. In (7), the first difference, $|\tilde{\mathbf{s}}_{\tau\pm L}| - |\tilde{\mathbf{s}}'_{t\pm L}|$, exists due to different gains between the corpus and underlying speech; the second difference, $|\tilde{\mathbf{x}}_{t\pm L}|^2 - |\tilde{\mathbf{s}}'_{t\pm L}|^2$, is proportional to the noise power in the observation. In the oracle experiment (Fig. 1), the first difference is zero because of the perfectly matching estimate. But the second difference caused by noise will not necessarily decrease by extending the length of speech segment matching. This can explain why using minimum Euclidean distance was much less robust than using maximum ZNCC in the oracle experiment, for finding perfectly matching long speech segments in noise.

The above oracle experiment, and theory, have motivated this research. If we can calculate the correlation between two very long speech segments (one being noisy speech and the other a potential matching clean speech estimate), then we should be able to obtain an accurate speech estimate without requiring specific knowledge about the noise. This method can thus be used to deal with untrained noise, or noise difficult

to train or estimate. Of course, the oracle experiment has one major obstacle to its practical implementation – it is almost impossible to collect enough training data to include the matching estimates for all possible very long speech segments, e.g., full speech sentences, which are typically hundreds to thousands of frames long. Current segment-based speech enhancement methods (e.g., inventory, LMS and feedforward DNN) can only model speech segment estimates typically below 20 frames – the size of a large phoneme or syllable. As shown in Fig 1, this just is not long enough to identify the best matching speech segments, which explains why these methods require noise estimation or noise training.[1] Significantly lengthening the speech segments for these methods does not seem to be feasible due to the sparsity of training data. In the next section, we describe our approach to try to break the segment length barrier. With limited training/corpus data, the new approach, namely wide matching, tries to implement *full-sentence* correlation for speech estimation, i.e., to go as far to the right in Fig. 1 as possible, to maximize immunity to noise.

III. Wide Matching for Speech Enhancement

As described above, we face the problem of generating matching estimates for full speech sentences. These estimates must be generated from limited training data, and must ideally cover *all* speech sentences, and *only* speech sentences. We formulate this as a constrained optimization problem. Given a noisy, unseen speech sentence, we seek the best full-sentence chain of clean corpus segments as an estimate. The optimal estimate has maximum sentence-level ZNCC with the noisy speech, subject to maximum recognizability to be valid speech.

A. Full Sentence Correlation

Our approach is essentially to replace a very long speech segment of contiguous frames with a chain of short speech segments. Let $X = (x_1, x_2, ..., x_T)$ be a noisy speech sentence with T frames in short-time power spectral vectors, and with an underlying speech sentence which is unseen in the corpus. To derive the enhancement algorithm, we use a segment-chain expression for X, and for the underlying speech sentence. Suppose we can divide X into some K consecutive segments, denoted by a segment chain $\mathbf{X} = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, ..., \mathbf{x}_{t_K})$, where each element segment $\mathbf{x}_{t_k}$ is centered at some frame time t_k with frames from $x_{t_k-\gamma}$ to $x_{t_k+\gamma}$, and the length of the element segment is $2\gamma + 1$. For simplicity, we assume a common γ being used for all the element segments and so γ can be implied in the expression. For example, a possible segmentation is $\mathbf{X} = X$, i.e., X is evenly divided into K consecutive segments and there is no overlap between adjacent segments [assuming $(2\gamma + 1)K = T$]. But in practice, to improve the segment smoothness, adjacent segments $\mathbf{x}_{t_k}$ are normally partially overlapped. In the following, unless otherwise indicated, we will use $\mathbf{X}$ to represent a noisy sentence and use a corresponding chain of clean speech segments to represent the underlying speech sentence to be estimated. Assume that the element segments are short (γ is small), such that we can find from the corpus a corresponding full-sentence chain of clean corpus segments as an estimate of the underlying speech sentence in $\mathbf{X}$. Denote by $\mathbf{S} = (g_{\tau_1}\mathbf{s}_{\tau_1}, g_{\tau_2}\mathbf{s}_{\tau_2}, ..., g_{\tau_K}\mathbf{s}_{\tau_K})$ such a chain of K clean corpus segments, where each element corpus 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 corpus segment in forming the sentence estimate. In $\mathbf{S}$, different corpus segments $\mathbf{s}_{\tau_k}$ can come from different corpus sentences/contexts to simulate arbitrary unseen test speech. Given the noisy sentence $\mathbf{X}$, we obtain the optimal clean speech sentence estimate $\mathbf{S}$ based on the *full-sentence* ZNCC $R(\mathbf{X}, \mathbf{S})$. This can be written as

$$\begin{aligned} R(\mathbf{X}, \mathbf{S}) &= R(\mathbf{x}_{t_1}\mathbf{x}_{t_2}...\mathbf{x}_{t_K}, g_{\tau_1}\mathbf{s}_{\tau_1}g_{\tau_2}\mathbf{s}_{\tau_2}...g_{\tau_K}\mathbf{s}_{\tau_K}) \\ &= \frac{\sum_{k=1}^{K}[\mathbf{x}_{t_k} - \mu(\mathbf{X})]^{\mathrm{T}}[g_{\tau_k}\mathbf{s}_{\tau_k} - \mu(\mathbf{S})]}{|\tilde{\mathbf{X}}||\tilde{\mathbf{S}}|} \\ &= \frac{\sum_{k=1}^{K}\sum_{l=-\gamma}^{\gamma}[x_{t_k+l} - \mu(\mathbf{X})]^{\mathrm{T}}[g_{\tau_k}s_{\tau_k+l} - \mu(\mathbf{S})]}{|\tilde{\mathbf{X}}||\tilde{\mathbf{S}}|} \\ &= \frac{\sum_{k=1}^{K} g_{\tau_k}\sum_{l=-\gamma}^{\gamma} x_{t_k+l}^{\mathrm{T}}s_{\tau_k+l} - L\mu(\mathbf{X})^{\mathrm{T}}\mu(\mathbf{S})}{|\tilde{\mathbf{X}}||\tilde{\mathbf{S}}|} \end{aligned} \tag{8}$$

where $L = (2\gamma+1)K$ is the length (number of frames) of the two full sentences $\mathbf{X}$ and $\mathbf{S}$ being correlated; $\mu(\mathbf{S})$ and $|\tilde{\mathbf{S}}|$ are the global mean frame vector and zero-mean Euclidean norm of the corpus segment chain based speech sentence estimate $\mathbf{S}$, respectively, i.e.,

$$\mu(\mathbf{S}) = \frac{1}{L}\sum_{k=1}^{K} g_{\tau_k}\sum_{l=-\gamma}^{\gamma} s_{\tau_k+l} \tag{9}$$

$$|\tilde{\mathbf{S}}|^2 = \sum_{k=1}^{K} g_{\tau_k}^2\sum_{l=-\gamma}^{\gamma} s_{\tau_k+l}^{\mathrm{T}}s_{\tau_k+l} - L\mu(\mathbf{S})^{\mathrm{T}}\mu(\mathbf{S}) \tag{10}$$

The above expressions (9) and (10) apply to $\mu(\mathbf{X})$ and $|\tilde{\mathbf{X}}|$, the global mean frame vector and zero-mean Euclidean norm of the noisy sentence $\mathbf{X}$ (without the gain terms). It should be noted that the above representation of unseen speech sentences in chains or sequences of short training speech segments is common in the segment-based speech enhancement methods (e.g., inventory, LMS and DNN). One major difference between these previous methods and the proposed new method is that the previous methods tend to estimate each (short) speech segment independently of the other segments in the sentence. Because of this, they have lacked the ability to capture the longer, cross-segment correlation of speech for speech separation. But in the full-sentence ZNCC $R(\mathbf{X}, \mathbf{S})$ defined above, there is no assumption about the temporal or spectral independence of speech within the segments, across the segments or anywhere in the sentence.

