Wide Matching - An Approach to Improving Noise Robustness for Speech Enhancement


Published in:
Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing

Document Version:
Peer reviewed version

Queen's University Belfast - Research Portal:
Link to publication record in Queen's University Belfast Research Portal

Publisher rights
© 2016 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

General rights
Copyright for the publications made accessible via the Queen's University Belfast Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Queen's institutional repository that provides access to Queen's research output. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact openaccess@qub.ac.uk.
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 discriminat ing 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 speech estimator. 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 \( s_i \) 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 \( X = (x_1, x_2, ..., x_T) \), where each noisy STPS vector \( x_i \) can be approximately expressed as \( x_i = s_i' + n_i \), with \( s_i' \) representing the underlying speech STPS vector and \( n_i \) representing the noise STPS vector. For each noisy test sequence \( X \), we aimed at finding a clean speech STPS sequence from the training data that matches the underlying speech STPS sequence \( S' = (s_1', s_2', ..., s_T) \). We obtained an estimate \( \hat{s}_i \) for each \( s_i' \) by maximizing the following normalized sample correlation coefficient

\[
\hat{s}_i' = \arg \max_{s_i} R(x_i, s_i') \\
= \max_{s_i} \frac{\sum_{l=-L}^{L} (x_{i,l} - m)\tilde{R} (s_{i,l} - m)}{\sigma_x \sigma_{s_i}}
\]  

where \( x_{i,l} \) denotes a segment of noisy STPS vectors centered at frame \( x_i \) from \( x_{i-L} \) to \( x_{i+L} \), \( m \) is the mean vector of \( x_{i-L} \) and \( x_{i+L} \), and \( \sigma_x \) is the mean-removed Euclidian norm of \( x_{i-L} \) and \( x_{i+L} \). The same definition applies to the training STPS segment \( s_{i-L} \), with mean vector \( m_{s_i} \) and mean-removed Euclidian norm \( \sigma_{s_i} \). In this experiment, we included the clean test sentence (i.e., the perfect match) in the training data, to examine under what condition it would be chosen. Fig. 1 shows the accuracy rates of finding the perfectly matching speech frames as a function of the segment length \( 2L + 1 \), averaged over all the frames of all the core test sentences. We see that as \( L \) increases, the perfect match can be found with a rapidly increasing probability, regardless of the types of noise. This oracle experiment, and the theory described below, suggest the potential to obtain accurate speech estimates just by correlating between very long speech segments, without requiring estimation of the noise that is independent of the speech. The remainder of the paper is aimed at generalizing the experiment 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 speech 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-range temporal dependence of speech has been an active research topic. This is important because it is easier to dis-
tistinguish speech from non-speech noise on a longer time basis. Cur-
rent methods were able to model the temporal dependence within
some speech segments, for example, the speech segments corre-
sponding to some phonetic classes [12–14], and the speech segments
about 9-31 frames long in DNN systems [6–10]. Our longest match-
ing segment (LMS) approach [15–17] was able to find the presum-
ably longest individual speech segments between the training and
test speech that match. However, these segment-based methods pro-
cess the individual segments either independently or with limited
correlation (limited by the available training data) as in some recurr-
ent DNNs [1, 2, 4, 5], with limited effect in capturing the longer-
distance, cross-segment dependence of speech for noise separation.
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 robust-
ness, 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 follow-
ing, 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 STPS vectors
\( x_t = s'_t + n_t \), we can decompose the normalized sample correlation
coefficient \( R(x_{t \pm L}, s_{r \pm l}) \), defined in (1), into two terms
\[
R(x_{t \pm L}, s_{r \pm l}) = \frac{\sigma_x}{\sigma_s} R(s'_{t \pm L}, s_{r \pm l}) + \frac{\sigma_n}{\sigma_s} R(n_{t \pm L}, s_{r \pm l}) \tag{2}
\]
where \( s'_{t \pm L} \) represents the underlying speech segment in the noisy
observation segment \( x_{t \pm L} \), from \( s'_{t-l} \) to \( s'_{t+l} \), and \( n_{t \pm L} \) repre-
sents the corresponding noise segment from \( n_{t-l} \) to \( n_{t+l} \), with \( m_x \), \( m_n \) (implied) and \( \sigma_x \), \( \sigma_n \) representing the mean vector and
mean-removed Euclidean norm of \( s'_{t \pm L} \) and \( n_{t \pm L} \), respectively. The
first term is the normalized sample correlation between the underly-
ing speech segment \( s'_{t \pm L} \) and the training speech segment \( s_{r \pm l} \),
weighted by \( \sigma_x/\sigma_s \) 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 \( \sigma_n/\sigma_s \) which is again inde-
dependent of the training speech segment, subject only to the SNR in
the observation. For independent noise and large \( L \), we may assume
\[
R(n_{t \pm L}, s_{r \pm l}) = \sum_{l=0}^{L} (m_{t+l} - m_n) \sum_{l=0}^{L} s_{r+l} - m_s
\]
\[ \propto \mathbb{E}[(n_{t-l} - m_n)^T (s_{r+l} - m_s)] \tag{3} \]
(Average over all possible concatenations of the training segments,
for perfectly matching speech \( S \).

