An iterative longest matching segment approach to speech enhancement with additive noise and channel distortion

Ji, M., & Crookes, D. (2014). An iterative longest matching segment approach to speech enhancement with additive noise and channel distortion. *Computer Speech & Language*, *28*(6), 1269-1286. Advance online publication. https://doi.org/10.1016/j.csl.2014.04.003

Published in:
Computer Speech & Language

Document Version:
Peer reviewed version

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

Publisher rights
Copyright 2014 Elsevier Ltd.
This manuscript version is made available under the Creative Commons Attribution-NonCommercial-NoDerivs License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits distribution and reproduction for non-commercial purposes, provided the author and source are cited.

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.

Open Access
This research has been made openly available by Queen's academics and its Open Research team. We would love to hear how access to this research benefits you. – Share your feedback with us: http://go.qub.ac.uk/oa-feedback

Download date:29. Feb. 2024

An Iterative Longest Matching Segment Approach to Speech Enhancement with Additive Noise and Channel Distortion

Ji Ming[1], Danny Crookes

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

Abstract

This paper presents a new approach to speech enhancement from single-channel measurements involving both noise and channel distortion (i.e., convolutional noise), and demonstrates its applications for robust speech recognition and for improving noisy speech quality. 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 for speech estimation. Third, we present an iterative algorithm which updates the noise and channel estimates of the corpus data model. 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.

Keywords: Corpus-based speech modeling, longest matching segment, noisy speech, channel distortion, speech enhancement, speech recognition

[1]Corresponding author. Email: j.ming@qub.ac.uk; Tel: +44 (0)28 90971705.

1. Introduction

This paper presents a new approach to speech enhancement from single-channel measurements involving noise, channel distortion (i.e., convolutional noise) and their combination, and demonstrates its application to improving speech recognition and to improving noisy speech quality. Modeling combined noise and channel distortion has been a major challenge in robust speech recognition. Research has been conducted in two main directions. The first is robust features; examples include speech enhancement (Couvreur and van Hamme, 2000; Deng et al., 2004; Logan and Robinson, 1997; Stouten et al., 2004), RASTA filtering (Hermansky and Morgan, 1994), feature normalization (De la Torre, 2005; Furui, 1981; Viikki and Laurila, 1998), SPLICE (Deng et al., 2000), and feature space adaptation (Li et al., 2002; Saon et al., 2001). The second research direction is robust acoustic models; examples include adaptive model training such as MAP (Gauvain and Lee, 1994), MLLR and CMLLR (Gales, 1998; Kim and Gales, 2011), predictive noise compensation such as parallel model combination (Gales and Young, 1995) and vector Taylor series compensation (Acero et al., 2000), joint uncertainty decoding (Liao and Gales, 2007) which combines feature transformation and model compensation, and missing-feature theory (Raj and Stern, 2005). Recent new developments include discriminative models (Ragni and Gales, 2011) and deep neural network based techniques (Seltzer et al., 2013). The work described in this paper is a complement to the robust feature approaches. We present a new approach to extracting clean speech features from single-channel measurements with both background noise and channel distortion. In extracting the features we focus on the reduction of the training and testing data mismatch which is critical to the success of speech recognition.

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 discrete Fourier transform (DFT) coefficients or spectral amplitudes (e.g., Cohen, 2005; Ephraim and Malah, 1984; Lotter and Vary, 2005; Martin, 2002; Martin and Breithaupt, 2003). 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 (e.g., Cohen, 2003; Hendriks et al., 2010; Lin et al., 2003; Martin, 2001; Rangachari and Loizou, 2006; Sohn and Kim, 1999). Some recent studies (e.g., Chinaev et al., 2012) have considered noise estimation based on some initial estimate of the speech power. 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 (e.g., Naidu and Srinivasan, 2012; Srinivasan et al., 2006), Gaussian mixture models (GMM) (e.g., Kundu et al., 2008), hidden Markov models (HMM) (e.g., Ephraim et al., 1989; Sameti and Deng, 2002; Zhao and Kleijn, 2007), and inventory-based models

which use prerecorded phonetic-class speech segments to restrict the enhanced signals (e.g., Nickel et al., 2012; Raj et al., 2011; Xiao et al., 2009). Some of the modeling techniques used in robust speech recognition have also found use in data-driven models for speech enhancement (e.g., Roux and Hershey, 2012; Seltzer et al., 2005). 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 in terms of improved speech recognition and speech enhancement performance.

This work is an extension of our previous work described in Ming et al. (2011, 2013). In Ming et al. (2011), we described a corpus-based approach for speech enhancement from additive noise. In Ming et al. (2013), we extended this approach to addressing the problem of separating two simultaneous speakers (i.e., speech separation). In this paper, we further extend this approach in three aspects. First, we extend the approach to single-channel speech enhancement with both additive noise and channel distortion (i.e., convolutional noise). Second, in Ming et al. (2011) we modeled unknown noise using a combination of multicondition model training and missing-data decoding; in this extended research we present an improved method to model noise for speech estimation, which shares some characteristics with the speech separation method described in Ming et al. (2013). Finally, we further extend the single-pass estimation algorithm to an iterative estimation algorithm; the new algorithm uses the previous corpus-based noise and channel estimates to update the corpus speech model for improved speech estimates. We demonstrate the improved performance for the new approach through experiments for speech recognition and speech enhancement.

The remainder of this paper is organized as follows. Section 2 outlines the assumptions made in this research for modeling noisy speech, and the key idea of the proposed approach for speech estimation. Section 3 introduces the first part of the proposed algorithm, including corpus-based modeling of speech with noise and channel distortion, and the longest matching segment algorithm for speech estimation. Section 4 describes a further development of the algorithm, including the refinement of the initial estimates and an iterative estimation algorithm for new, improved speech estimates. Experimental studies of the new approach for speech enhancement and as a preprocessor for feature extraction for speech recognition are described in Section 5. Finally, conclusions are presented in Section 6.

2. Assumptions and Key Idea

Let $\mathbf{X}_{1:T} = \{\mathbf{x}_t : t = 1, 2, ..., T\}$ represent a wideband, clean speech signal, expressed as the time series of the signal's logarithmic short-time power spectra (STPS) $\mathbf{x}_t$, where t is the discrete frame time. Consider a single-channel measurement of $\mathbf{X}_{1:T}$ in an adverse condition, with both

background noise and channel distortion. Let $\mathbf{Y}_{1:T} = \{\mathbf{y}_t : t = 1, 2, ..., T\}$ be the measured signal. In this study, we assume no 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 communication channel assumption forms the basis of most current methods for speech enhancement and speech recognition (for example, spectral subtraction, RASTA filtering, minimum-statistics based noise predication, cepstral feature normalization, model/feature space adaptation and prediction, etc.). In this paper, we describe a novel way of applying this assumption to speech estimation from noise and channel distortion. Specifically, we assume that real-world, slowly-varying noises can be approximated by piecewise stationary random processes. Assuming independence between the speech and noise, the noisy speech signal can be expressed as

$$\mathbf{y}_t = \ln(e^{\mathbf{x}_t+\mathbf{h}} + e^{\mathbf{n}_t}) \tag{1}$$

where we use $\mathbf{h}$ to represent the log channel characteristic assuming it is fixed during the utterance, and $\mathbf{n}_t$ to represent the log STPS of the noise assuming it is piecewise stationary. Assume that $\mathbf{n}_t$ is subject to a Gaussian distribution. By piecewise stationarity we mean

$$\mathbf{n}_\epsilon \sim N(\mu_{\mathbf{n}_t}, \Sigma_{\mathbf{n}_t}) \quad \text{for } \epsilon \in [t, \tau] \tag{2}$$

That is, from t the noise statistics (mean vector and covariance matrix) $\lambda_{\mathbf{n}_t} = (\mu_{\mathbf{n}_t}, \Sigma_{\mathbf{n}_t})$ will remain invariant for a segment of consecutive frames from time t to τ, as a function of t, while the speech statistics may change on a frame-by-frame basis. But $\lambda_{\mathbf{n}_t}$ can change across the segments to model nonstationary noise. Except for this local stationarity, we do not assume specific knowledge about the noise, i.e., the value of $\lambda_{\mathbf{n}_t}$ and the length of the measurement segment τ on which the noise can be assumed stationary. Nor do we assume specific knowledge about the channel characteristic $\mathbf{h}$.

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 assume that we have 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 $y_{k,t}$ for a specific frequency bin k sampled at consecutive discrete frame times t, consisting of the clean signal PSD $x_{k,t}$ and some unknown noise PSD $n_{k,t}$. Below the noisy signal, Fig. 1 shows, on the left, a corpus of pre-recorded sample PSD $s_{k,t}$ of the clean signal $x_{k,t}$, and on the right, examples of stationary noise PSD of variable noise levels used to model the piecewise stationary measurement noise assuming the noise model (2). As mentioned previously, 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 $y_{k,t}$. Knowing the make-up of this matched combination we can obtain 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 measurements, e.g., 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* noisy segment from t that can assume stationary noise and has an accordingly matched corpus speech segment, subject to a constant channel factor. As illustrated in the above example, if the noise and the channel change slower than the speech, and can be approximated with piecewise stationarity or invariance, this approach may lead to the estimates of the matched corpus speech segments with the least uncertainty. Since it is difficult to obtain accurate PSD estimates for nonstationary speech and noise, we implement the LMS approach for the log STPS features using the above assumed statistics for the noise, and using the corpus-based statistics, described below, for the speech. The following section details the modeling of the speech log STPS features and the basic LMS algorithm for speech estimation.