B. Incorporating Clean Speech Recognition

Because we have no knowledge about the underlying speech, in theory we may have to consider all possible chains

[1] It is noted that in neural networks similar correlation operations are performed for the input data (i.e., segments), for example, zero-mean normalization, and inner products between the normalized input data and weights [53].

of corpus segments $\mathbf{S}$ to search for the potential matching estimate. However, not all of the possible chains constitute realistic speech; some chains may well be semantically meaningless sequences of segments, even if they have a larger ZNCC. For example, there could be chains that match the noise better [i.e., they maximize the second term in (2), thinking of the appropriate segments as segment chains]. When noise dominates the observed data, this could lead to an estimation error. So to constrain the estimate to valid speech, we propose to use the estimate's *recognizability*, by using a *clean speech recognizer*, to regularize the formation of the potential matching estimate. Thus, we can formulate the problem of obtaining an optimal speech sentence estimate as constrained maximization of the full-sentence ZNCC subject to *maximum recognizability* of the estimate. We use the expression

$$\hat{\mathbf{S}} = \arg\max_{\mathbf{S}} [RH(\mathbf{X}, \mathbf{S}) = \log R(\mathbf{X}, \mathbf{S}) + \lambda \log H(\mathbf{S})] \quad (11)$$

where $\hat{\mathbf{S}}$ represents the optimal corpus segment chain based speech sentence estimate, $H(\mathbf{S})$ is a clean speech recognizer's likelihood of the corpus segment chain $\mathbf{S}$ to be valid speech, λ is a Lagrange multiplier, and $RH(\mathbf{X}, \mathbf{S})$ is an abbreviation for the recognizability-constrained full-sentence ZNCC. As an example, in this paper we implemented a simple HMM-based clean speech recognizer which produces a log likelihood score for a given $\mathbf{S}$ as follows:

$$\log H(\mathbf{S}) = [\log h(\mathbf{S}) + \sum_{i=1}^{I} \log p_i(d_i) + \sum_{u=1}^{U} \log p_u(d_u)]/T \quad (12)$$

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 (see Sections IV and V for more details about how to implement this recognizer). Because the recognizer is trained with clean speech, we can assume that among all the possible corpus segment chains, the chains resembling clean, valid speech are most recognizable to the recognizer in terms of achieving large likelihoods $H(\mathbf{S})$ (this is because clean, valid speech is most likely to simultaneously fulfill the acoustic, language, state duration and phone duration constraints of clean speech learned by the recognizer). If a sentence-long chain with a large noise-independent likelihood of being clean, valid speech *simultaneously* has a large ZNCC with the noisy sentence, or vice versa, then we can assume that it is an optimal estimate of the underlying speech sentence. We assume that (11) partially emulates what humans sometimes do in trying to pick out speech in strong noise – humans try to make sense of the speech, by recognizing parts or even the whole of the sentence, as part of our method of noise removal. Specifically, we call (11) the wide matching approach.

C. An Iterative Solution

We propose a computationally efficient iterative algorithm to solve the constrained maximization problem (11). Given a noisy sentence $\mathbf{X} = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, ..., \mathbf{x}_{t_K})$, we seek an optimal full-sentence chain of corpus speech segments $\hat{\mathbf{S}} = (\hat{g}_{\tau_1}\hat{\mathbf{s}}_{\tau_1}, \hat{g}_{\tau_2}\hat{\mathbf{s}}_{\tau_2}, ..., \hat{g}_{\tau_K}\hat{\mathbf{s}}_{\tau_K})$ that maximizes the recognizability-constrained full-sentence ZNCC $RH(\mathbf{X}, \hat{\mathbf{S}})$. We start with an initial estimate $\hat{\mathbf{S}}$ by separately estimating each element corpus segment $\hat{\mathbf{s}}_{\tau_k}$ through maximizing the segment-level ZNCC $R(\mathbf{x}_{t_k}, \mathbf{s}_{\tau_k})$ based on (1), assuming a unit gain $\hat{g}_{\tau_k}$. Then we update this initial estimate by alternately re-estimating each element corpus segment with gain to maximize the appropriate $RH(\mathbf{X}, \hat{\mathbf{S}})$; in re-estimating a specific element corpus segment, the other element corpus 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 corpus segments $\hat{g}_{\tau_k}\hat{\mathbf{s}}_{\tau_k}$ in the order from $k = 1$ to K. In the jth iteration, to obtain a new estimate of the optimal kth element corpus segment, denoted by $\hat{g}^j_{\tau_k}\hat{\mathbf{s}}^j_{\tau_k}$, we maximize the recognizability-constrained full-sentence ZNCC with respect to $g_{\tau_k}\mathbf{s}_{\tau_k}$, with the succeeding element corpus segments $\hat{g}_{\tau_m}\hat{\mathbf{s}}_{\tau_m}$ $(m > k)$ taken from the $(j-1)$th iteration, and the preceding element corpus segments $\hat{g}_{\tau_m}\hat{\mathbf{s}}_{\tau_m}$ $(m < k)$ taken from the jth iteration. Therefore in the jth iteration with the kth element corpus segment to be re-estimated, the optimal speech sentence estimate can be written 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 corpus segments fixed to their latest optimal estimates from the appropriate iterations. A new estimate of $\hat{g}_{\tau_k}\hat{\mathbf{s}}_{\tau_k}$, and a corresponding new speech sentence estimate, are obtained by maximizing $RH[\mathbf{X}, \hat{\mathbf{S}}^j(g_{\tau_k}\mathbf{s}_{\tau_k})]$ with respect to $g_{\tau_k}\mathbf{s}_{\tau_k}$, i.e.,

$$\begin{aligned}\hat{\mathbf{S}}^j(\hat{g}^j_{\tau_k}\hat{\mathbf{s}}^j_{\tau_k}) \\ = \arg\max_{g_{\tau_k}\mathbf{s}_{\tau_k}} \{RH[\mathbf{X}, \hat{\mathbf{S}}^j(g_{\tau_k}\mathbf{s}_{\tau_k})] = \log R[\mathbf{X}, \hat{\mathbf{S}}^j(g_{\tau_k}\mathbf{s}_{\tau_k})] \\ + \lambda \log H[\hat{\mathbf{S}}^j(g_{\tau_k}\mathbf{s}_{\tau_k})]\} \\ k = 1, 2, ..., K; j = 1, 2, ...\end{aligned} \quad (13)$$

with $\hat{g}^0_{\tau_k}\hat{\mathbf{s}}^0_{\tau_k}$ corresponding to the initial estimates. Eq. (13) represents an iterative algorithm to solve the constrained maximization problem (11). It manages to estimate the optimal element corpus segments one segment at a time, subject to the constraints of all the other segments in the sentence, and hence can be calculated efficiently. Let $\hat{\mathbf{S}}^j$ denote the speech sentence estimate at the end of the jth iteration after all the element corpus segments have been updated. In our experiments, we have seen that this algorithm converges in terms of generating speech sentence estimates that always increase $RH(\mathbf{X}, \hat{\mathbf{S}}^j)$ with each iteration. A more detailed step-by-step implementation of the algorithm is presented in the next section along with an algorithm for integrating the clean speech recognizer.

IV. An Example Implementation

A. Integrating Clean Speech Recognition

Eq. (11) shows that, with full-sentence correlation, we may effectively reduce the problem of noisy speech enhancement to a problem of *clean speech recognition*, and thus remove the requirement for noise estimation or noise training.

We could integrate, for example, DNN-based clean speech recognition (e.g., [41]) into the system to provide advanced speech recognition accuracy, and hence achieve both robust and accurate speech estimation from noise. In this paper, however, we describe an alternative method of embedding clean speech recognition into the system. This method may not necessarily provide optimal clean speech recognition but it should require less computation, and hence is important in terms of implementation and applications. We use the clean speech recognizer that is used in our experiments to describe this method.

