Speech Enhancement from Additive Noise and Channel Distortion -
a Corpus-Based Approach

Ji Ming, Danny Crookes

Institute of ECIT, Queen’s University Belfast, Belfast BT7 1NN, UK
j.ming@qub.ac.uk, d.crookes@qub.ac.uk

Abstract

This paper presents a new approach to single-channel speech enhancement involving both noise and channel distortion (i.e., convolutive noise). The approach is based on finding longest matching segments (LMS) from a corpus of clean, wideband speech. The approach adds three novel developments to our previous LMS research. First, we address the problem of channel distortion as well as additive noise. Second, we present an improved method for modeling noise. Third, we present an iterative algorithm for improved speech estimates. In experiments using speech recognition as a test with the Aurora 4 database, the use of our enhancement approach as a preprocessor for feature extraction significantly improved the performance of a baseline recognition system. In another comparison against conventional enhancement algorithms, both the PESQ and the segmental SNR ratings of the LMS algorithm were superior to the other methods for noisy speech enhancement.

Index Terms: corpus-based speech model, longest matching segment, speech enhancement, speech recognition

1. Introduction

In speech enhancement, most current approaches impose few or very loose constraints on the underlying speech to be estimated. As a result, they require specific knowledge about the noise for noise removal and hence speech recovery. The typical constraint or prior for the underlying speech is the probability distribution of the speech short-time DFT coefficients or spectral amplitudes [1]-[5]. The common methods for noise estimation include prediction by using neighboring measurements without significant speech content based on voice activity detection, minimum statistics, time-recursive averaging, MMSE-based high-resolution noise DFT estimation and their combination [6]-[11]. Data-driven speech models, built on the training data of real speech, represent a different way of imposing prior or constraint on the speech to be estimated. Common speech models include vector-quantization (VQ) codebooks, Gaussian mixture models, hidden Markov models (HMM), and inventory-based models [12]-[19]. Some of the modeling techniques used in robust speech recognition have also found use in data-driven models for speech enhancement [20], [21]. In this research, we further tighten the constraint for the speech to be estimated. We use a corpus consisting of complete speech utterances with little manipulation to provide examples of both short-time spectral shapes and up to sentence-long spectral variation for the speech to be extracted from noise and channel distortion. We show that the tightened constraint with long speech segments for the underlying speech could help to reduce the requirement for specific knowledge about the noise and channel, and could help to obtain an improved speech estimate.

This work extends our previous work [22], [23]. The extensions include: (1) addressing noisy speech with both additive noise and channel distortion, (2) an improved method for modeling noise, and (3) an iterative estimation algorithm for improved speech estimates. We demonstrate the improved performance for speech recognition and speech enhancement.

2. Assumptions and key idea

Let $X_{1:T} = \{x_t : t = 1, 2, ..., T\}$ be a wideband, clean speech signal, with $x_t$ being the logarithmic short-time power spectra (STPS) at time $t$. Consider a single-channel measurement of $X_{1:T}$ with both background noise and channel distortion. Let $Y_{1:T} = \{y_t : t = 1, 2, ..., T\}$ be the measured signal. In this study, we do not assume specific knowledge about the noise and channel. We only assume that the noise statistics and channel frequency characteristic change slower than the speech. This slowly-varying noise and channel assumption forms the basis of most current methods for speech enhancement and recognition. Specifically, we assume that real-world, slowly-varying noises can be approximated by piecewise stationary random processes. The noisy speech signal can be expressed as

$$y_t = \ln(e^{x_t} + e^{\eta_t})$$ (1)

where $h$ represents the log channel characteristic assuming it is fixed during the utterance, and $n_t$ represents the log STPS of the noise assuming it is piecewise stationary. Assume $n_t$ is subject to a Gaussian distribution. By piecewise stationarity we mean

$$n_t \sim N(iu, \Sigma_u) \text{ for } \epsilon \in [t, \tau]$$ (2)

That is, from $t$ the noise statistics (mean vector and covariance matrix) will remain invariant for a segment of consecutive frames from $t$ to $\tau$, as a function of $t$, while the speech statistics may change on a frame-by-frame basis. But the noise statistics can change across the segments to model nonstationary noise.