3. WIDE MATCHING FOR SPEECH ENHANCEMENT

3.1. A constrained optimization problem

Let \( X = (x_1, x_2, \ldots, x_T) \) be a noisy test sentence with the un-
derlying speech sentence \( S = (s'_1, s'_2, \ldots, s'_T) \) unseen in the training
data. We seek an approach to concatenating a number of short train-
ing 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 signal to be cor-
related and hence the robustness to independent noise, i.e., to best
fulfill Condition 1 as required in the above theory.

Suppose we can divide \( X \) into some \( K \) consecutive segments,
denoted by \( X = (x_{11}, x_{12}, \ldots, x_{1K}) \), where each segment \( x_{1k} \) is
centered at some time \( t_k \) with frames from \( t_{k-\gamma} \) to \( t_{k+\gamma} \),
where \( \gamma \) defines the length of the element segment. For simplicity,
we assume a common \( \gamma \) for all the element segments and so \( \gamma \) can
be implied in the expression. Adjacent element segments can have
some overlap to improve the smoothness. In a similar way, denote by
\( S = (s_1, s_2, s_3, \ldots, s_K) \) a concatenation of \( K \) clean train-
ing segments as an estimate of the underlying speech sentence in
\( X \), where each element training segment \( s_k \) consists of consecutive
frames from \( s_{k-\gamma} \) to \( s_{k+\gamma} \), and \( g_{k-1} \) is the gain of the element
training segment in forming the sentence estimate. In \( S \), different training
segments \( s_k \) can come from different training sentences/contexts to
simulate unseen test speech. We estimate the optimal \( S \) based on the
sentence-wide, normalized sample correlation coefficient between
\( X \) and \( S \). After some manipulation, this can be written as
\[
R(X, S) = \frac{\sum_{k=1}^{K} g_k \sum_{l=0}^{\gamma} s_{k+l} - m_S\bar{m}_S - Lm_S m_S}{\sigma_X \sigma_S} \tag{6}
\]
where \( L = (2\gamma + 1)K \) is the sample length of the two full sentences
being correlated, \( m_S \) and \( \sigma_S \) are the global mean vector and
mean-removed Euclidean norm of the training segment concatenation \( S \),
\[ m_S = \frac{1}{\gamma} \sum_{k=1}^{K} g_k \sum_{l=0}^{\gamma} s_{k+l} - m_S \tag{7} \]
\[ \sigma_S^2 = \sum_{k=1}^{K} g_k^2 \sum_{l=0}^{\gamma} s_{k+l} - m_S \tag{8} \]
The above expressions apply to \( m_X \) and \( \sigma_X \), the global mean vector and
mean-removed Euclidean norm of the noisy segment sequence
\( X \) (without the gain terms). In the above sentence-wide correlation
coefficient \( R(X, S) \), 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 \( S \) by
maximizing the normalized correlation coefficient \( R(X, S) \) over all
possible concatenations of the training segments \( g_k s_k \).
However, not all of the concatenations constitute realistic speech;
some concatenations with larger \( R(X, S) \) may simulate the original
noisy speech \( X \) well (as indicated in (2), the correlation coefficient
\( R(X, S) \) for perfectly matching speech \( S' \) and \( S \) is confined around