The recognizer is effectively an HMM-based phone recognizer, in which three states, each associated with a frame-emitting GMM, are used to model the acoustics [e.g., the mel-frequency cepstral coefficients (MFCCs) and derivatives sequences] of a phone. We train the recognizer using the corpus data. After training, we force align each corpus sentence to the corresponding phonetic HMMs. Thus, for each corpus segment $\mathbf{s}_{\tau_k} = (s_{\tau_k-\gamma}, ..., s_{\tau_k}, ..., s_{\tau_k+\gamma})$ we can obtain a corresponding sequence of triplets

$$\begin{aligned}\mathbf{v}_{\tau_k} = \{&[Q(s_{\tau_k-\gamma}), \log b_Q(s_{\tau_k-\gamma}), u_Q(s_{\tau_k-\gamma})],\\ &...,\\ &[Q(s_{\tau_k}), \log b_Q(s_{\tau_k}), u_Q(s_{\tau_k})],\\ &...,\\ &[Q(s_{\tau_k+\gamma}), \log b_Q(s_{\tau_k+\gamma}), u_Q(s_{\tau_k+\gamma})]\}\end{aligned} \tag{14}$$

in which each triplet $[Q(s_\tau), \log b_Q(s_\tau), u_Q(s_\tau)]$ records the most-likely state Q, the state-based log likelihood $\log b_Q$ and the state-based phonetic label u_Q of a corresponding corpus frame s_τ, based on the maximum-likelihood frame-to-state alignment. In the following, we show that we can use the pre-recorded triplet sequences of the corpus segments to approximately express the result of the Viterbi search of a given speech sentence estimate, and thus to accomplish the required clean speech recognition of the estimate with minimal calculation.

Given a corpus segment chain based speech sentence estimate $\mathbf{S} = (g_{\tau_1}\mathbf{s}_{\tau_1}, g_{\tau_2}\mathbf{s}_{\tau_2}, ..., g_{\tau_K}\mathbf{s}_{\tau_K})$, we will calculate the log likelihood $\log H(\mathbf{S})$ to decide if $\mathbf{S}$ is valid speech. As shown in (12), this includes calculating the log likelihood $\log h(\mathbf{S})$ associated with the clean speech acoustic and language models, and measuring the durations of all the states and phones in $\mathbf{S}$ to obtain the probabilities based on clean speech statistics. We assume that the latter should be straightforward. Without loss of generality, assuming that a speech sentence always begins with silence, we can write $\log h(\mathbf{S})$ as

$$\begin{aligned}&\log h(\mathbf{S})\\ &= \max_{\mathbf{q}} \sum_{k=1}^{K} \sum_{\tau=\tau_k-\gamma'}^{\tau_k+\gamma'} \{\log a[q(s_{\tau-1}), q(s_\tau)] + \log b_{q(s_\tau)}(s_\tau)\}\\ &\approx \sum_{k=1}^{K} \sum_{\tau=\tau_k-\gamma'}^{\tau_k+\gamma'} \{\log a[Q(s_{\tau-1}), Q(s_\tau)] + \log b_{Q(s_\tau)}(s_\tau)\}\end{aligned} \tag{15}$$

where $\gamma' \leq \gamma$ represents the net length of each corpus segment $\mathbf{s}_{\tau_k}$ in forming the (non-overlapping) frame sequence of the speech sentence corresponding to $\mathbf{S}$, $\mathbf{q} = [q(s_{\tau_1-\gamma'}), q(s_{\tau_1-\gamma'+1}), ..., q(s_{\tau_K+\gamma'})]$ represents a possible state sequence of the frame sequence of the speech sentence, $a(i,j)$ are the state transition probabilities, and, by definition, $q(s_{\tau_k-\gamma'-1}) = q(s_{\tau_{k-1}+\gamma'})$ for $k > 1$ and $q(s_{\tau_1-\gamma'-1}) = q_0$, where q_0 represents the initial state of the initial (silence) acoustic model. As shown in (15), for any given corpus segment chain $\mathbf{S} = (g_{\tau_1}\mathbf{s}_{\tau_1}, g_{\tau_2}\mathbf{s}_{\tau_2}, ..., g_{\tau_K}\mathbf{s}_{\tau_K})$, we use its corresponding chain of pre-stored triplet sequences $(\mathbf{v}_{\tau_1}, \mathbf{v}_{\tau_2}, ..., \mathbf{v}_{\tau_K})$ to obtain an approximate most-likely state sequence $[Q(s_{\tau_1-\gamma'}), Q(s_{\tau_1-\gamma'+1}), ..., Q(s_{\tau_K+\gamma'})]$ and further to compose the corresponding likelihood, as an approximation of the Viterbi search. This approximation reduces the calculation for the recognition of $\mathbf{S}$ to just $O(T)$ look-up table operations and additions, where T is the number of frames in the sentence. In (15), the state transition probabilities $a(i,j)$ typically encode both acoustic constraints (e.g., left-to-right topology) and language constraints (e.g., bigram phonetic language model) on the state transition of valid speech within and across phonetic units. In other words, only the chains $\mathbf{S}$ with a state sequence that fulfills the appropriate acoustic and language constraints of speech can score highly.

B. The Step-by-Step Algorithm

The following summarizes the overall wide matching algorithm used in our experiments, for searching for the optimal corpus segment chains for speech sentence estimation.

1) *Initialization*. Given a noisy sentence $\mathbf{X} = (\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, ..., \mathbf{x}_{t_K})$, find an initial corpus segment chain $\hat{\mathbf{S}}^0 = (\hat{g}^0_{\tau_1}\hat{\mathbf{s}}^0_{\tau_1}, \hat{g}^0_{\tau_2}\hat{\mathbf{s}}^0_{\tau_2}, ..., \hat{g}^0_{\tau_K}\hat{\mathbf{s}}^0_{\tau_K})$ in which each corpus segment $\hat{\mathbf{s}}^0_{\tau_k}$ is obtained by maximizing the segment-level ZNCC $R(\mathbf{x}_{t_k}, \mathbf{s}_{\tau_k})$ over all possible corpus segment candidates $\mathbf{s}_{\tau_k}$, assuming $\hat{g}^0_{\tau_k} = 1$.

2) *Iterative re-estimation*. For each iteration j ($j >= 1$), for each initial optimal corpus segment $\hat{g}^{j-1}_{\tau_k}\hat{\mathbf{s}}^{j-1}_{\tau_k}$ ($k = 1, 2, ..., K$) to be re-estimated, for each possible corpus segment candidate $\mathbf{s}_{\tau_k}$ do:

- Search for an optimal segment gain $\hat{g}_{\tau_k}$ for $\mathbf{s}_{\tau_k}$ to maximize the full-sentence ZNCC $R[\mathbf{X}, \hat{\mathbf{S}}^j(\hat{g}_{\tau_k}\mathbf{s}_{\tau_k})]$ based on (8), where $\hat{\mathbf{S}}^j(\hat{g}_{\tau_k}\mathbf{s}_{\tau_k}) = (\hat{g}^j_{\tau_1}\hat{\mathbf{s}}^j_{\tau_1}, ..., \hat{g}^j_{\tau_{k-1}}\hat{\mathbf{s}}^j_{\tau_{k-1}}, \hat{g}_{\tau_k}\mathbf{s}_{\tau_k}, \hat{g}^{j-1}_{\tau_{k+1}}\hat{\mathbf{s}}^{j-1}_{\tau_{k+1}}, ..., \hat{g}^{j-1}_{\tau_K}\hat{\mathbf{s}}^{j-1}_{\tau_K})$.
- Calculate the likelihood $\log H[\hat{\mathbf{S}}^j(\hat{g}_{\tau_k}\mathbf{s}_{\tau_k})]$ based on (12) and (15).
- Combine $R[\mathbf{X}, \hat{\mathbf{S}}^j(\hat{g}_{\tau_k}\mathbf{s}_{\tau_k})]$ and $\log H[\hat{\mathbf{S}}^j(\hat{g}_{\tau_k}\mathbf{s}_{\tau_k})]$ to obtain the recognizability-constrained full-sentence ZNCC score $RH[\mathbf{X}, \hat{\mathbf{S}}^j(\hat{g}_{\tau_k}\mathbf{s}_{\tau_k})]$, and take the $\hat{g}_{\tau_k}\mathbf{s}_{\tau_k}$ that has the maximum score among all the candidates as the new optimal corpus segment $\hat{g}^j_{\tau_k}\hat{\mathbf{s}}^j_{\tau_k}$, as indicated in (13).