3. Longest Matching Segment (LMS) Approach

3.1. Modeling Speech, Noise and Channel Distortion

Assume that we have a clean, wideband speech corpus. We build the speech enhancement system by first normalizing all the corpus speech utterances to a common gain. Let $\Omega = \{\mathbf{S}_{1:\Gamma}\}$ represent the corpus, consisting of gain-normalized sample speech utterances $\mathbf{S}_{1:\Gamma} = \{\mathbf{s}_t : t = 1, 2, ..., \Gamma\}$, where $\mathbf{s}_t$ represents the log STPS of the speech frame at time t. As in Ming et al. (2011, 2013), we model the whole corpus Ω by using a GMM, and model each sample utterance $\mathbf{S}_{1:\Gamma}$ by using a corresponding Gaussian sequence $\lambda_{\mathbf{S}_{1:\Gamma}} = \{\lambda_{\mathbf{s}_t} : t = 1, 2, ..., \Gamma\}$, where $\lambda_{\mathbf{s}_t} = (\mu_{\mathbf{s}_t}, \Sigma_{\mathbf{s}_t})$ is a Gaussian taken from the corpus GMM that produces maximum likelihood for the frame $\mathbf{s}_t$. The corpus utterance model $\lambda_{\mathbf{S}_{1:\Gamma}}$ can be viewed as a template-based statistical model for speech; it captures all spectral temporal variations in $\mathbf{S}_{1:\Gamma}$, and yet it models each frame with a smoothed Gaussian distribution. We use such models of corpus speech utterances to provide temporal-spectral examples for the speech to be extracted.

Given a noisy speech signal, we normalize its gain to the gain of the corpus speech data. This gain normalization may be performed in two steps: (a) normalize the average gain of the noisy signal to that of the corpus data, and (b) detect the frequency bands in the normalized noisy

signal with a higher average gain than that of the corresponding frequency bands of the corpus data, if found further adjust the gain of the normalized noisy signal so that these frequency bands will have the same average gain as that of the corresponding corpus data. Therefore, in such a gain-normalized noisy 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 $\mathbf{S}_{1:\Gamma}$ 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 $\lambda_{\mathbf{S}_{1:\Gamma},g+\mathbf{h}} = \{\lambda_{\mathbf{s}_t,g+\mathbf{h}} : t = 1, 2, ..., \Gamma\}$, where $\lambda_{\mathbf{s}_t,g+\mathbf{h}} = (\mu_{\mathbf{s}_t} + g\mathbf{1} + \mathbf{h}, \Sigma_{\mathbf{s}_t})$ is the Gaussian for the corpus frame $\mathbf{s}_t$ with a gain change g and a channel change $\mathbf{h}$, where $\mathbf{1}$ denotes a unit vector.

We model the unknown, piecewise stationary measurement noise by first generating a stationary zero-mean white noise with the same gain as the corpus speech data. We obtain a model for this noise by estimating a Gaussian density (i.e., (2)) with a diagonal-covariance matrix for the log power spectrum of the simulated noise data. From this gain-normalized noise model, noise models at other gain levels can be obtained conveniently by adding for each level a corresponding gain change to the mean vector of the gain-normalized noise model. We use $\lambda_{\mathbf{n}} = (\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}})$ to represent the statistics of this gain-normalized stationary white noise, and $\lambda_{\mathbf{n},\mathbf{q}} = (\mu_{\mathbf{n}} + \mathbf{q}, \Sigma_{\mathbf{n}})$ to represent the noise with a gain change vector $\mathbf{q}$. By allowing different gain levels for different frequency bands in $\mathbf{q}$, the noise model $\lambda_{\mathbf{n},\mathbf{q}}$ is capable of simulating stationary colored noise based on the stationary white noise model. Additionally, as in most systems, in our experiments for each given noisy utterance, we collect some measurements from the beginning and end of the signal, which we assume does not contain speech, to obtain a Gaussian density estimate for the noise. This alternative noise model was used along with the white noise model as the noise candidates.

3.2. A Posterior Probability Formulation

Given a gain-normalized noisy utterance $\mathbf{Y}_{1:T}$, we use $\mathbf{Y}_{t:\tau} = \{\mathbf{y}_\epsilon : \epsilon = t, t+1, ..., \tau\}$ to represent a segment from time t consisting of consecutive frames from t to τ. In a similar way, we use $\lambda_{\mathbf{S}_{\zeta:\eta}} = \{\lambda_{\mathbf{s}_\epsilon} : \epsilon = \zeta, \zeta+1, ..., \eta\}$ to model a corpus speech segment $\mathbf{S}_{\zeta:\eta}$ taken from a corpus speech utterance $\mathbf{S}_{1:\Gamma}$ and consisting of the consecutive frames from time ζ to η, and use $\lambda_{\mathbf{S}_{\zeta:\eta},g+\mathbf{h}} = \{\lambda_{\mathbf{s}_\epsilon,g+\mathbf{h}} : \epsilon = \zeta, \zeta+1, ..., \eta\}$ to model the same corpus speech segment with a gain change g and a channel change $\mathbf{h}$. Assume that each log STPS vector consists of K frequency-band components. Let $y_{k,t}$, h_k and $s_{k,t}$ represent the k'th frequency-band component of the noisy measurement $\mathbf{y}_t$, channel characteristic $\mathbf{h}$ and corpus speech frame $\mathbf{s}_t$, respectively, and let $\lambda_{s_{k,t}} = (\mu_{s_{k,t}}, \Sigma_{s_{k,t}})$ represent the corresponding Gaussian statistics for $s_{k,t}$, $\lambda_{s_{k,t},g+h_k} = (\mu_{s_{k,t}} + g + h_k, \Sigma_{s_{k,t}})$ represent the corresponding Gaussian statistics modeling $s_{k,t}$ with a gain change g and a channel change h_k, and $\lambda_{n_k,q_k} = (\mu_{n_k} + q_k, \Sigma_{n_k})$ represent the k'th component of the statistics $\lambda_{\mathbf{n},\mathbf{q}}$ of the simulated noise, modeling the noise at the k'th frequency-band with a gain change q_k. In these expressions, we assume diagonal covariance matrices for both the speech and noise log STPS vectors.

Consider a statistical approach to compare the noisy segment $\mathbf{Y}_{t:\tau}$ and a corpus segment $\mathbf{S}_{\zeta:\eta}$. Assume stationary noise in $\mathbf{Y}_{t:\tau}$ and assume that, compared to the corresponding corpus speech segment, the speech segment in $\mathbf{Y}_{t:\tau}$ is subject to a fixed gain change and a fixed channel change, which we do not assume specific knowledge of. We write the likelihood function of $\mathbf{Y}_{t:\tau}$ associated with $\mathbf{S}_{\zeta:\eta}$ as

$$\begin{aligned} p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}_{\zeta:\eta}}) &\simeq \max_{g,\mathbf{h},\mathbf{q}} p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}_{\zeta:\eta}, g+\mathbf{h}}, \lambda_{\mathbf{n},\mathbf{q}}) \\ &= \max_{g\le 0} \prod_{k=1}^{K} \max_{h_k \le 0} \max_{q_k \le \ln(1-\exp(g))} \prod_{\epsilon=t}^{\tau} p(y_{k,\epsilon}|\lambda_{s_{k,w(\epsilon)}, g+h_k}, \lambda_{n_k,q_k}) \end{aligned} \tag{3}$$

In (3), we assume conditional independence between the frames and the frequency-band components in $\mathbf{Y}_{t:\tau}$ (conditioned on the segment $\mathbf{S}_{\zeta:\eta}$), and $p(y_{k,\epsilon}|\lambda_{s_{k,w(\epsilon)}, g+h_k}, \lambda_{n_k,q_k})$ is the likelihood of the noisy measurement $y_{k,\epsilon}$ given the measurement model (1) and the statistics of the speech (represented by the corpus model), gain, channel and noise. In our experiments presented in this paper, we use a linear time-warping function $w(\epsilon) = \zeta + \epsilon - t$ to compare the two segments and compare only equal-length segments (in our previous experiments with the LMS method for speech segment match, we have often found that using dynamic time warping (DTW) to calculate the match likelihood is insignificant in improving the match accuracy). For the noisy utterance with the gain normalized to the corpus data, as described above, the inside speech gain $g \le 0$ due to the existence of noise; $g = 0$ means there is no noise in $\mathbf{Y}_{t:\tau}$. Given a speech gain g, the maximum allowable noise gain can be approximately written as $\ln(1 - \exp(g))$ for the noise model $\lambda_{\mathbf{n}}$ also with the gain normalized to the corpus data, so that the noise power plus the speech power do not exceed the noisy utterance power; [2] colored noises are accounted for with the white noise model by selecting different noise gain levels q_k in different frequency bands to match the given measurement. The negative channel characteristic h_k in each frequency band represents the distortion of the wide-band speech signal in that band caused by the channel effect; $h_k = 0$ means there is no channel distortion in the frequency band. The likelihood of the match between the two segments $\mathbf{Y}_{t:\tau}$ and $\mathbf{S}_{\zeta:\eta}$ is decided through optimizing the parameters g, q_k and h_k on the segment level assuming stationary noise and constant channel characteristic in the segment. In other words, given a noisy segment $\mathbf{Y}_{t:\tau}$, $p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}_{\zeta:\eta}})$ indicates the likelihood of the noisy segment with stationary noise and with an accordingly matched corpus segment $\mathbf{S}_{\zeta:\eta}$, subject to a time-invariant channel factor. In our experiments, the maximization in (3) is performed by selecting the parameters g and h_k from a set of predefined statistics modeling a range of possible signal-to-noise ratios (SNRs) and channel distortion, with details given later. Given the speech and noise statistics, we use the log-normal approximation (Gales and Young, 1993) to calculate the likelihood $p(y_{k,\epsilon}|\lambda_{s_{k,w(\epsilon)}, g+h_k}, \lambda_{n_k,q_k})$ which

[2] For illustration, suppose $P^2_{\mathbf{y}}$, $P^2_{\mathbf{s}}$, and $P^2_{\mathbf{n}}$ represent the gain-normalized average power of the noisy measurement, corpus speech and noise, respectively. The gain normalization leads to $P^2_{\mathbf{y}} = P^2_{\mathbf{s}} = P^2_{\mathbf{n}}$. Therefore for additive noise we may assume that $P^2_{\mathbf{y}} \simeq GP^2_{\mathbf{s}} + (1-G)P^2_{\mathbf{n}} = \exp(g)P^2_{\mathbf{s}} + (1-\exp(g))P^2_{\mathbf{n}}$, where $g = \ln G$ is the logarithmic speech gain. Hence the corresponding logarithmic noise gain is approximately limited by $\ln(1-\exp(g))$.

takes the form of a Gaussian function (see Equations (11)–(16) in Section 4.1 for details of how to calculate this likelihood with the appropriate model parameters). At each time t, we aim to find the longest noisy segment $\mathbf{Y}_{t:\tau}$ (by extending τ) that can assume stationary noise and has an accordingly matched corpus segment (Fig. 1), subject to a time-invariant channel factor. We achieve this through maximizing the likelihood (3) among all the other likelihoods. This can be formulated as a maximum a posteriori (MAP) problem (i.e., Equation (8)). The following presents the details.