We propose a new approach for speech estimation based on the time-variation differences between the speech, noise, and channel, as assumed above. In our approach, we use a clean, wideband speech corpus to provide temporal-spectral examples of the speech to be extracted. We use a simplified example to illustrate our idea. Consider the power spectral density (PSD) as the statistics of a signal in the linear-spectral domain. Suppose Fig. 1 shows, on the top, the noisy signal PSD $y_{h,t}$ for a specific frequency bin $h$ sampled at consecutive frame times $t$, consisting of the clean signal PSD $x_{h,t}$ and some unknown noise PSD $n_{h,t}$. Below the noisy signal, Fig. 1 shows, on the left, a corpus of pre-recorded sample PSD $s_{h,t}$ of the clean signal $x_{h,t}$, and on the right, examples of stationary noise PSD of variable

This work was supported by the UK EPSRC grant EP/G001960.
noise levels used to model the piecewise stationary measurement noise assuming (2). The noise PSD can change from one level to another, on a segment-by-segment basis, to model globally nonstationary noise. For \( y_{k,t} \) at each \( t \), we aim to find a corpus sample and a noise candidate which, when added, match the given \( y_{k,t} \), thereby obtaining an estimate of the clean signal using the matched corpus sample. However, this won’t easily work if we focus on matching short measurements. As illustrated in the upper part of Fig. 1, given a single short-time noisy PSD measurement, there can be many different matched combinations between the corpus sample and the noise candidate. This explains why speech enhancement based on short measurements (e.g., single frames) requires specific knowledge about the noise for resolving the uncertainty. But, if we focus on matching longer segments of consecutive frames, and assume stationary noise (and hence a constant noise PSD) in the segment, then the number of possible matched combinations reduces (see the lower part of Fig. 1), subject to the nonnegative, constant noise PSD constraint. The longer the stationary noise segment and the matched combination found, the more specific the matched corpus sample segment and hence the signal estimate. This example can be extended to include a channel change in the measurement, which, in our assumption, only introduces a time-invariant gain change in each frequency bin in the corpus samples to form the match. Therefore, we propose the longest matching segment (LMS) approach: at each time \( t \), we find the longest noise segment from \( t \) that can assume stationary noise and has an accordingly matched corpus speech segment, subject to a constant channel factor. Since it is difficult to obtain accurate PSD estimates for nonstationary speech and noise, we implement the LMS approach for the log STPS features of the speech and noise within a statistical framework.

3. Longest matching segment approach

3.1. The posterior probability formulation

Assume we have a clean, wideband speech corpus. We build the enhancement system by first normalizing all the corpus speech utterances to a common gain. Let \( \Omega = \{ S_{\lambda,t} \} \) represent the corpus, consisting of gain-normalized sample speech utterances \( S_{\lambda,t} = \{ s_t : t = 1, 2, \ldots, \Gamma \} \). As in [22], we model the corpus \( \Omega \) by using a GMM, and model each sample utterance \( S_{\lambda,t} \) by using a Gaussian sequence \( \lambda_{S_{\lambda,t}} = \{ \lambda_{s_t} : t = 1, 2, \ldots, \Gamma \} \), where \( \lambda_{s_t} \sim \Sigma_{s_t} \) is a Gaussian taken from the corpus GMM that produces maximum likelihood for the frame \( s_t \). Given a noisy speech signal, we normalize its gain to the gain of the corpus speech data, such that in the signal the underlying speech signal’s gain may be smaller than the matched corpus speech signal’s gain due to the existence of noise, and due to channel distortion which can cause a loss of speech energy at certain frequency bands. To model the corpus utterance \( S_{\lambda,t} \) with a gain change, which is common to all the frequency bands, and a channel change, which can be different for different frequency bands, we use the model \( \lambda_{S_{\lambda,t}} = \{ \lambda_{s_t} : t = 1, 2, \ldots, \Gamma \} \), where \( \lambda_{s_t} \sim \Sigma_{s_t} \) is the Gaussian for the corpus frame \( s_t \) with a gain change \( g \) and a channel change \( h \), where \( I \) denotes a unit vector.

We model variable-level, piecewise stationary measurement noise by first generating a stationary zero-mean white noise with the same gain as the corpus speech data. We use \( \lambda_n = (\mu_n, \Sigma_n) \) to represent the statistics of this gain-normalized stationary white noise, and then use \( \lambda_{n,q} = (\mu_n + q, \Sigma_n) \) to represent the noise at a different gain level \( q \). Stationary colored noise can be simulated based on this stationary white noise model by allowing different gain levels in different frequency bands in the gain change vector \( q \).