In our experiments, we stop iterating when there is no change in the estimate $\hat{\mathbf{S}}^j$ between successive iterations. The above Step 2), iterative re-estimation, takes most of the computational time. In our experiments, we accelerate the computation by only considering in Step 2) the most-likely corpus segment candidates $\mathbf{s}_{\tau_k}$ for each noisy element segment $\mathbf{x}_{t_k}$. The most-likely corpus segment candidates for each noisy segment are selected in Step 1) which satisfy $R(\mathbf{x}_{t_k}, \mathbf{s}_{\tau_k}) >= R_{min}$, where R_{min} is a threshold used to prune unlikely matching

corpus segments because of their extremely low correlation values. In our experiments, we choose $R_{min} = 0.1$. The algorithm is found to be faster than our previous iterative LMS algorithm [28].

C. Reconstructing Speech Based on Corpus Segment Estimates

Since the corpus speech and test speech may differ in speaker characteristics, and since there may be phone or word recognition errors, we do not directly output the matching corpus segment chain, but use it to reconstruct the underlying speech through optimal noise filtering. This approach is similar to what was used previously in our LMS-based methods for speech enhancement and speech separation (e.g., [26], [27]), with the aim of better retaining the speech's intelligibility and the speakers' characteristics while reducing the noise. Let $X = (x_1, x_2, ..., x_T)$ be the frame sequence of a noisy sentence, let $S' = (s'_1, s'_2, ..., s'_T)$ be the frame sequence of the underlying speech sentence, and let $\hat{\mathbf{S}} = (\hat{g}_{\tau_1}\hat{\mathbf{s}}_{\tau_1}, \hat{g}_{\tau_2}\hat{\mathbf{s}}_{\tau_2}, ..., \hat{g}_{\tau_K}\hat{\mathbf{s}}_{\tau_K})$ be a matching corpus segment chain. Due to the overlap between adjacent speech segments, each underlying speech frame can be included in a number of matching corpus segments, or in other words, multiple overlapping matching corpus segments can each contain an estimate of the same speech frame. We can take an average between these individual segment-based estimates to form an overall estimate of the corresponding underlying speech frame. We use the underlying speech frame s'_t as an example. Let $P(s'_t)$ denote the short-time DFT (discrete Fourier transform) power spectrum of s'_t to be sought (while it is assumed that s'_t is in a power-spectrum form, it may be in a different format – e.g., Mef-frequency format – which is not directly suitable for speech waveform reconstruction). We use the following expression to obtain an estimate $\hat{P}(s'_t)$

$$\hat{P}(s'_t) = \frac{\sum_{\hat{\mathbf{s}}_{\tau_k}} \hat{g}_{\tau_k} P[\hat{s}_{\tau_k(t)}]}{N_{s'_t}} \tag{16}$$

where $\hat{s}_{\tau_k(t)}$ denotes the corpus frame corresponding to s'_t taken from the matching corpus segment $\hat{\mathbf{s}}_{\tau_k}$, $P[\hat{s}_{\tau_k(t)}]$ is the short-time DFT power spectrum associated with $\hat{s}_{\tau_k(t)}$, and the sum is over all the matching corpus segments that contain the estimate for s'_t, assuming that there are $N_{s'_t}$ such segments. Taking $\hat{P}(s'_t)$ as a clean speech power spectrum estimate, we can obtain a corresponding noise DFT power spectrum estimate, denoted by $\hat{P}(n_t)$, using a smoothed recursion

$$\hat{P}(n_t) = \alpha\hat{P}(n_{t-1}) + (1-\alpha)\max[P(x_t) - \hat{P}(s'_t), 0] \tag{17}$$

where $P(x_t)$ represents the noisy speech periodogram at time t, and α is a smoothing constant ($\alpha = 0.95$ in our experiments). Thus, we can form a Wiener filter with a time-varying transfer function F_t as follows

$$F_t = \frac{\hat{P}(s'_t)}{\hat{P}(s'_t) + \hat{P}(n_t)} \tag{18}$$

This filter takes the noisy frame DFT magnitude spectra as input and produces the corresponding speech frame DFT magnitude spectral estimates as output. Given the estimates, in our experiments, we use the corresponding noisy frame DFT phase spectra to build the speech frame waveform estimates. It is noted that phase estimation or phase-aware processing has become increasingly important in recent speech enhancement studies (e.g., [54], [55]). A cleaner phase should lead to improved speech quality.

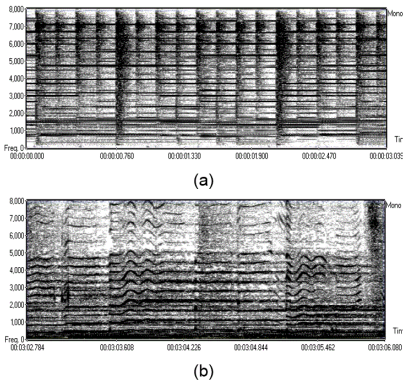


Fig. 2. Noises used in the WSJ0 test data, showing the noise spectra over a period of about three seconds. (a) Polyphonic musical ring. (b) Pop song.

V. Experimental Studies

A. Experimental Data, Models and Parameters

We have conducted speech enhancement experiments to evaluate the wide matching approach, which was implemented using the algorithm described in Section IV-B. The evaluation was focused on its performance without any noise estimation. Two databases, TIMIT and WSJ0, were used in the experiments. TIMIT contains a standard training set consisting of 3696 speech sentences from 462 speakers (326 male, 136 female). This training set was used as our corpus to provide element corpus speech segments to model free-text, free-speaker speech. TIMIT contains a standard core test set consisting of 192 speech sentences from 24 speakers (16 male, 8 female), all unseen to the training set. WSJ0 contains a 5K-vocabulary speaker-independent test set (SI_ET_05) consisting of 330 test sentences from eight speakers (four male, four female). These two test sets were used for enhancement experiments based on the same TIMIT training set. The additional WSJ0 test set was used to further evaluate the ability of the wide matching system to generalize to unseen speech, speakers and acoustic environments. The two test sets were added with variable noises to form the noisy test sentences. The TIMIT test sentences have an average duration about 2.8 s (or 280 frames), and the WSJ0 test sentences have an average duration about 7.3 s (or 730 frames).

For the TIMIT test data, we used six different types of noise from Aurora 4 [56]: airport, babble, car, restaurant, street and train station. These were each added to each test sentence at four different SNRs: 10, 5, 0 and -5 dB, respectively. The SNR was measured on each sentence basis. For the WSJ0 test data, we used two new types of noise showing some greater non-stationarity than the Aurora 4 noises: a polyphonic musical ring and a pop song with mixed music and voice of a female singer. The spectra of these two new noises are shown in Fig. 2. The speech signals were sampled

TABLE I
STATISTICS OF THE TEST SENTENCES, SHOWING THE MINIMUM, MAXIMUM AND AVERAGE LENGTH $L = (2\gamma + 1)K$ OF THE TEST SENTENCE SEGMENT CHAINS BEING CORRELATED (UNIT: FRAME).