Assume an equal prior probability P for all possible speech segments. We define the posterior probability of the match of a corpus speech segment $\mathbf{S}_{\zeta:\eta}$ given the noisy segment $\mathbf{Y}_{t:\tau}$, assuming stationary noise and a fixed channel change in $\mathbf{Y}_{t:\tau}$, as

$$\begin{aligned} P(\lambda_{\mathbf{S}_{\zeta:\eta}}|\mathbf{Y}_{t:\tau}) &= \frac{p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}_{\zeta:\eta}})P}{p(\mathbf{Y}_{t:\tau})} \\ &\simeq \frac{p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}_{\zeta:\eta}})}{\sum_{\mathbf{S}^{\dagger}_{\theta:\vartheta}\in\Omega} p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}^{\dagger}_{\theta:\vartheta}}) + p(\mathbf{Y}_{t:\tau}|\phi)} \end{aligned} \tag{4}$$

where $p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}_{\zeta:\eta}})$ is the likelihood function defined in (3). The denominator, the average likelihood of the noisy segment $p(\mathbf{Y}_{t:\tau})$, is expressed as a sum of two terms. The first term is the average likelihood of the given noisy segment $\mathbf{Y}_{t:\tau}$ assuming that it contains stationary noise and has an accordingly matched corpus speech segment; the second term, $p(\mathbf{Y}_{t:\tau}|\phi)$, represents the average likelihood of $\mathbf{Y}_{t:\tau}$ when the previous assumption does not hold. The conditions that violate the assumption may include: $\mathbf{Y}_{t:\tau}$ is too long to find a matched speech segment in the corpus, or $\mathbf{Y}_{t:\tau}$ is too long to be modeled by stationary noise, or both. We use the following expression to model the likelihood of $\mathbf{Y}_{t:\tau}$ associated with unseen speech segments and/or nonstationary noise

$$p(\mathbf{Y}_{t:\tau}|\phi) = \max_{g\leq 0} \prod_{k=1}^{K} \max_{h_k\leq 0} \prod_{\epsilon=t}^{\tau} \sum_{s_{k,w(\epsilon)}\in\Omega} \sum_{q_k\leq \ln(1-\exp(g))} p(y_{k,\epsilon}|\lambda_{s_{k,w(\epsilon)},g+h_k,\lambda_{n_k,q_k}})P(s_{k,w(\epsilon)})P(q_k) \tag{5}$$

In (5), for each frame $\mathbf{y}_{\epsilon}$ in $\mathbf{Y}_{t:\tau}$, an average likelihood is calculated over all corpus speech frames and all different noise statistics, to account for the unseen speech segment and/or nonstationary noise in $\mathbf{Y}_{t:\tau}$, where the different noise statistics are simulated by the white noise statistics with variable gain levels in each frequency band, and $P(s_{k,w(\epsilon)})$ and $P(q_k)$ represent the prior probabilities of the individual speech frames and noise statistics, respectively. We can see the resemblance of (5) with a GMM used to model text-independent speech (the term "text" corresponds to the segmental dynamics of the speech and noise in discussion). In our experiments, we use uniform priors $P(s_{k,w(\epsilon)})$ and $P(q_k)$.

Noisy segments with mismatched corpus speech segments and/or with nonstationary noise are likely to result in low likelihoods of match defined by (3) but not necessarily low likelihoods of mismatch defined by (5), and hence are likely to result in low posterior probabilities of match based on (4). For the noisy segment $\mathbf{Y}_{t:\tau}$ which contains stationary noise and has an accordingly matched corpus speech segment $\hat{\mathbf{S}}_{\zeta:\eta}$, we can assume that the corresponding likelihood of match based on (3) is greater than the corresponding likelihood of mismatch based on (5) (explained below) and hence,

we will have a large posterior probability based on (4). Thus, the posterior probability (4) can be used to identify the matched stationary noise and corpus speech segment combination. Given a noisy segment, a large posterior probability will be obtained for a corpus speech segment if the noisy segment contains stationary noise and is matched accordingly by the corpus speech segment; a small posterior probability may indicate a mismatched corpus speech segment and/or nonstationary noise in the given noisy segment. The reason that we can assume $p(\mathbf{Y}_{t:\tau}|\phi) \leq p(\mathbf{Y}_{t:\tau}|\lambda_{\hat{\mathbf{S}}_{\zeta:\eta}})$, where $p(\mathbf{Y}_{t:\tau}|\lambda_{\hat{\mathbf{S}}_{\zeta:\eta}})$ represents the likelihood of match associated with the matched corpus segment $\hat{\mathbf{S}}_{\zeta:\eta}$ with a correspondingly matched stationary noise segment represented by the gain vector $\hat{\mathbf{q}}$, is that, based on (5), we can write $p(\mathbf{Y}_{t:\tau}|\phi)$ approximately as

$$\begin{aligned} p(\mathbf{Y}_{t:\tau}|\phi) &\simeq \max_{g\leq 0} \prod_{k=1}^{K} \max_{h_k \leq 0} \prod_{\epsilon=t}^{\tau} p(y_{k,\epsilon}|\lambda_{\hat{s}_{k,w(\epsilon)},g+h_k}, \lambda_{n_k,\hat{q}_k}) P(\hat{s}_{k,w(\epsilon)}) P(\hat{q}_k) \\ &\leq \max_{g\leq 0} \prod_{k=1}^{K} \max_{h_k \leq 0} \prod_{\epsilon=t}^{\tau} p(y_{k,\epsilon}|\lambda_{\hat{s}_{k,w(\epsilon)},g+h_k}, \lambda_{n_k,\hat{q}_k}) \\ &= p(\mathbf{Y}_{t:\tau}|\lambda_{\hat{\mathbf{S}}_{\zeta:\eta}}) \end{aligned} \tag{6}$$

The first approximation is based on the assumption that the matched and hence highly likely stationary noise and corpus speech segment combination dominates the mixture-based likelihood.

To locate the *longest* noisy segment $\mathbf{Y}_{t:\tau}$ with stationary noise and with a matched corpus speech segment, hence to remove as much of the uncertainty of the estimate of the matched corpus speech segment as possible as illustrated in Fig. 1, we point out an important property of the posterior probability (4): its value increases when a longer noisy segment, with stationary noise, is matched. Assume that the noisy segment $\mathbf{Y}_{t:\tau}$ with stationary noise is matched by the corpus speech segment $\hat{\mathbf{S}}_{\zeta:\eta}$, in the sense that the likelihood of match $p(\mathbf{Y}_{t:\tau}|\lambda_{\hat{\mathbf{S}}_{\zeta:\eta}}) \geq p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}^{\dagger}_{\theta:\vartheta}})$ for any $\mathbf{S}^{\dagger}_{\theta:\vartheta} \neq \hat{\mathbf{S}}_{\zeta:\eta}$, and $p(\mathbf{Y}_{t:\tau}|\lambda_{\hat{\mathbf{S}}_{\zeta:\eta}}) \geq p(\mathbf{Y}_{t:\tau}|\phi)$. Then we can have the following inequality concerning the posterior probabilities of the match of variable-length corpus segments and noisy segments with stationary noise

$$P(\lambda_{\hat{\mathbf{S}}_{\zeta:w(\epsilon)}}|\mathbf{Y}_{t:\epsilon}) \leq P(\lambda_{\hat{\mathbf{S}}_{\zeta:\eta}}|\mathbf{Y}_{t:\tau}) \tag{7}$$

where $\mathbf{Y}_{t:\epsilon}$ with $\epsilon \leq \tau$ is a noisy segment starting at the same time as $\mathbf{Y}_{t:\tau}$ but not lasting as long, and $\hat{\mathbf{S}}_{\zeta:w(\epsilon)}$ is the corresponding corpus subsegment matching the shorter noisy segment $\mathbf{Y}_{t:\epsilon}$. This inequality can be proved conveniently (see Ming et al., 2011, 2013). Based on (7), therefore, we can obtain an estimate of the longest noisy segment $\mathbf{Y}_{t:\tau}$ from t with stationary noise and with a matched corpus speech segment, through maximizing the posterior probability $P(\lambda_{\mathbf{S}_{\zeta:\eta}}|\mathbf{Y}_{t:\tau})$ with respect to τ and the corpus segment candidate $\mathbf{S}_{\zeta:\eta}$. We express the estimates as

$$\hat{\mathbf{S}}_{\zeta(t):\eta(\tau_{\max})}, \hat{\mathbf{h}}_t, \hat{\mathbf{q}}_t, \hat{g}_t = \arg\max_{\tau} \max_{\mathbf{S}_{\zeta:\eta}\in\Omega} P(\lambda_{\mathbf{S}_{\zeta:\eta}}|\mathbf{Y}_{t:\tau}) \tag{8}$$

where $\tau_{\max}$ denotes the maximum τ found, and $\mathbf{Y}_{t:\tau_{\max}}$ corresponds to the longest noisy segment found from t which can assume stationary noise and has an accordingly matched corpus segment