Consider a statistical approach to compare the noisy speech segment \( Y_{l+r} \) and a corpus speech segment \( S_{\lambda,n} \). Assume stationary noise in \( Y_{l+r} \) with statistics \( \lambda_{n,q} \), and assume that compared to the corresponding corpus speech segment, the speech segment in \( Y_{l+r} \) is subject to a fixed gain change \( g \) and a fixed channel change \( h \). Thus, the likelihood of \( Y_{l+r} \) associated with \( S_{\lambda,n} \) can be written as

\[
p(Y_{l+r} | S_{\lambda,n}) = \sum_{g \in G, h \in H} \max_{\epsilon \in \mathbb{R}} \prod_{t=1}^{r} p(y_t | \lambda_{w(\epsilon)(t) + \epsilon g + h}, \lambda_{n,q})
\]

where we assume conditional independence between the frames in \( Y_{l+r} \) given \( S_{\lambda,n} \), \( p(y_t | \lambda_{w(\epsilon)(t) + \epsilon g + h}, \lambda_{n,q}) \) is the likelihood of the noisy frame \( y_t \) given (1) and the statistics of the speech (represented by the corpus model), gain, channel noise, and \( w(\epsilon) \) is a time warping function between \( Y_{l+r} \) and \( S_{\lambda,n} \). In our experiments, we use the log-normal approximation [24] to calculate \( p(y_t | \lambda_{w(\epsilon)(t) + \epsilon g + h}, \lambda_{n,q}) \), and use a simple linear \( w(\epsilon) = \epsilon \) to compare the two segments and compare only equal-length segments. The likelihood of the match between the two segments is decided through optimizing the parameters \( g, h \) and \( q \) on the segment level assuming stationary noise and constant channel in the segment. For the noisy utterance with the gain normalized to the corpus data, as described above, the inside speech gain \( g \leq 0 \) due to the existence of noise, \( g = 0 \) means there is no noise in \( Y_{l+r} \); given the speech gain \( g \), the maximum allowable noise gain in each frequency band can be approximately written as \( \ln(1 - e^g) \) for the noise model \( \lambda_n \) with the gain normalized to the corpus data, such that the noise power plus the speech power do not exceed the noisy utterance power. Colored noises are accounted for with the white noise model by selecting different noise gain levels in different frequency bands to match the given measurement. The negative channel characteristic in each frequency band represents the distortion of the wideband speech signal in that band caused by the channel effect; \( h = 0 \) means there is no channel distortion in the measurement. At each time \( t \), we aim to find the longest noisy segment \( Y_{l+r} \) (by extending \( r \)) that can assume stationary noise and has an accordingly matched corpus speech segment (Fig. 1), subject to a constant channel. We achieve this by maximizing the likelihood (3) among all the other likelihoods. This can be formulated as a maximum a posteriori problem described below.

We define the posterior probability of the match of a corpus speech segment \( S_{\lambda,n} \) given the noisy segment \( Y_{l+r} \), assuming
stationary noise and a fixed channel change in \( Y_{t:τ} \), as

\[
P(S_{ζ:η}|Y_{t:τ}) \approx p(Y_{t:τ}|S_{ζ:η}) \sum_{S^\dagger θ:ϑ ∈Ω} p(Y_{t:τ}|S^\dagger θ:ϑ)
\]

(4)

The denominator, the average likelihood \( p(Y_{t:τ}|S_{ζ:η}) \), is expressed as a sum of two terms. The first term is the average likelihood (3) of \( Y_{t:τ} \) over the corpus assuming that it contains stationary noise and has an accordingly matched corpus segment; the second term, \( p(Y_{t:τ}|φ) \), represents the average likelihood of \( Y_{t:τ} \) when the previous assumption does not hold, e.g., \( Y_{t:τ} \) is too long to find a matched corpus segment, or too long to be modeled by stationary noise, or both. We use the expression

\[
p(Y_{t:τ}|φ) = \max_{g ≤ b ≤ T} \left( \sum_{\sigma = \lambda_g ≤ τ} \sum_{q ≤ \lambda_q |S_{ζ:η}|} p(Y_{t:τ}|λ_g h_q, λ_q q) P(s) P(q) \right)
\]

(5)

In (5), for each frame \( y_t \) in \( Y_{t:τ} \), an average likelihood is calculated over all corpus speech frames and all different noise levels, to account for the unseen speech segment and/or nonstationary noise in \( Y_{t:τ} \). We assume uniform priors \( P(0) \) and \( P(q) \).