Test set	Min	Max	Average
TIMIT	440	2233	1023
WSJ0	718	5166	2642

at 16 kHz and divided into frames of 25 ms with a frame period of 10 ms. In our experiments, for identifying matching segments, we represented each frame using its short-time Mel-frequency power spectral vector. This was obtained by first obtaining each frame's short-time DFT power spectrum, and then obtaining the corresponding Mel-frequency power spectrum by passing the DFT power spectrum through a Mel-frequency filterbank. We have tested filterbanks of variable numbers of channels within the range from 30 to some higher resolutions up to 128. In general, a higher-resolution power spectrum representation gave improved results, but also resulted in higher computational load. For the experiments in this paper, we used a 40-channel filterbank representation, which appeared to provide a good balance. As described in Section IV-C, when matching corpus segments were found, to reconstruct the clean speech waveform, we switched back to the highest-resolution DFT representation to perform the appropriate noise filtering and waveform reconstruction.

In our experiments, each clean corpus sentence was divided into short, consecutive, partially overlapped segments, to be used to form full-sentence speech estimates for arbitrary noisy sentences through segment chaining. More specifically, we formed the element corpus segments by taking each corpus frame in each corpus sentence and forming a segment around the frame with a fixed length of 11 frames [i.e., $\gamma = 5$ in (8), a figure borrowed from the previous DNN-based speech recognition studies on the TIMIT database [53]]. This ended up with about 1.1 million element corpus segments. As the TIMIT training set is small, we did not try to further compress the corpus data. With larger speech corpora, one may consider using the method described in [26], [28] to compactly represent the corpus data to improve the scalability. The noisy test sentences were each divided into a chain of consecutive segments each with the same length of 11 frames and with 8-frame overlap between adjacent segments. As indicated in (11) or (13), the underlying speech is estimated based on full-sentence segment chain correlation with a length of $L = (2\gamma + 1)K$ frames, where K is the number of segments in the chain. Table I shows the statistics of L of the noisy sentence chains $\mathbf{X}$ that have been correlated to derive the underlying speech estimates $\hat{\mathbf{S}}$, for the 192 TIMIT test sentences and 330 WSJ0 test sentences, respectively. We have found that some overlap between successive noisy segments helps to improve the estimation accuracy or smoothness, as found in DNN-based speech recognition studies. The large L of correlation, along with the constraint on the estimate's recognizability, contributed importantly to improving noise robustness without having noise estimation, to be seen later.

To form the recognizability score (12) for a speech sentence estimate, we trained a simple HMM-based phone recognizer using the TIMIT 3696 training sentences. The recognizer contains 61 three-state HMMs for the 61 TIMIT monophones, and a bigram phonetic language model trained with the phonetic transcripts of the 3696 training sentences. This recognizer was used to generate the triplets (14) for each training sentence/segment through forced frame-state alignment. As shown in (15), we used these triplets of the corpus data to form an approximate, but highly efficient, acoustic and language (bigram phonetic) constraint for the underlying speech estimate. Additionally, based on the state sequences of the training sentences, we created 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 appropriate histograms.

We used the iterative algorithm (13), with the step-by-step details given in Section IV-B, to derive the optimal estimate for each element corpus segment and hence for the whole sentence. In determining an optimal corpus segment estimate, we assumed that the gain for the corpus segment (i.e., g_{τ_k}) could vary within the range [0.5, 2]. We have tested some larger ranges, for example, [0.33, 3] and [0.25, 4], and found some small improvement. For each possible corpus segment candidate, we used a fast algorithm (the golden section method) to search for its optimal gain within the range to maximize the full-sentence ZNCC. The extra computation was found to be minimal. Unless otherwise indicated, throughout the experiments we assumed $\lambda = 0.1$, which is the constraining Lagrange multiplier in (13).

B. Experimental Results on TIMIT Test Data

In this section we present the experimental results on the TIMIT test set with six types of Aurora 4 noise, and comparison with existing speech enhancement methods. We used three standard objective measures for the evaluation, which were Segmental SNR (SSNR), PESQ (perceptual evaluation of speech quality) and STOI (short-time objective intelligibility) [57]. We compare wide matching with some popular and important *frame*-based speech enhancement methods, which include LogMMSE [4], LogMMSE [5], [17], Wiener filtering [58] and KLT [59]. The two different versions of LogLMMSE mainly differ in their methods of tracking the background noise. For convenience, we note them as LogMMSE-1 and LogMMSE-2, respectively. Additionally, we also include a reduced form of the wide matching method in the comparison. This reduced method performs full-sentence correlation alone without constraint of the recognizability [i.e., (11) or (13) with $\lambda = 0$]. For convenience, we call it the *full-sentence correlation* (FSC) method. This method is included to show the significance of the recognizability constraint in terms of improving the speech estimation accuracy.

For clarity, we compare these methods by using their average scores across the different types of noise, as a function of the input noisy sentence SNR. Thus, each score for each SNR condition is obtained by averaging over 1152 noisy test sentences (192 test sentences per noise type $\times$ 6 noise types). The wide matching method did not use any noise estimation

TABLE II
RESULTS ON THE TIMIT TEST SET, COMPARING BETWEEN WIDE MATCHING AND SOME EXISTING FRAME-BASED AND SEGMENT-BASED ENHANCEMENT METHODS, AND FULL-SENTENCE CORRELATION (FSC) WHICH IS WIDE MATCHING WITH $\lambda = 0$, ON SCORES OF THE ENHANCED SPEECH AVERAGED OVER SIX TYPES OF NOISE, AS A FUNCTION OF THE INPUT NOISY SENTENCE SNR.

	Method\SNR (dB)	-5	0	5	10	Clean
SSNR	Unprocessed	-6.79	-4.26	-1.11	2.48	
	LMMSE-1	-2.59	-0.31	2.23	4.96	19.28
	LMMSE-2	-4.37	-2.21	0.13	3.01	12.02
	Wiener filter	-3.94	-1.42	1.40	4.38	**19.69**
	KLT	-1.62	0.36	2.64	5.05	15.28
	LMS	**0.61**	1.90	2.93	3.62	17.74
	FSC	-0.35	1.47	3.17	4.48	18.16
	Wide matching	0.51	**2.33**	**3.96**	**5.16**	18.11
PESQ	Unprocessed	1.39	1.74	2.09	2.44	
	LMMSE-1	1.60	2.01	2.38	2.72	**4.39**
	LMMSE-2	1.56	2.01	2.42	2.78	4.35
	Wiener filter	1.53	1.93	2.31	2.67	4.41
	KLT	1.25	1.75	2.21	2.63	4.32
	LMS	1.71	2.09	2.48	2.76	4.26
	FSC	1.51	1.94	2.30	2.58	4.05
	Wide matching	**1.76**	**2.19**	**2.54**	**2.79**	4.22
STOI	Unprocessed	0.58	0.69	0.79	0.88	
	LMMSE-1	0.54	0.67	0.78	0.86	**0.99**
	LMMSE-2	0.54	0.67	0.78	0.87	0.99
	Wiener filter	0.56	0.68	0.79	0.88	0.99
	KLT	0.56	0.70	0.81	0.89	0.99
	LMS	0.68	0.76	0.81	0.85	0.97
	FSC	0.59	0.73	0.81	0.84	0.94
	Wide matching	**0.68**	**0.79**	**0.86**	**0.90**	0.98

while the conventional methods each used an algorithm to estimate the noise. Table II presents the comparison results for each of the three quality measures. The new method outperformed all these frame-based enhancement methods on all the three measures for each noise type in all the noisy conditions, and the improvement appeared to be more significant for the low SNR conditions. The wide matching method also outperformed the FSC method in all the noisy conditions (see more discussions on this later). In comparison to some of the conventional methods, we experienced a slight drop in performance when taking clean speech as input for enhancement. We observed the similar phenomenon in our earlier corpus-based LMS method (e.g., [28]). This is because in the corpus-based approaches, alien (i.e., corpus) data are used as part of the method to reconstruct the underlying speech, while in the conventional frame-based methods only the original (high-SNR) data are used in the reconstruction.