$\hat{\mathbf{S}}_{\zeta(t):\eta(\tau_{\max})}$, in terms of the maximum posterior probability. As indicated, this longest match is found by first finding for each fixed-length noisy segment $\mathbf{Y}_{t:\tau}$ the most-likely match, and then finding the $\mathbf{Y}_{t:\tau}$ with maximum length $\tau_{\max}$ that results in the maximum posterior probability. Along with the estimate of the matched corpus speech segment, we can also obtain the estimates of the corresponding channel characteristic $\hat{\mathbf{h}}_t$, stationary noise statistics $\lambda_{\mathbf{n},\hat{\mathbf{q}}_t}$ and gain of the matched corpus speech segment $\hat{g}_t$ from (3) which form the longest segment match. The following outlines the algorithm for solving the estimation problem (8):

For each test segment $\mathbf{Y}_{t:\tau}$ from t
  For each segment length τ
    Calculate $p(\mathbf{Y}_{t:\tau}|\phi)$ using (5)
    For each corpus segment $\mathbf{S}_{\zeta:\eta}$
      Calculate $p(\mathbf{Y}_{t:\tau}|\lambda_{\mathbf{S}_{\zeta:\eta}})$ using (3) and record the optimal parameters $\hat{\mathbf{h}}_t$, $\hat{\mathbf{q}}_t$ and $\hat{g}_t$
    For each corpus segment $\mathbf{S}_{\zeta:\eta}$
      Calculate posterior $P(\lambda_{\mathbf{S}_{\zeta:\eta}}|\mathbf{Y}_{t:\tau})$ using (4)
    Obtain the matched corpus segment and parameters with max $P(\lambda_{\mathbf{S}_{\zeta:\eta}}|\mathbf{Y}_{t:\tau})$ at given τ
  Obtain the longest matched corpus segment and parameters with max $P(\lambda_{\mathbf{S}_{\zeta:\eta}}|\mathbf{Y}_{t:\tau})$ over τ

In the above, we write $\hat{\mathbf{h}}_t$, $\lambda_{\mathbf{n},\hat{\mathbf{q}}_t}$ and $\hat{g}_t$ as a function of t to indicate that they are the estimates associated with the longest matched noisy segment $\mathbf{Y}_{t:\tau_{\max}}$ starting from time t. Given a noisy utterance, we conduct the estimation (8) at every frame time t. This provides the initial estimates of the matched corpus speech segments and of the corresponding channel characteristic and noise statistics for the whole utterance. In the following section, we extend the above single-pass LMS estimation algorithm to an iterative estimation algorithm, for obtaining an improved speech estimate.

4. LMS-Based Iterative Estimation and Reconstruction

4.1. Smoothing the Estimates and Iteration

Since we assume the channel frequency characteristic remains invariant during an utterance, we can obtain a smoothed channel estimate by averaging the individual segment-based estimates $\hat{\mathbf{h}}_t$ over the whole utterance; on average, each segment-based estimate is weighted by the corresponding segment posterior probability. Let $\tilde{\mathbf{h}}$ represent the smoothed channel estimate; we use the expression

$$\tilde{\mathbf{h}} = \frac{1}{\bar{P}} \sum_{t=1}^{T} \hat{\mathbf{h}}_t P(\lambda_{\hat{\mathbf{S}}_{\zeta(t):\eta(\tau_{\max})}}|\mathbf{Y}_{t:\tau_{\max}}) \tag{9}$$

where the posterior probability obtained in (8) is used as a confidence score for each segment-based channel estimate, and $\bar{P}$ is a normalization factor equalling the sum of the posterior probabilities across the utterance. In our experiments, we have found that the above smoothing operation is useful to correct the channel estimation biases, which may arise from those matched segments which have little speech content, or are short and hence have small posterior probabilities.

A similar operation can be applied to the segment-based stationary noise statistics estimates $\lambda_{\mathbf{n},\hat{\mathbf{q}}_t}$. This may lead to more accurate noise estimates, especially for the noise with nonstationary characteristics. 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 may each contain some information about the on-going nonstationary noise and can be averaged to obtain a smoothed noise estimate. Fig. 2 presents an example, showing the temporal power variation of a segment of restaurant noise (taken from the Aurora 4 database) during a nonstationary event, measured at the output of a mel-frequency filter centered at 1622 Hz. In Fig. 2, each horizontal straight-line segment corresponds to the mean estimate of the log power of the noise in an appropriate longest matched noisy segment based on (8); the dotted curve corresponds to a smoothed mean estimate obtained by averaging at each frame time the multiple noise mean estimates for the time from the different longest matched segments. In general, denoting by $\tilde{\lambda}_{\mathbf{n}_\epsilon} = (\tilde{\mu}_{\mathbf{n}_\epsilon}, \tilde{\Sigma}_{\mathbf{n}_\epsilon})$ the smoothed noise statistics estimate at time ϵ, we use the expression

$$\tilde{\lambda}_{\mathbf{n}_\epsilon} = \frac{1}{\tilde{P}} \sum_{t \text{ if } \epsilon \in [t, \tau_{\max}]} \lambda_{\mathbf{n},\hat{\mathbf{q}}_t} P(\lambda_{\hat{\mathbf{S}}_{\zeta(t):\eta(\tau_{\max})}} | \mathbf{Y}_{t:\tau_{\max}}) \tag{10}$$

where, as defined earlier in Section 3.1, $\lambda_{\mathbf{n},\hat{\mathbf{q}}_t} = (\mu_{\mathbf{n}} + \hat{\mathbf{q}}_t, \Sigma_{\mathbf{n}})$ with $\lambda_{\mathbf{n}} = (\mu_{\mathbf{n}}, \Sigma_{\mathbf{n}})$ being the gain-normalized white noise statistics, and the sum is taken over all longest matched noisy segments $\mathbf{Y}_{t:\tau_{\max}}$ that contain the noise frame $\mathbf{n}_\epsilon$, with $\tilde{P}$ being a normalization factor equalling the sum of the posterior probabilities associated with the segments included in the average. Note that as an estimate of nonstationary noise, the smoothed noise statistics estimate $\tilde{\lambda}_{\mathbf{n}_\epsilon}$ can change with time on a frame-by-frame basis, as shown in Fig. 2. The same expression (10) can be used to obtain a smoothed gain estimate for the speech frame in each noisy frame, by replacing $\lambda_{\mathbf{n},\hat{\mathbf{q}}_t}$ with the segment-based speech gain estimate $\hat{g}_t$. Let $\tilde{g}_\epsilon$ represent the smoothed estimate, for the gain of the speech frame in the noisy frame $\mathbf{y}_\epsilon$. This estimate will be used later for updating the corpus speech model.

After obtaining the smoothed channel, noise and speech gain estimates $\tilde{\mathbf{h}}$, $\tilde{\lambda}_{\mathbf{n}_\epsilon}$ and $\tilde{g}_\epsilon$, we consider how to incorporate them into the LMS system for a new search for the longest matched corpus speech segments, with the aim of improving the speech estimate. The smoothed channel and noise estimates can be used to modify the wideband, clean corpus speech model to reduce the mismatch against the noisy measurement, or used to reduce the level of distortion in the noisy measurement, thereby reducing the error in segment matching. In the following, we describe an algorithm to add compensation into the corpus speech model, and an iterative LMS algorithm for estimating the underlying speech.

As described, we model each corpus speech utterance by using a sequence of frame-based, maximum-likelihood Gaussians taken from the corpus GMM. By introducing channel and noise compensation into the corpus GMM, we therefore introduce the compensation into all the corpus

utterances built on the GMM used for finding the matched segments. Let $\{\lambda_m = (\mu_m, \Sigma_m) : m = 1, 2, ..., M\}$ represent the corpus GMM with M Gaussian density functions λ_m (note that earlier for clarity we have addressed a corpus Gaussian by using the corpus speech frame it models, e.g., $\lambda_{\mathbf{s}_t}$; but it should be understood that $\lambda_{\mathbf{s}_t} \in \{\lambda_m\}$). We use $\lambda_m(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon})$ to represent the modified λ_m which includes appropriate compensations for the channel distortion $\tilde{\mathbf{h}}$ and additive noise with statistics $\tilde{\lambda}_{\mathbf{n}_\epsilon}$. In the new search for the longest matching segments, we will model the corpus speech utterances/frames by replacing the clean Gaussians λ_m with the corresponding channel and noise compensated $\lambda_m(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon})$, for comparing with the noisy frame $\mathbf{y}_\epsilon$ for $\epsilon = 1, 2, ..., T$. Based on the log-normal approximation, the corrupted speech frames are approximately log normally distributed, and hence $\lambda_m(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon}) = (\mu_m(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon}), \Sigma_m(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon}))$, where $\mu_m(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon})$ and $\Sigma_m(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon})$ represent the mean vector and covariance matrix of the channel and noise compensated Gaussian, respectively. Assuming a diagonal covariance matrix, the k'th mean and variance element, denoted by $\mu_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})$ and $\Sigma_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})$, can be expressed as (Gales and Young, 1993)

$$\mu_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}}) = \ln(\bar{\mu}_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})) - \frac{1}{2}\ln(\frac{\bar{\Sigma}_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})}{\bar{\mu}_m^2(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})} + 1) \tag{11}$$

$$\Sigma_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}}) = \ln(\frac{\bar{\Sigma}_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})}{\bar{\mu}_m^2(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})} + 1) \tag{12}$$

where $\bar{\mu}_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})$ and $\bar{\Sigma}_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}})$ represent the linear-spectral domain statistics to form the channel and noise compensation for the clean Gaussian λ_m, which can be expressed as

$$\bar{\mu}_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}}) = e^{\mu_m + \Sigma_m/2 + \tilde{h}_k} + \bar{\mu}_{n_{k,\epsilon}} \tag{13}$$

$$\bar{\Sigma}_m(\tilde{h}_k, \tilde{\lambda}_{n_{k,\epsilon}}) = e^{2\mu_m + \Sigma_m + 2\tilde{h}_k}(e^{\Sigma_m} - 1) + \bar{\Sigma}_{n_{k,\epsilon}} \tag{14}$$