image
\( \sigma_s / \sigma_g < 1 \) 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 concatenating some very short speech segments. To make the estimate to be valid speech, which is independent of the noise and hence fulfills Condition 2 of the above theory, we use the estimate’s recognizability, to a speech recognizer trained with clean data, to regularize the formation of the optimal estimate \( \hat{S} \). 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

\[
\hat{S} = \arg \max_S [\log R(X, S) + \lambda \log H(S)]
\]

where \( H(S) \) represents the confidence score of the estimate \( S \) to be valid speech, and \( \lambda \) is a Lagrange multiplier. For the proof of the concept, this paper uses an HMM-based phone recognizer to provide \( H(S) \). The recognizer is trained with clean speech and learns the acoustic HMMs for context-independent phones, a bigram language model, and 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 concatenations \( S \), the concatenation constituting a valid clean speech is most recognizable to the recognizer, in terms of achieving a large score \( H(S) \) (this is because valid clean speech is most likely to simultaneously fulfill the acoustic, phone language, state duration and phone duration constraints of clean speech learned by the recognizer). If such a long concatenation with a large noise-independent speech confidence score \( H(S) \) simultaneously has a large correlation coefficient \( R(X, S) \) with the noisy signal \( X \), or vice versa, then it can be assumed that this is an optimal estimate of the underlying speech in \( X \). 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 \( \hat{S} \). Given a noisy sentence \( X \), we start with an initial estimate \( \hat{S} \) by separately estimating each element training segment \( s_{r_k} \) based on maximizing the segment-level constraint correlation coefficient \( R(x_{s_{r_k}}, s_{r_k}) \) with a unit gain \( g_{r_k} \). 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 \( g_{r_k} s_{r_k} \) in the order from \( k = 1 \) to \( K \). In the \( j \)th iteration, to obtain a new estimate of the optimal \( k \)th element training segment, denoted by \( \hat{g}_{s_{r_k}} \), we maximize (9) with respect to \( g_{r_k} s_{r_k} \), with the preceding element training segments \( g_{r_m} s_{r_m} \) (\( m > k \)) from the \( (j - 1) \)th iteration, and the preceding element training segments \( g_{r_m} s_{r_m} \) (\( m < k \)) from the \( j \)th iteration. Therefore in the \( j \)th iteration and \( k \)th stage, the optimal speech sentence estimate to be determined can be expressed as

\[
\hat{S}^j(g_{r_k} s_{r_k}) = (\hat{g}_{r_k}^1 s_{r_k}^1, \ldots, \hat{g}_{r_k}^{j-1} s_{r_k}^{j-1}, g_{r_k} s_{r_k}, \hat{g}_{r_k}^{j-1} s_{r_k}^{j-1}, \ldots, \hat{g}_{r_k}^0 s_{r_k}^0),
\]

which is only a function of \( g_{r_k} s_{r_k} \), 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

\[
\hat{S}^j(\hat{g}_{r_k}^j s_{r_k}^j) = \arg \max_{g_{r_k} s_{r_k}} \log R(X, S^j(g_{r_k} s_{r_k})) + \log H(S^j(g_{r_k} s_{r_k}))
\]

where \( \hat{g}_{r_k} s_{r_k} \) 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 3969 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 \( X \) that have been correlated to derive the speech estimates \( \hat{S} \), for the 192 test sentences. We take the overlap between element segments as effective signals as we found that some overlap did help improve

<table>
<thead>
<tr>
<th>Min</th>
<th>Max</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>440</td>
<td>2233</td>
<td>1023</td>
</tr>
</tbody>
</table>
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 recognizer. It can 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 $\lambda$ 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 $R(X, S)$, respectively, averaged overall the test sentences and noise types.
6. REFERENCES