Further, we compared wide matching with our latest, *segment*-based LMS method [28], which uses an iteration algorithm to alternately estimate the noise and the longest underlying speech segments with matching clean training segments until convergence. The longest matching training segments found are used to form the clean speech estimates. The scores for the enhanced speech by the LMS method are included in Table II. The wide matching method, without using any noise estimation, outperformed the LMS method in all the noisy conditions except in one case, SNR = -5 dB, in which LMS scored slightly higher in SSNR than wide matching. As indicated in Table II, LMS performed better than the frame-based enhancement methods in the low SNR conditions, in terms of LMS achieving higher scores in all the three measures. LMS also outperformed FSC in many noisy cases, especially in terms of the PESQ and STOI scores.

TABLE III
RESULTS ON THE TIMIT TEST SET, SHOWING SCORES OF THE ENHANCED SPEECH BY WIDE MATCHING WITH OR WITHOUT INCLUDING THE DELTA POWER SPECTRA, AVERAGED OVER SIX TYPES OF NOISE AS A FUNCTION OF THE INPUT NOISY SENTENCE SNR.

Measure	With delta\SNR (dB)	-5	0	5	10
SSNR	No	0.51	2.33	3.96	5.16
	Yes	0.40	2.30	3.95	5.18
PESQ	No	1.76	2.19	2.54	2.79
	Yes	1.81	2.23	2.57	2.83
STOI	No	0.68	0.79	0.86	0.90
	Yes	0.69	0.80	0.86	0.91

Finally, we conducted experiments to test if we could further improve wide matching's performance by adding the short-term dynamic power spectra into the speech representation for correlation based matching. The short-term dynamic spectra, often called derivative or delta spectra in speech recognition, are typically calculated as linear regression coefficients of the static spectra over a short segment of a time trajectory. They are suitable to be added in the new method because the assumption that environmental noise is approximately additive in the power spectra is also applicable to the delta power spectra. Therefore our hypothesis in Section II for the convergence of the long-segment ZNCC to noise immunity should also hold for the augmented speech representation, in which each frame is represented by a combination of static and dynamic power spectra. We further added the so-called delta-delta power spectra. So in the revised representation each frame contained 120 coefficients (the original 40 plus the corresponding first-order and second-order delta coefficients). Table III shows the results and comparison. The addition of the dynamic power spectra did bring some small improvement, mainly in the PESQ and STOI scores. Later we will show its significance in improving the phone matching rate.

C. Experimental Results on WSJ0 Test Data

The WSJ0 test data were used to provide greater uncertainty of speech, speaker and acoustic conditions with respect to the TIMIT training set. As mentioned earlier, we also used two new types of noise (Fig. 2), which generally exhibit greater time variation than the Aurora 4 noises, to generate the noisy WSJ0 sentences. For the clarity of presentation, we show the average scores across these two types of noise as a function of the input noisy sentence SNR. Thus, each score for each SNR condition is obtained by averaging over 660 noisy test sentences (330 test sentences per noise type $\times$ 2 noise types). For this experiment, the wide matching system used the delta and delta-delta power spectra. We chose the conventional frame-based LMMSE-1 (LMMSE-2 performed poorer than LMMSE-1 for this case) and the segment-based LMS for comparison.

Table IV presents the comparison results for each of the three speech quality measures. The wide matching method

TABLE IV
RESULTS ON THE WSJ0 TEST SET, COMPARING BETWEEN WIDE MATCHING, FRAME-BASED LMMSE-1 AND SEGMENT-BASED LMS ON SCORES OF THE ENHANCED SPEECH AVERAGED OVER TWO NEW TYPES OF NONSTATIONARY NOISE, AS A FUNCTION OF THE INPUT NOISY SENTENCE SNR.

	Method\SNR (dB)	0	5	10
S	Unprocessed	-4.24	-1.40	1.74
S	LMMSE-1	-3.39	-0.82	1.92
N	LMS	0.67	1.60	2.26
R	Wide matching	**2.00**	**3.17**	**3.74**
P	Unprocessed	1.90	2.23	2.56
E	LMMSE-1	1.88	2.24	2.60
S	LMS	2.35	2.56	2.74
Q	Wide matching	**2.38**	**2.61**	**2.75**
S	Unprocessed	0.83	0.89	**0.94**
T	LMMSE-1	0.78	0.87	0.91
O	LMS	0.83	0.88	0.89
I	Wide matching	**0.85**	**0.89**	0.90

TABLE V
RESULTS ON THE WSJ0 TEST SET, AUTOMATIC SPEECH RECOGNITION WORD ACCURACY (%) FOR THE NOISY SPEECH AND ENHANCED SPEECH, AS A FUNCTION OF THE INPUT NOISY SENTENCE SNR.

Method\SNR (dB)	0	5	10
Unprocessed	9.65	25.7	52.8
LMMSE-1	7.35	19.7	41.7
LMS	36.0	52.9	65.1
Wide matching	45.3	59.7	68.9

scored higher than all the other methods, except in the high SNR (10 dB) case, in which the unprocessed noisy speech scored highest in STOI over all the enhanced speech. Due to the highly nonstationary characteristics of the noise, the LMMSE method based on noise tracking failed to effectively remove much of the noise, indicated by its lower SSNR. While SSNR may not be a reliable indicator of speech quality, but in this case, low SSNR did lead to poor accuracy of automatic speech recognition, as shown below.

Next, we passed the enhanced speech, produced by the various methods, to an automatic speech recognizer trained using *clean* speech data for the WSJ0 database. The recognition accuracy, thus, can be used as an indicator of the distortion of the enhanced speech as against the clean speech. The recognizer took a GMM-HMM architecture, was built following the HTK WSJ Training Recipe [60], and achieved over 92% word accuracy on the clean WSJ0 test set used in our enhancement experiment. In the recognition evaluation, we used the same word insertion/deletion penalties for all the methods. Table V shows the word accuracy results. The LMMSE method failed to improve the word accuracy from the unprocessed noisy speech, while the wide matching method delivered significant improvement in all the noisy conditions. Wide matching (based on sentences) also outperformed LMS (based on segments), especially in the low SNR conditions.

D. Wide Matching Analysis

We try to understand how the wide matching method improved noise robustness without having noise estimation, based on the TIMIT test set. Table VI uses an example (SNR = 0 dB) to show the importance of the length of the segments being correlated and the constraint on the estimate's recognizability. It shows a comparison between wide matching with a sentence-long correlation length as shown in Table I, and segment correlation with a fixed length of 11 frames, averaged over all the test sentences from all the six noise types. In between is the full-sentence correlation (FSC) based on segment chains without constraint of their recognizability [i.e., wide matching with $\lambda = 0$ in (11) or (13)]. Segment correlation assumes independence between element speech segments, is prone to noise corruption, and was used to provide the initial estimates for wide matching (see Section IV-B). With full-sentence correlation one may obtain better-quality enhanced speech, as shown in Table VI. However, when noise dominates the observed data the matching segment chain can become noisy (see the example below). Table VI shows that by forcing the potential matching segment chain to be most recognizable as speech, as implemented in wide matching with $\lambda > 0$, we can relieve this problem, as demonstrated by the improved quality scores. The maximum recognizability constraint helps obtain noise-independent speech estimates, and this constraint may only be effectively applied on long speech segments (e.g., phrases or sentences). In our experiments, we did not optimize the λ value for the wide matching algorithm. Instead, we tested a range of λ values and found that its performance is quite stable across different noise types, SNR levels and speech databases. Table VI shows an example.