The statistics in (13) and (14) are each expressed in two terms: the first term shows the channel compensation $\tilde{h}_k$ and the second term shows the noise compensation $\bar{\mu}_{n_{k,\epsilon}}$ and $\bar{\Sigma}_{n_{k,\epsilon}}$ to the clean corpus Gaussian λ_m; the noise compensation is calculated from the noise estimate $\tilde{\lambda}_{n_{k,\epsilon}} = (\tilde{\mu}_{n_{k,\epsilon}}, \tilde{\Sigma}_{n_{k,\epsilon}})$ with the well-known relation

$$\bar{\mu}_{n_{k,\epsilon}} = e^{\tilde{\mu}_{n_{k,\epsilon}} + \tilde{\Sigma}_{n_{k,\epsilon}}/2} \tag{15}$$

$$\bar{\Sigma}_{n_{k,\epsilon}} = \bar{\mu}_{n_{k,\epsilon}}^2(e^{\tilde{\Sigma}_{n_{k,\epsilon}}} - 1) \tag{16}$$

After forming the new corpus GMM based on (11)–(12), we re-normalize the gain of the noisy utterance $\mathbf{Y}_{1:T}$, and the gain of the stationary white noise model $\lambda_{\mathbf{n}}$, to the gain of the new corpus model, as described in Section 3.1. Hence, by replacing each corpus speech frame Gaussian $\lambda_{\mathbf{s}_{w(\epsilon)}}$ in (3) and (5) with the corresponding channel and noise compensated Gaussian $\lambda_{\mathbf{s}_{w(\epsilon)}}(\tilde{\mathbf{h}}, \tilde{\lambda}_{\mathbf{n}_\epsilon})$, we can rerun the longest matching segment based estimation (8) to obtain new estimates of the matched corpus speech segments. In the new search, the to-be-determined channel change $\mathbf{h}$ and noise statistics $\lambda_{\mathbf{n},\mathbf{q}}$ in (3) and (5) model the *residual* channel change and additive noise in the noisy utterance as compared to the compensated corpus speech 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 longest matching segment based speech, channel and noise estimation, and the formation of the noise and channel compensated corpus speech model based on the smoothed channel and noise estimates, can be alternated to form an iterative algorithm. A new iteration starts with the update of the clean corpus Gaussians λ_m through (11) to (14) with the smoothed channel characteristic and noise statistics $\tilde{h}_k$, $\bar{\mu}_{n_{k,\epsilon}}$ and $\bar{\Sigma}_{n_{k,\epsilon}}$ accumulated over all previous iterations. The accumulated statistics for the i'th iteration, denoted by $\tilde{h}_k(i)$, $\bar{\mu}_{n_{k,\epsilon}}(i)$ and $\bar{\Sigma}_{n_{k,\epsilon}}(i)$, can be computed using recursion

$$\tilde{h}_k(i) = \tilde{h}_k(i-1) + \tilde{h}_k \tag{17}$$

$$\bar{\mu}_{n_{k,\epsilon}}(i) = \bar{\mu}_{n_{k,\epsilon}}(i-1) + \bar{\mu}_{n_{k,\epsilon}}/e^{\tilde{g}_\epsilon} \tag{18}$$

$$\bar{\Sigma}_{n_{k,\epsilon}}(i) = \bar{\Sigma}_{n_{k,\epsilon}}(i-1) + \bar{\Sigma}_{n_{k,\epsilon}}/e^{2\tilde{g}_\epsilon} \tag{19}$$

where $\tilde{h}_k$, $\bar{\mu}_{n_{k,\epsilon}}$ and $\bar{\Sigma}_{n_{k,\epsilon}}$ are the smoothed channel and noise estimates and $\tilde{g}_\epsilon$ is the smoothed corpus frame gain estimate, generated after the $(i-1)$'th iteration based on (9), (10), (15) and (16). The division of the noise estimates by the corpus speech gain estimate is needed to make the compensated corpus speech model have approximately the same SNR as the noisy utterance as indicated in the estimates. The complete iteration algorithm can be summarized as follows.

***Initialization*:** Set the iteration index $i = 0$; set the accumulated channel characteristic and noise statistics $\hat{h}_k(0)$, $\bar{\mu}_{n_{k,\epsilon}}(0)$, $\bar{\Sigma}_{n_{k,\epsilon}}(0)$ to zero.

***Step 1*:** Perform the LMS-based estimation (8). Stop, or go to Step 2 with $i = i + 1$.

***Step 2*:** Update the clean corpus GMM.

- Obtain smoothed channel, noise and corpus frame gain estimates based on (9), (10).
- Update the accumulated channel characteristic and noise statistics using the smoothed channel, noise and corpus frame gain estimates obtained above, based on (17)–(19).
- Update the clean corpus GMM using the accumulated channel characteristic and noise statistics obtained above, based on (11)–(14). Go to Step 1.

In our experiments, for each test utterance in each iteration, we calculate the average length of the longest matched segment found over all the test frames. We stop the iterations when there is no significant change in this average segment length between successive iterations. For more discussion see Section 5.3.

4.2. Reconstructing Speech Based on Segment Estimates

Based on the longest matched corpus speech segments found at each time t (i.e., (8)), there can be several ways to build the estimates of the underlying speech frames. In this paper, we consider two different applications of the above system: speech enhancement and feature extraction

for speech recognition. We use the Aurora 4 database in our experiments, and have found the following methods produce the best results.

As we estimate a matched corpus speech segment from each noisy frame, each underlying speech frame can be included in a number of adjacent matched corpus speech segments, each segment providing an estimate of the frame (Fig. 2 shows the same situation for the estimation of the underlying noise frames). We can obtain an estimate of the underlying speech frame at t by using the matched corpus speech frame chosen from the matched corpus speech segment that has the longest left and right contexts about t. We have considered other methods, including taking the average of the corresponding estimates from the different matched segments, but found that the estimates with the longest and most balanced left and right contexts demonstrate the desirable quality in terms of the individual frame sharpness and the cross-frame continuity (measured by several objective tests including speech recognition and PESQ, for example). Given the estimate of the matched corpus speech frame for each noisy frame, we can reconstruct the underlying clean speech frame by forming a Wiener filter as in (Ming et al., 2011). However, we found that this may not be the best method for speech recognition, because of the likely mismatch of the enhanced speech features against the training data. 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 (Ming et al., 2011), can be used to suppress the additive noise, it is not effective for recovering speech from channel distortion. Therefore, in our speech enhancement experiments with the Aurora 4 database, we reconstruct the waveform for each underlying speech frame by using the magnitude spectrum of the 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 waveforms of the speech being estimated. Although the noisy measurements phase spectra have proven to be usable for speech enhancement from noise, we have experienced poor performance on the Aurora 4 database for reconstructing the speech waveforms with the noisy measurements phase spectra with both noise and channel distortion. One possible reason is that some channel distortion (e.g., bandwidth reduction) can significantly reduce the speech energies in certain frequency bands and hence cause the phase spectra in these bands to become unusable or dominated by noise. In our experiments when this becomes a problem, we take the phase spectra from the matched corpus speech frames as an alternative. This was found to give better performance for speech enhancement.

5. Experimental Studies

5.1. Experimental Database and Systems

The experiments were conducted on the Aurora 4 database (Hirsch, 2002), which contains speech data with additive noise and combined additive noise and channel distortion. Aurora 4 is generated from the test data set of the WSJ0 database for a 5K-word speaker-independent speech recognition task. Table 1 summarizes the data used in our experiments. We built an HTK-based 5k-word speech recognition system following the HTK WSJ Training Recipe (Vertanen, 2006) using the training data shown in Table 1, and using a bigram language model, for speech recognition experiments. The recognition system used 13 static MFCC (mel-frequency cepstral coefficients) plus the first and second order derivatives as the feature vector for each frame. In a slight difference from the recipe system, in our system we dropped the zero'th cepstral coefficient (C0) to account for the variable gain changes of the reconstructed speech. Then, we used a subset of the WSJ0 training data set (SI-TR-S) as the wideband, clean speech corpus to build the proposed LMS enhancement system as a preprocessor for providing clean speech feature estimates for the recognition experiments, as well as for speech enhancement experiments. As mentioned earlier, for speech recognition, the enhanced speech features built on the training data reduces the training and testing data mismatch. We only considered the training and test speech data sampled at 16 kHz.

The corpus WSJ0 training set (SI-TR-S) we used to build the LMS speech enhancement system consists of 12776 utterances from 101 speakers (roughly balanced in gender) and was recorded with a Sennheiser microphone in quiet environments. In our LMS based speech enhancement experiments, for identifying matching speech segments, we divided speech signals into frames of 20 ms with a frame period of 10 ms, and then represented each frame using the Mel-frequency log filterbank power spectrum with 50 channels. We built the LMS enhancement system by first normalizing all the corpus utterances to a common gain, then using all the corpus utterances to train a GMM with 4096 Gaussian densities with diagonal covariance matrices, and finally obtaining a statistical model for each training utterance by associating each frame in the utterance with a Gaussian density chosen from the GMM which produces maximum likelihood for the frame, as described in Section 3.1.

Aurora 4 consists of 330 test speech utterances from eight speakers not included in the training/corpus data set. Each test utterance is recorded with a Sennheiser microphone (and hence contains no channel distortion compared to the corpus data), and also with one of three other microphones each introducing a different type of channel distortion compared to the corpus data. Aurora 4 is divided into two parts. The first part contains test data with additive noise only, which is formed by adding noise to the test utterances recorded with a Sennheiser microphone. Six different types of noise are used: car, babble, restaurant, street, airport and train station, each being added to the 330 test utterances at a randomly chosen SNR between 5 and 15 dB for each test utterance. This forms six test sets of noisy speech plus one test set without noise corruption for experiments, each test set containing 330 utterances.