Using the method described in [22], [23], we can show that the posterior probability \( P(S_{ζ:η}|Y_{t:τ}) \) of a matched \( S_{ζ:η} \) increases when a longer noisy segment \( Y_{t:τ} \), with stationary noise, is matched. Therefore, we can obtain an estimate of the longest \( Y_{t:τ} \) from \( t \) with stationary noise and with a matched corpus segment by maximizing \( P(S_{ζ:η}|Y_{t:τ}) \) with respect to \( τ \) and \( S_{ζ:η} \). We have the estimates

\[
\hat{S}_{ζ(1):η(τ_{max})}, \hat{h}_t, \hat{q}_t = \arg \max_{τ, S_{ζ:η} ∈Ω} P(S_{ζ:η}|Y_{t:τ})
\]

(6)

where \( τ_{max} \) denotes the maximum \( τ \) found, and \( Y_{t:τ_{max}} \) corresponds to the longest noisy segment found from \( t \) which assumes stationary noise and has an accordingly matched corpus segment \( \hat{S}_{ζ(1):η(τ_{max})} \). Along with \( \hat{S}_{ζ(1):η(τ_{max})} \), we can also obtain the estimates of the corresponding channel \( \hat{h}_t \) and stationary noise statistics \( λ_{q(t)}, λ_q \) from (3) which form the longest segment match. We write these as a function of \( t \) to indicate they are associated with the longest matched noisy segment \( Y_{t:τ_{max}} \). We conduct the estimation (6) at every time \( t \). As each underlying speech frame \( x_t \) can be included in a number of matched corpus segments, we choose the matched corpus frame \( \hat{s}_t \) as an estimate of \( x_t \), \( \hat{s}_t \) being taken from the longest matched corpus segment with the longest left and right contexts about \( t \).

### 3.2. Smoothing the estimates and iteration

Since we assume the channel remains invariant during an utterance, we can obtain a smoothed channel estimate by averaging the individual segment-based estimates \( \hat{h}_t \) over the whole utterance. Let \( \tilde{h} \) denote the smoothed channel estimate, we take

\[
\tilde{h}_t = \frac{1}{P} \sum_{t=1}^{T} \hat{h}_t P(\hat{S}_{ζ(1):η(τ_{max})}|Y_{t:τ_{max}})
\]

(7)

where the posterior probability obtained in (6) is used as a confidence score for each segment-based estimate, and \( P \) is a normalization factor. A similar operation can be applied to the segment-based stationary noise estimates \( λ_{q(t)}, λ_q \). While we assume locally stationary noise in each matched segment, we assume that the noise statistics can change across the segments to model nonstationary noise. The noise estimates for the same noisy frame from different longest matched segments can be averaged to obtain a smoothed estimate for the nonstationary noise. Denote by \( λ_{n(t)}, λ_n \) the smoothed noise statistics estimate at time \( t \); we use the expression

\[
\tilde{λ}_n = \frac{1}{P} \sum_{t ≤ t_{max}} λ_{n(q)} P(\hat{S}_{ζ(1):η(τ_{max})}|Y_{t:τ_{max}})
\]

(8)

where the sum is taken over all longest matched noisy segments containing noise frame \( n \), and \( P \) is a normalization factor. Note that as an estimate of nonstationary noise, the smoothed noise estimate \( λ_n \) can change with time on a frame-by-frame basis.

After obtaining the smoothed channel and noise estimates, we can use the log-normal approximation [24] to combine these estimates with the clean speech corpus Gaussians, thereby obtaining noise and channel compensated corpus Gaussians. By replacing each clean corpus Gaussian \( λ_c \) in (3) and (5) with the corresponding channel and noise compensated Gaussian, we can rerun the LMS-based estimation (6) to obtain new estimates of the matched corpus speech segments. In the new search, the to-be-determined channel change and noise statistics in (3) and (5) model the residual channel change and noise in the noisy utterance as compared to the compensated speech corpus model, assuming that the residual channel change is fixed during the utterance and the residual noise is piecewise stationary. The above two processes of the LMS-based estimation and the formation of the noise and channel compensated corpus model based on the estimates can be alternated to form an iterative algorithm, which is implemented in our experiments.

In this paper, we consider applications of the LMS algorithm for speech recognition and speech enhancement. An advantage of the corpus-based system is that it can effectively connect, through the corpus data, the often separately implemented speech recognition and speech enhancement tasks, to achieve joint optimization for reducing the training and testing data mismatch. In our experiments for speech recognition, we build the enhanced speech features by directly taking the matched corpus speech features as the enhanced features; the same corpus speech features are also used to train the speech recognizer, thereby achieving a degree of matched condition training and testing. For speech enhancement, while Wiener filtering based on the matched corpus speech frame, as described in [22], can be used to suppress the additive noise, it is not effective for recovering speech from channel distortion. Therefore, we reconstruct the waveform for each underlying speech frame by using the magnitude spectrum of the wideband, matched corpus speech frame. A further advantage of corpus-based speech enhancement is that we have the option of using the phase spectra of the matched corpus speech signals to reconstruct the speech waveforms being estimated. Although the noisy signals phase spectra have proven to be usable for speech enhancement from noise, we have experienced poor performance on the Aurora 4 database when using the noisy signals phase spectra to reconstruct the speech with both noise and channel distortion. Therefore, in our experiments we take the phase spectra from the matched corpus speech frames for the reconstruction.