TABLE VI
IMPORTANCE OF THE CORRELATION LENGTH AND RECOGNIZABILITY CONSTRAINT (UNCONSTRAINED WHEN $\lambda = 0$), AVERAGED OVER SIX TYPES OF NOISE WITH SNR=0 DB.

Method\Measure	SSNR	PESQ	STOI
Segment correlation	1.01	1.82	0.69
Full-sentence correlation (wide matching with $\lambda = 0$)	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

Since the whole wide matching system effectively performed noisy speech recognition based on clean speech training, we can use the phone matching rate between the underlying speech and the matching corpus speech as part of the evaluation criteria. Specifically, based on the (rather simple) clean phone recognizer we implemented for the system, we chose the frame-level phone matching rate as the criterion, i.e., a matching corpus frame is considered to be correct if it has the same phone label as the underlying speech frame (this criterion was used previously in [61]). We folded the original 61 TIMIT phone labels into the standard 39 phone classes to calculate the matching rate. Fig. 3 shows the results, comparing full-sentence correlation without constraint on recognizability and wide matching with this constraint ($\lambda = 0.1$), and further with dynamic spectral features. The results once again demonstrate the importance of recognizability optimization in improving noise robustness, especially in the low SNR conditions, in terms of achieving higher frame-level phone matching accuracy in the enhanced speech. The results also

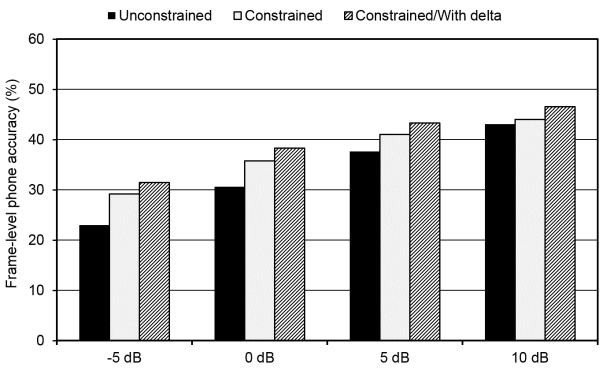


Fig. 3. Frame-level phone matching rate for full-sentence correlation without constraint on recognizability, wide matching with constraint on recognizability and wide matching also with delta power spectra.

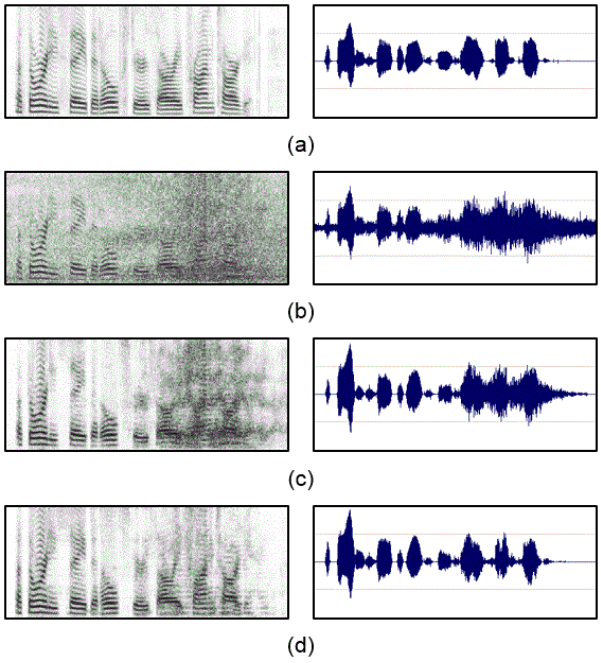


Fig. 4. A speech enhancement example. (a) Clean speech sentence. (b) Noisy speech sentence with nonstationary street noise at a sentence-level SNR = 0 dB. (c) Estimate based on full-sentence correlation (FSC) without constraint on recognizability. (d) Estimate based on wide matching.

show the significance of including the dynamic power spectra in improving the frame-level phone matching accuracy, though this extra improvement did not appear to lead to further sizable improvement in the other three quality measures as shown in Table III.

Fig. 4 shows an example, for recovering a speech sentence [Fig. 4(a)] from nonstationary street noise with a sentence-level SNR = 0 dB. Fig. 4(b) shows the noisy sentence. In this example, apparently, the first half of the noisy sentence was dominated by speech and the second half of the noisy sentence was dominated by noise. For nonstationary speech and noise, it is typical to have varying local SNRs in a sentence. Fig. 4(c) shows the estimate based on full-sentence correlation (FSC) without constraint on recognizability, and Fig. 4(d) shows the estimate based on wide matching. While

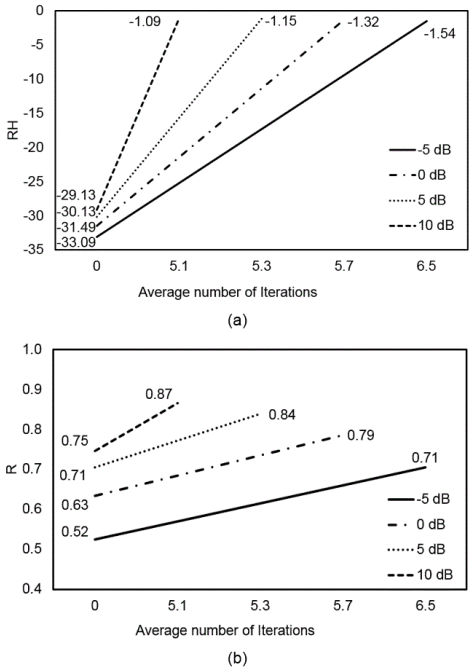


Fig. 5. Convergence of the iterative estimation algorithm in the experiments, showing the average number of iterations used to reach convergent estimates, and the increases in (a) the recognizability-constrained full-sentence ZNCC (RH) value and (b) the corresponding full-sentence ZNCC (R) value after convergence, averaged over the six types of noise as a function of the input noisy sentence SNR.

both methods appear to be effective in reproducing the first half of the speech sentence, only wide matching appears to be effective in recovering the second half of the speech sentence from the overwhelming noise. Without constraint on the estimate's recognizability, the correlation based on short segment chaining apparently produced an estimate that tended to follow the dominant signal (whether it is speech or noise) as a result of maximum correlation.

Finally, we show the convergence of the iteration algorithm used to implement wide matching, described in Section IV-B. Fig. 5 summarizes the average number of iterations used in our experiments to reach convergent estimates, and the initial and convergent values of the iteration of the recognizability-constrained full-sentence ZNCC (RH) and the corresponding full-sentence ZNCC (R), respectively, as a function of the input noisy sentence SNR averaged overall all the noise types. These statistics are accumulated from the experimental results with $\lambda = 0.1$ and without using dynamic power spectra. We have not seen a single case in which the iteration decreased the appropriate full-sentence ZNCC values.

VI. Concluding Remarks

Methods for single-channel speech enhancement are mainly frame, multi-frame or segment based. All require knowledge about the noise. The method described in this paper is a complement to existing methods and may be viewed as *sentence based*. This research aims to reduce or effectively remove the requirement for knowledge of noise. This is significant as it could lead to a technique that is capable of retrieving speech

from noisy data without requiring vast amounts of noisy training data or noise statistics. With an oracle experiment, and with some hypothetical studies, we first showed that by directly matching long speech segments based on ZNCC we could potentially increase significantly noise immunity for speech enhancement without requiring noise knowledge. This paper described a realization of this approach, namely wide matching, for practical use. The core part of this realization is the integration of a clean speech recognizer to regularize the formation of the potential matching estimate subject to maximum recognizability. Our conjecture is that this could correspond to one technique used by humans for picking out speech in strong noise – to try to make sense of the speech. For computational reasons, we chose to use a simple, approximate clean speech recognizer in our experiments. But as indicated by our experimental results, for moderate-length speech sentences, and for a family of difficult real-world noise, the new method could outperform a range of state-of-the-art speech enhancement methods without any estimation of the noise. Presently, we are studying the integration of more advanced clean speech recognition into the system for further improved performance, and the possible extension of the new method for robust speech recognition.