The second part of Aurora 4 contains test data with both additive noise and channel distortion, compared to the clean corpus data recorded with a Sennheiser microphone. The test data are generated by adding the same types of noise, at a randomly chosen SNR between 5 and 15 dB, to the test utterances recorded with one of the three other microphones: a Shure SM91, a RadioShack Pro-Unidirectional Highball, and a AT&T 720 Handset. Like the first part of test data, the second part of test data includes six test sets with both noise and channel distortion, plus one test set without noise and with channel distortion only; each test set contains 330 utterances.

As described in Section 3.1, we simulated the piecewise stationary noise by generating stationary zero-mean white noise with the same gain as the corpus speech data. Additionally, for each given test utterance, we used the first and last 20 frames of the signal to obtain a Gaussian density estimate for the noise. This new noise model was used as an alternative to the white noise model - in calculating the likelihood of the measurement in (3) and (5), the noise model of the two which produced a larger likelihood would be used. Given a noisy test utterance, we normalized its gain to the gain of the corpus data. Taking the gain-normalized noisy utterances as input, 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. Based on our experiments, we found that modeling this range of gain losses was suitable for the Aurora 4 data with variable noise and channel distortions, and that the 25-level quantization offered a good balance between the modeling accuracy and the computational efficiency. From this range, we used the segment-level gain losses from 0 dB to -20 dB to model the speech gain loss due to the existence of noise in the segment (corresponding to a local, or segment-level, SNR from $+\inf$ to about -20 dB), and used the rest of the segment-level gain losses to model the channel distortion. This corresponds to the speech gain g in the maximization in (3) and (5) taking a value from $G = [0.0, -0.2, -0.4, ..., -2.0] \times \ln 10$, with a total of eleven levels. Given a speech gain $g = G[v]$ where v is the index of the gain value set G, the corresponding noise gain q_k in (3) and (5) for each frequency band takes a value from the set $[\ln(1 - \exp(G[v'])) : v' \leq v]$. This is subject to the constraint that the power of the model of speech plus noise should not exceed the power of the noisy measurement; the use of the speech gain resolution to quantize the noise gain range for the search reduces the amount of computation for (3) and (5).

From each g that models the noise-caused gain loss which applies to all speech frequency bands, we further modeled the gain loss in each frequency band caused by the channel distortion, by selecting the channel characteristic h_k for each frequency band, in (3) and (5), from the current g (i.e., no channel distortion) to $g - 28$ dB with a 2-dB resolution, giving fourteen further levels. As shown in the above, in the implementation of the LMS enhancement system for the experiments, we computed only 25-level gain changes in each frequency band for the corpus speech segment model, and the corresponding number of gain variations for the noise segment model, to model a wide range of unknown noise and channel distortions in the measurement. Also note that the above full-

range search for the gains of the matched corpus speech segments may only be needed in the initial estimation of the longest matching segments. In the subsequent iterations based on the previous estimates, we can reduce the search range accordingly to account for the reduced variations of the residual channel distortion and noise. This was implemented in our experiments, and caused no performance degradation. When dealing with the test speech without channel distortion, we set $h_k = 0$ for all frequency bands. When the longest matched corpus segments were found, the clean speech frames were reconstructed using the DFT magnitudes of the corresponding corpus speech frames.

5.2. Speech Recognition Results

First, we evaluated the proposed LMS enhancement system by performing speech recognition experiments. In these experiments, the LMS system was used as a preprocessor for clearing the noise and channel distortion from the input signals before passing them for recognition by the HTK baseline recognition system described above. Table 2 shows the word error rates (WER) produced by the HTK baseline recognition system when taking (a) the unprocessed noisy speech as input and (b) the reconstructed speech features from the LMS enhancement system as input, respectively. The effect of the channel distortion on the recognition accuracy can be clearly seen in Table 2, particularly for the "clean" speech recognition. For this wideband, clean speech trained baseline recognition system, the channel distortion alone had significantly increased the WER. We have studied the data, and found that some of the alternative microphones introduced not only spectral distortion, but also significant electrical noise, to the speech signal. As described earlier, we did not use the reconstructed waveforms from the LMS system to calculate the features for recognition (we found this produced poorer results, possibly due to the discontinuity of the adjacent frames which can cause some distortion in calculating the dynamic features for recognition). Instead, we took both the static and dynamic features directly from the matched corpus speech frames. This is found to be helpful in reducing the training and testing data mismatch.

In Table 3, we compare the recognition results obtained above with the results obtained by some of the other systems performing speech recognition on Aurora 4 published recently in the literature, to show the effect of the LMS enhancement system as a preprocessor for feature extraction for robust speech recognition. The results in Table 3 show that, among the selected recognition systems, using the LMS enhancement 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 applications of the LMS-based preprocessing for feature extraction would also help improve the performance of the other recognition systems.

In obtaining the above recognition results with the LMS system, we performed four iterations of the LMS-based estimation for each test utterance with noise only, and six iterations of the estimation for each test utterance with both noise and channel distortion, based on the iterative algorithm described in Section 4.1. In each iteration, a new corpus model was formed based on

the previous accumulated noise and channel estimates for a new search of the longest matching segments. In our experiments, 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. Fig. 3 shows the effect of the iterations on the speech recognition WER obtained on Aurora 4, averaged over the seven test conditions (six with noise, one noise free) with and without channel distortion. The iteration reduced the WER by absolute 4.7% for the test data without channel distortion (test conditions A and B combined), and by absolute 9.9% for the test data with channel distortion (test conditions C and D combined). After convergence, we did see further iterations might lead to random, but extremely small fluctuations in some performance measures. In our experiments, we have observed a strong correlation between the speech recognition accuracy and the enhanced speech ratings for the LMS algorithm. The results presented below for speech enhancement were produced based on the same number of iterations for each test utterance.

5.3. Speech Enhancement Results

Next, we evaluated the proposed LMS system for speech enhancement applications. Table 4 shows the PESQ scores for the unprocessed noisy speech and for the reconstructed speech from the LMS enhancement system. Again, we see that the channel distortion alone had significantly degraded the speech quality, in comparison to the original wideband clean speech. We conducted experimental comparisons with other conventional speech enhancement algorithms. Since many of these algorithms do not include a component for processing channel distortion, we compare with these conventional algorithms only on the part of the Aurora 4 test data without channel distortion. Fig. 4 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 and one for the noisy test data; for the latter, the scores are averaged over the six types of noise. As indicated in Fig. 4, 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. 4 also shows the PESQ scores for the reconstructed speech from the test data with channel distortion and with combined noise and channel distortion, obtained by the LMS algorithm compared to the PESQ scores for the unprocessed data. Further evaluation of the LMS-based speech enhancement performance was conducted using the objective measure segmental SNR, with the results presented in Table 5 and Fig. 5. Table 5 shows the detailed segmental SNR ratings obtained by the LMS algorithm for all the test conditions and Fig. 5 shows the comparison of the average segmental SNR ratings between the LMS algorithm and other conventional enhancement algorithms. Based on the comparisons, we can draw similar conclusions for the LMS algorithm in comparison to other conventional algorithms.

As mentioned earlier, the LMS enhancement system has the option to use the phase spectra of

the matched corpus speech data to reconstruct the underlying speech waveform. This contributed to the better PESQ and segmental SNR scores for the LMS-based reconstruction, for dealing with the test data with both noise and channel distortion, in comparison to the reconstruction with the noisy measurements phase spectra. For example, to reconstruct the speech based on the measurements with both noise and channel distortion, the use of the matched corpus speech phase spectra resulted in an average PESQ score 2.8, as shown in Fig. 4 Condition D. However, the reconstruction with the noisy measurements phase spectra only produced an average PESQ score 2.0, which is lower than the average PESQ score 2.2 for the unprocessed noisy speech. Similar observations were also obtained for the segmental SNR measure. For the measurements with both noise and channel distortion, the reconstruction with the matched corpus speech phase spectra resulted in an average segmental SNR of about 0.7 dB, as shown in Fig. 5 Condition D; but the reconstruction with the noisy measurements phase spectra only produced an average segmental SNR -3.9 dB.

Fig. 6 shows the histograms of the length of the longest matched segments found by the LMS algorithm as a function of the iteration index, with a total of six iterations performed for each test utterance with both noise and channel distortion (test condition D). In the initial iteration, the matched segments found are short because of the high noise level and potentially nonstationary noise content in the raw measurements, such that the measurement segments that can assume stationary noise could be short. The subsequent iterations each dealt with the residual noise from the previous estimation and compensation. We assume that the residual noise would have a reduced level than the initial noise to model, and that after the compensation for a nonstationary noise estimate (10), the residual noise could be more accurately modeled by a piecewise stationary noise model. Indeed, the algorithm was converging with the iteration by finding longer matched segments between the noisy measurements and the compensated corpus model, as shown in Fig. 6. For our test data, the histogram became largely stable after four iterations. In our experiments, we stopped the iterations for each test utterance if successive iterations produced longest matched segment estimates with similar average lengths. The mean length of the longest matched segments found after six iterations over all the test utterances is about nineteen frames.

Finally, we returned to the speech recognition experiments. We compared the proposed LMS algorithm with the other conventional speech enhancement algorithms as a preprocessor for generating enhanced speech features from noisy signals for speech recognition. As above, we conducted the experiments on the Aurora 4 data without channel distortion, and in the experiments we optimized the word insertion/deletion penalties for each individual enhancement algorithm. Table 6 shows the results. While the conventional enhancement algorithms produced significant improvement in the SNR (Fig. 5), they offered rather limited improvement in the recognition accuracy. A reason for this is the lack of joint optimization between the enhancement and recognition tasks, which creates the chance of mismatch between the training and test data for speech recognition.