### 4. Experiments

Experiments were conducted on Aurora 4 [25], a database for 5k-word speech recognition with additive noise and combined additive noise and channel distortion. We built an HTK-based speech recognizer following [26]. Then, we used a subset of the WSJ0 training set (SI-TR-S), with 12776 utterances from
101 speakers recorded with a Sennheiser microphone, as the wideband clean speech corpus to build the LMS enhancement system. Aurora 4 has 14 test data sets, seven containing additive noise (clean, airport, babble, car, restaurant, street or train station at a randomly chosen SNR between 5 and 15 dB), and seven containing the same noise with further channel distortion (recorded with one of three other microphones) compared to the clean corpus data. In the LMS enhancement system, we divided the speech signals, sampled at 16 kHz, into frames of 20 ms with a period of 10 ms, and then represented each frame using the Mel-frequency log filterbank power spectrum with 50 channels. The corpus speech data was represented with a GMM of 4096 Gaussians with diagonal covariance matrices.

In the maximization in (3) and (5), we considered a range of segment-level speech gain losses to account for the noise and channel effects in the segment, from 0 dB (i.e., no gain loss) to −48 dB divided uniformly into 25 levels. From this range, we used the gain losses from 0 dB to −20 dB, with a total of eleven levels, to model $g \leq 0$ due to the existence of noise. Given a value of $g$ in this range, we estimated the corresponding noise gain for each frequency band subject to the constraint $q \leq \ln(1 - e^g)$. Finally, from each $g$, we further modeled the channel distortion $h \leq 0$ by selecting the channel characteristic for each frequency band from the current $g$ (i.e., no noise distortion) to $g = 28$ dB, giving fourteen further levels. We performed four iterations of the LMS-based estimation for the test utterances with only noise, and six iterations for the test utterances with both noise and channel distortion. We found that the iterations converged and always led to improved speech estimates in terms of improved speech recognition and speech enhancement performance compared to without iteration.

First, we used the LMS system as a preprocessor for feature extraction for speech recognition. Table 1 presents the word error rate (WER) produced by the HTK baseline recognition system when treating (a) the unprocessed noisy speech as input and (b) the enhanced speech features from the LMS enhancement system as input, respectively, averaged over the six noise conditions. To have an idea of how significant the improvement is, we include in Table 1 the WER published recently in the literature by some other systems performing speech recognition on Aurora 4. Table 1 shows that, among the selected recognition systems, using the LMS system to extract the acoustic features for speech recognition has raised the baseline recognition system performance from last position to around third position. We see no reason not to suppose that the application of the LMS-based preprocessing for feature extraction would also help improve the performance of the other recognition systems.

Next, we evaluated the LMS system for speech enhancement. We conducted experimental comparisons with other conventional speech enhancement algorithms. Since many of these conventional algorithms do not include a component for processing channel distortion, we compare with these algorithms only on the part of the Aurora 4 data without channel distortion. Fig. 2 shows the comparison of the PESQ scores between the LMS algorithm and four other enhancement algorithms, which we found produced better results among other algorithms. Two sets of scores are shown: one for the clean test data (A) and one for the noisy test data (B); for the latter, the scores are averaged over the six types of noise. As indicated in Fig. 2, for the clean speech test data, many of the conventional algorithms produced higher PESQ scores than LMS algorithm. This is because the LMS algorithm reconstructed the speech using different speech data from the corpus. However, for the noisy speech test data, the LMS algorithm performed rather better than all the other algorithms. Fig. 2 also shows the PESQ score improvement obtained by the LMS algorithm, for reconstructing speech from the test data with channel distortion (C) and with combined noise and channel distortion (D). Further evaluation of the LMS-based speech enhancement was conducted using the segmental SNR measure. Fig. 3 shows the comparison of the average segmental SNR ratings between the LMS algorithm and other conventional algorithms. Based on the comparisons, we conclude that the LMS algorithm gives similar improvements compared to other conventional enhancement algorithms.

## 5. Conclusions

This paper has focused on the modeling of the time variation differences between speech, noise and channel distortion for speech estimation. We described a novel corpus-based, iterative LMS approach for extracting speech signals from slowly-varying noise and channel distortion. The new approach was evaluated on the Aurora 4 database. The use of our enhancement approach as a preprocessor for feature extraction significantly improved the performance of a baseline recognition system. In another comparison for speech enhancement, both the PESQ and the segmental SNR ratings of the LMS algorithm were superior to the other enhancement methods.
6. References