REFERENCES

[1] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, pp. 113–120, 1979.

[2] J. S. Lim and A. V. Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of IEEE*, vol. 67, pp. 1586–1604, 1979.

[3] R. J. McAulay and K. L. Malpass, "Speech enhancement using a soft-decision noise suppression filter," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, pp. 137–145, 1980.

[4] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, pp. 443–445, 1985.

[5] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, pp. 2403–2418, 2001.

[6] T. Lotter and P. Vary, "Speech enhancement by MAP spectral amplitude estimation using a super-gaussian speech model," *EURASIP J. Appl. Signal Process.*, vol. 2005, pp. 1110–1126, 2005.

[7] S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, pp. 163–176, 2006.

[8] D. H. R. Naidu and S. Srinivasan, "A Bayesian framework for robust speech enhancement under varying contexts," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2012, pp. 4557–4560.

[9] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman filter for speech enhancement in cocktail party scenarios using a codebook-based approach," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 191–195.

[10] A. Kundu, S. Chatterjee, A. S. Murthy, and T. V. Sreenivas, "GMM based Bayesian approach to speech enhancement in signal/transform domain," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2008, pp. 4893–4896.

[11] M. E. Deisher and A. S. Spanias, "HMM-based speech enhancement using harmonic modeling," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 1997, pp. 1175–1178.

[12] H. Sameti and L. Deng, "Nonstationary-state hidden Markov model representation of speech signals for speech enhancement," *Signal Processing*, vol. 82, pp. 205–227, 2002.

[13] N. Mohammadiha, R. Martin, and A. Leijon, "Spectral domain speech enhancement using HMM state-dependent super-gaussian priors," *IEEE Signal Process. Lett.*, vol. 20, pp. 253–256, 2013.

[14] F. Deng, C. Bao, and W. B. Kleijn, "Sparse hidden markov models for speech enhancement in non-stationary noise environments," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1973–1987, 2015.

[15] P. K. Ghosh, A. Tsiartas, and S. S. Narayanan, "Robust voice activity detection using long-term signal variability," *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, pp. 600–613, 2011.

[16] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, 2001.

[17] I. Cohen, "Noise spectrum estimation in adevrse environments: improved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, pp. 466–475, 2003.

[18] S. Rangachari and P. Loizou, "A noise estimation algorithm for highly nonstationary environments," *Speech Communication*, vol. 28, pp. 220–231, 2006.

[19] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1383–1393, 2012.

[20] X. Xiao and R. M. Nickel, "Speech enhancement with inventory style speech resynthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, pp. 1243–1257, 2010.

[21] R. Nickel, R. 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.

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

[23] J. T. Geiger, J. F. Gemmeke, B. Schuller, and G. Rigoll, "Investigating NMF speech enhancement for neural network based acoustic models," in *Proceedings Interspeech*. ISCA, 2014, pp. 2405–2409.

[24] D. Baby, T. Virtanen, J. F. Gemmeke, and H. V. hamme, "Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1788–1799, 2015.

[25] Z.-Q. Wang, Y. Zhao, and D. L. Wang, "Phoneme-specific speech separation," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 146–149.

[26] 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.

[27] 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.

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

[29] M. Delcroix *et al.*, "Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds," *Comput. Speech Lang.*, vol. 27, pp. 851–873, 2013.

[30] A. Ogawa, K. Kinoshita, T. Hori, T. Nakatani, and A. Nakamura, "Fast segment search for corpus-based speech enhancement based on speech recognition technology," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 1557–1561.

[31] H. Seo, H. Kang, and F. Soong, "A maximum a posterior-based reconstruction approach to speech bandwidth expansion in noise," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2014, pp. 6128–6132.

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

[33] Y. Wang and D. L. Wang, "Towards scaling up classification-based speech separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, pp. 1381–1390, 2013.

[34] 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, pp. 65–68, 2014.

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

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

[37] 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.

[38] K. Han, Y. Wang, D. L. Wang, W. S. Woods, I. Merks, and T. Zhang, "Learning spectral mapping for speech dereverberation and denoising," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 982–992, 2015.

[39] X.-L. Zhang and D. L. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, pp. 967–977, 2016.

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

[41] A. Graves, A. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2013, pp. 6645–6649.

[42] J. 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.

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

[44] M. 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.

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

[46] F. Li, P. 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.

[47] T. Yoshioka, K. Ohnishi, F. Fang, and T. Nakatani, "Noise robust speech recognition using recent developments in neural networks for computer vision," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 5730–5734.

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

[49] Y. Huang and J. Benesty, "A multi-frame approach to the frequency-domain single-channel noise reduction problem," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1256–1269, 2012.

[50] A. Schasse and R. Martin, "Estimation of subband speech correlations for noise reduction via MVDR processing," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 20, pp. 1355–1365, 2014.

[51] J. Ming and D. Crookes, "Wide mathcing – an approach to improving noise robustness for speech enhancement," in *Proceedings Int. Conf. Acoust., Speech, Signal Process.* IEEE, 2016, pp. 5910–5914.

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

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

[54] P. Mowlaee and J. Kulmer, "Phase estimation in single-channel speech enhancement: Limits-potential," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, pp. 1283–1294, 2015.

[55] T. Gerkmann, M. Krawczyk-Becker, and J. L. Roux, "Phase processing for single channel speech enhancement: History and recent advances," *IEEE Signal Processing Magazine*, vol. 32, pp. 55–66, 2015.

[56] 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.

[57] C. Taal, R. 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.

[58] 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.

[59] 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.

[60] K. Vertanen, "Baseline WSJ acoustic models for htk and sphinx: training recipes and recognition experiments," *Tech. Rep., Cavendish Laboratory*, 2006.

[61] L. Deng and D. Yu, "Deep convex net: a scalable architecture for speech pattern classification," in *Proceedings Interspeech.* ISCA, 2011, pp. 2285–2288.

Ji Ming (M'97) received the B.Sc. degree from Sichuan University, Chengdu, China, in 1982, the M.Phil. degree from Changsha Institute of Technology, Changsha, China, in 1985, and the Ph.D. degree from Beijing Institute of Technology, Beijing, China, in 1988, all in electronic engineering.

He was Associate Professor with the Department of Electronic Engineering, Changsha Institute of Technology, from 1990 to 1993. Since 1993, he has been with the Queen's University Belfast, Belfast, U.K., where he is currently a Professor in the School of Electronics, Electrical Engineering and Computer Science. From 2005 to 2006, he was a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge. His research interests include speech and language processing, image processing, signal processing and pattern recognition.

Danny Crookes (SM'12) was appointed to the Chair of Computer Engineering in 1993 at Queen's University Belfast, Belfast, U.K., and was Head of Computer Science from 1993-2002. He is currently Director of Research for Speech, Image and Vision Systems at the Institute of Electronics, Communications and Information Technology, Queen's University Belfast. His current research interests include the use of novel architectures (especially GPUs) for high performance speech and image processing. Professor Crookes is currently involved in projects in automatic shoeprint recognition, speech separation and enhancement, and processing of 4D confocal microscopy imagery. Professor Crookes has over 200 scientific papers in journals and international conferences.