6. Conclusions

This paper has focused on the modeling of the time variation differences between speech, noise and channel for speech estimation. We described a novel corpus-based, iterative LMS approach for extracting speech signals from slowly-varying noise and channel distortion. The corpus speech signal segments provide examples for the time-varying speech signals to be estimated; finding the longest matched noisy segments, subject to the constraint of stationary noise and invariant channel effect in the segments (i.e., a model of the slowly-varying noise and channel distortion), could lead to an estimate of the matched corpus speech segments with the least uncertainty. To further improve the estimation accuracy, the new approach uses iterations between the LMS-based estimation, and the estimation-based corpus model updating, to improve the modeling accuracy for the noise and channel distortion and thereby to derive an improved speech estimate. The new approach was evaluated on the Aurora 4 database for both speech recognition and speech enhancement experiments, with test data with combined additive noise and channel distortion. The use of our enhancement approach as a preprocessor for feature extraction significantly improved the performance of a baseline recognition system for dealing with noisy speech with additive noise, channel distortion, and their combination. 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.

References

Acero, A., Deng, L., Kristjansson, T.T., Zhang, J., 2000. HMM adaptation using vector Taylor series for noisy speech recognition. In Proc. ICSLP, 2000, pp. 869-872.

Alam, M.J., Kenny, P., O'Shaughnessy, D., 2013. Speech recognition using regularized minimum variance distortionless response spectrum estimation-based cepstral features. In Proc. ICASSP, 2013, pp. 8071-8075.

Chinaev, A., Krueger, A., Vu, D.H.T., Haeb-Umbach, R., 2012. Improved noise power spectral density tracking by a MAP-based postprocessor. In Proc. ICASSP, 2012, pp. 4041-4044.

Cohen, I., 2003, Noise spectrum estimation in adevrse environments: improved minima controlled recursive averaging. IEEE Transactions on Speech and Audio Processing, vol. 11, pp. 466-475, 2003.

Cohen, I., 2005. Speech enhancement using super-Gaussian speech models and noncausal a priori SNR estimation. Speech Communication, vol. 47, pp. 336-350, 2005.

Couvreur, C., van Hamme, H., 2000. Model-based feature enhancement for noisy speech recognition. In Proc. ICASSP, 2000, pp. 1719-1722.

De la Torre, A., Peinado, A.M., Segura, J.C., Perez-Cordoba, J.L., Benitez, M.C., Rubio,A.J., 2005. Histogram equalization of speech representation for robust speech recognition. IEEE Transactions on Speech and Audio Processing, vol. 13, pp. 355-366, 2005.

Deng, L., Acero, A., Plumpe, M., Huang, X.D., 2000. Large vocabulary speech recognition under adverse acoustic environments. In Proc. ICSLP, 2000, pp. 806-809.

Deng, L., Droppo, J., Acero, A., 2004. Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. IEEE Transactions on Speech and Audio Processing, vol. 12, pp. 133-143, 2004.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 32, pp. 1109-1121, 1984.

Ephraim, Y., Malah, D., Juang, B.H., 1989. On the application of hidden Markov models for enhancing noisy speech. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 37, pp. 1846-1856, 1989.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verifcation. IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 29, pp. 254-272, 1981.

Gales, M.J.F., Young, S.J., 1993. Cepstral parameter compensation for HMM recognition in noise. Speech Communication, vol. 12, pp. 231-239, 1993.

Gales, M.J.F., Young, S.J., 1995. Robust speech recognition in additive and convolutional noise using parallel model combination. Computer Speech and Language, vol. 9, pp. 289-307, 1995.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language, vol. 12, pp. 75-98, 1998.

Gauvain, J. L., Lee, C. H., 1994. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. IEEE Transactions on Speech and Audio Processing, vol. 2, pp. 291-298, 1994.

Gonzalez, J.A., Peinado, A.M., Gomez, A.M., Ma, N., Barker, J., 2012. Combining missing-data reconstruction and uncertainty decoding for robust speech recognition. In Proc. ICASSP, 2012, pp. 4693-4696.

Hendriks, R.C., Heusdens, R., Jensen, J., 2010. MMSE based noise PSD tracking with low complexity. In Proc. ICASSP, 2010, pp. 4266-4269.

Hermansky, H., Morgan, N., 1994. RASTA processing of speech. IEEE Transactions on Speech and Audio Processing, vol. 2, pp. 578-589, 1994.

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

Joshi, V., Bilgi, R., Umesh, S., Garcia, L., Benitez, C., 2012. Noise and speaker compensation in the log filter bank domain. In Proc. ICASSP, 2012, pp. 4709-4712.

Kim, D.K., Gales, M.J.F., 2011. Noisy constrained maximum-likelihood linear regression for noise-robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing, vol. 19, pp. 315-325, 2011.

Kundu, A., Chatterjee, S., Murthy, A.S., Sreenivas, T.V., 2008. GMM based Bayesian approach to speech enhancement in signal/transform domain. In Proc. ICASSP, 2008, pp. 4893-4896.

Li, Y., Erdogan, H., Gao, Y., Marcheret, E., 2002. Incremental online feature space MLLR adaptation for telephony speech recognition. In Proc. Interspeech, 2002.

Liao, H., Gales, M.J.F., 2007. Adaptive training with joint uncertainty decoding for robust recognition of noisy data. In Proc. ICASSP, 2007, pp. 389-392.

Lin, L., Holmes, W., Ambikairajah, E., 2003. Subband noise estimation for speech enhancement using a perceptual Wiener filter. In Proc. ICASSP, 2003, pp. 80-83.

Logan, B., Robinson, A.,1997. Enhancement and recognition of noisy speech within an autoregressive hidden Markov model frameowrk using estimates from the noisy signal. In Proc. ICASSP, 1997, pp. 843-846.

Lotter, T., Vary, P., 2005. Speech enhancement by MAP spectral amplitude estimation using a super-Gaussian speech model. EURASIP Journal on Applied Signal Processing, vol. 2005, pp. 1110-1126, 2005.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 504-512, 2001.

Martin, R., 2002. Speech enhancement using MMSE short time spectral estimation with Gamma distributed speech priors. In Proc. ICASSP, 2002, pp. 253-256.

Martin, R., Breithaupt, C., 2003. Speech enhancement in the DFT domain using Laplacian speech priors. In Proc. International Workshop on Acoustic Echo and Noise Control (IWAENC'2003), 2003, pp. 87-90.

Ming, J., Srinivasan, R., Crookes, D., 2011. A corpus-based approach to speech enhancement from nonstationary noise. IEEE Transactions on Audio, Speech and Language Processing, vol. 19, pp. 822-836, 2011.

Ming, J., Srinivasan, R., Crookes, D., Jafari, A., 2013. CLOSE – a data-driven approach to speech separation. IEEE Transactions on Audio, Speech and Language Processing, vol. 27, pp. 1355-1368, 2013.

Naidu, D.H.R., Srinivasan, S., 2012. A Bayesian framework for robust speech enhancement under varying contexts. In Proc. ICASSP, 2012, pp. 4557-4560.

Nickel, R.M., Astudillo, R.F., Kolossa, D., Zeiler, S., Martin, R., 2012. Inventory-style speech enhancement with uncertainty-of-observation techniques. In Proc. ICASSP, 2012, pp. 4645-4648.

Ragni, A., Gales, M.J.F., 2011. Structured discriminative models for noise robust continuous speech recognition. In Proc. ICASSP, 2011, pp. 4788-4791.

Raj, B., Stern, R., 2005. Missing-feature approaches in speech recognition. IEEE Signal Processing Magazine, vol. 22, pp. 101-116, 2005.

Raj, B., Singh, R., Virtanen, T., 2011. Phoneme-dependent NMF for speech enhancement in monaural mixtures. In Proc. Interspeech, 2011, pp. 1217-1220.

Rangachari, S., Loizou, P., 2006. A noise estimation algorithm for highly nonstationary environments. Speech Communication, vol. 28, pp. 220-231, 2006.

Roux, J.L., Hershey, J.R., 2012. Indirect model-based speech enhancement. In Proc. ICASSP, 2012, pp. 4045-4048.

Sameti, H., Deng, L., 2002. Nonstationary-state hidden Markov model representation of speech signals for speech enhancement. Signal Processing, vol. 82, pp. 205-227, 2002.

Saon, G., Zweig, G., Padmanabhan, M., 2001. Linear feature space projections for speaker adaptation. In Proc. ICASSP, 2001, pp. 325-328.

Seltzer, M.L., Acero, A., Droppo, J., 2005. Robust bandwidth extension of noise-corrupted narrowband speech. In Proc. Interspeech, 2005, pp. 1509-1512.

Seltzer, M.L., Yu, D., Wang, Y., 2013. An investigation of deep neural networks for noise robust speech recognition. In Proc. ICASSP, 2013, pp. 7398-7402.

Sohn, J., Kim, N., 1999. "Statistical model-based voice activity detection," IEEE Signal Processing Letters, vol. 6, pp. 1-3, 1999.

Srinivasan, S., Samuelsson, J., Kleijn, W.B., 2006. Codebook driven short-term predictor parameter estimation for speech enhancement. IEEE Transactions on Audio, Speech and Language Processing, vol. 14, pp. 163-176, 2006.

Stouten, V., Van Hamme, H., Wambacq, P., 2004. Joint removal of additive and convolutional noise with model-based feature enhancement. In Proc. ICASSP, 2004, pp. 949-952.

Vertanen, K., 2006. Baseline WSJ acoustic models for HTK and Sphinx: training recipes and recognition experiments. Tech. Rep., Cavendish Laboratory, 2006.

Viikki. O., Laurila, K., 1998. Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Communication, vol. 25, pp. 133-147, 1998.

Wand, Y.Q., Gales, M.J.F., 2011. Speaker and noise factorisation on the Aurora 4 task. In Proc. ICASSP, 2011, pp. 4584-4587.

Xiao, X., Lee, P., Nickel, R.M., 2009. Inventory based speech enhancement for speaker dedicated speech communication systems. In Proc. ICASSP, 2009, pp. 3877-3880.

Xiao, X., Li, J., Chng, E.S., Li, H., 2011. Maximum likelihood adaptation of histogram equalization with constraint for robust speech recognition. In Proc. ICASSP, 2011, pp. 5480-5483.

Zhao, D.Y., Kleijn, W.B., 2007. HMM-based gain modeling for enhancement of speech in noise. IEEE Transactions on Audio, Speech and Language Processing, vol. 15, pp. 882-892, 2007.

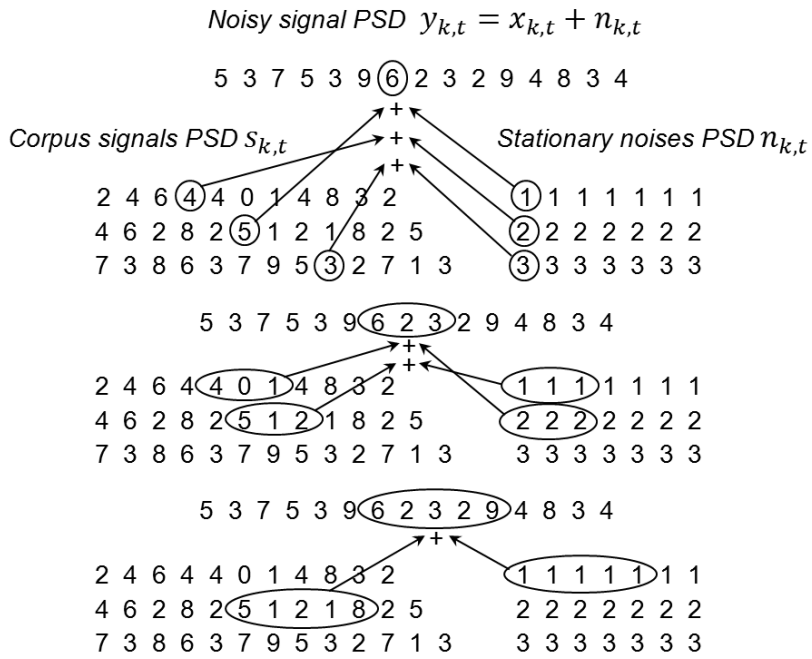

Figure 1: Illustration of the proposed longest matching segment (LMS) approach. Assume that shown on the top of each section is a noisy signal power spectral density (PSD) sequence for a specific frequency bin k sampled at consecutive discrete frame times t. The bottom of each section shows the combination of some corpus signal PSD segment and stationary noise PSD segment to match a noisy signal PSD segment assuming stationary measurement noise in the segment. The longer the matched segments found, the more specific the matched corpus signal, subject to the nonnegative, constant noise PSD constraint.

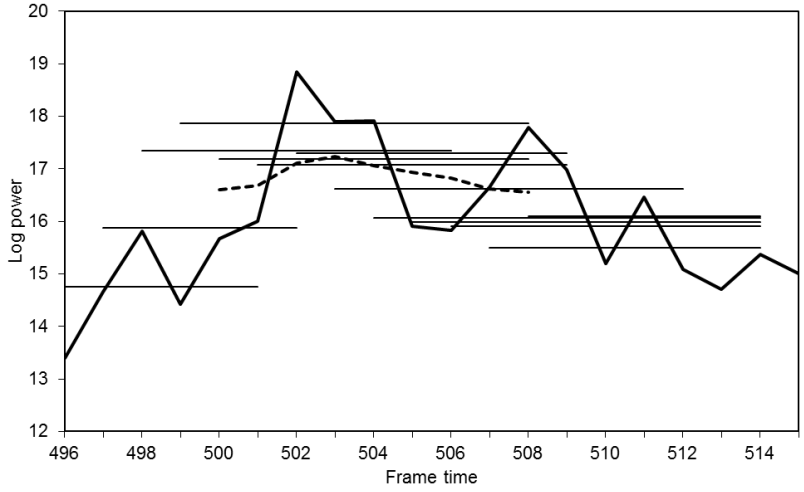

Figure 2: A smoothed mean estimate (dashed curve) of the noise power variation (solid curve) by averaging the segment-based stationary noise mean estimates (horizontal straight lines).

Table 1: Summary of the databases used in the experiments.

Training data used to build the baseline speech recognizer	The corpus used to build the proposed LMS system	Test data
Full set of WSJ0, WSJ1 training data, TIMIT-bootstrapped monophones	Subset of WSJ0 training set (SI-TR-S)	Aurora 4

Table 2: Aurora 4 word error rates (WER) for the unprocessed noisy speech and the reconstructed speech features from the LMS enhancement system, based on a HTK baseline speech recognition system trained using wideband, clean training data, for different test noise conditions with and without channel distortion.

Input	Channel distortion	Clean	Airport	Babble	Car	Restaurant	Street	Train
Unprocessed	No	9.4	53.2	57.7	37.3	49.8	49.5	54.3
	Yes	48.5	67.6	67.1	60.6	64.4	69.6	68.5
LMS output	No	10.3	20.5	22.5	11.7	24.6	20.3	23.2
	Yes	15.6	22.6	24.5	22.4	32.9	30.8	29.6

Table 3: Comparison of WER on Aurora 4 between some existing recognition systems and a baseline recognition system with LMS-based preprocessing, for test conditions: A – clean, B – with noise, C – with channel distortion, and D – with both noise and channel distortion.

Technique	A	B	C	D	Avg
DNN+NAT (Seltzer et al. 2013)	5.4	8.3	7.6	18.5	12.4
Joint+MLLR (Wand and Gales, 2011)	5.0	11.5	8.1	19.1	14.1
MBFE (Stouten et al., 2004)	4.9	20.8	19.2	37.7	26.8
Missing data+UD (Gonzalez et al., 2012)	12.6	35.7	27.4	47.4	38.5
VTS+TVTLN (Joshi et al., 2012)	9.4	27.5	13.7	32.9	27.5
Regularized MVDR (Alam et al., 2013)	9.9	45.8	21.8	59.4	47.3
HEQ-ML (Xiao et al., 2011)	12.6	31.6	19.8	41.1	33.4
HTK baseline	9.4	50.3	48.5	66.3	54.1
HTK baseline with LMS preprocessing	10.3	20.5	15.6	27.1	22.3

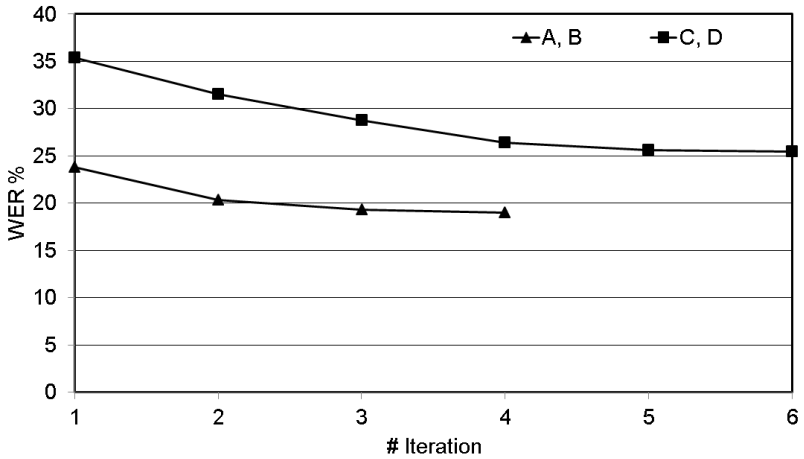

Figure 3: Average WER decreases with the iteration, based on the proposed iterative LMS algorithm, for the test data without channel distortion (test conditions A and B combined), and with channel distortion (test conditions C and D combined).

Table 4: PESQ scores for the unprocessed noisy speech and for the reconstructed speech from the LMS enhancement system, for different noise conditions with and without channel distortion.

Input	Channel distortion	Clean	Airport	Babble	Car	Restaurant	Street	Train
Unprocessed	No	4.5	2.4	2.3	2.6	2.3	2.2	2.2
	Yes	3.0	2.3	2.2	2.5	2.2	2.1	2.1
LMS	No	4.2	3.1	2.9	3.7	2.8	3.0	2.9
output	Yes	3.4	2.9	2.9	3.2	2.6	2.8	2.6

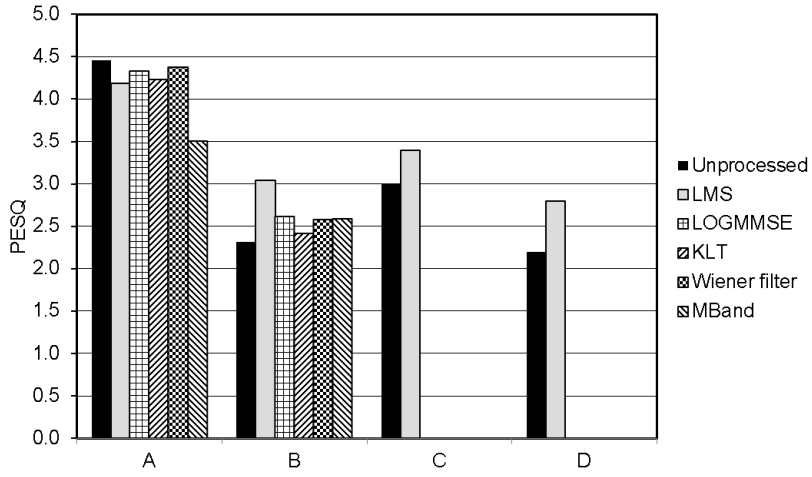

Figure 4: PESQ scores for the reconstructed speech by the LMS enhancement method compared to other speech enhancement methods, for test conditions: A – clean, B – with noise, C – with channel distortion, and D – with both noise and channel distortion.

Table 5: Segmental SNR ratings (dB) for the unprocessed noisy speech and for the reconstructed speech from the LMS enhancement system, for different noise conditions with and without channel distortion.

Input	Channel distortion	Clean	Airport	Babble	Car	Restaurant	Street	Train
Unprocessed	No	10.1	-0.9	-1.5	-5.8	-0.8	-3.3	-3.1
	Yes	-5.5	-6.5	-6.5	-7.7	-6.4	-6.9	-6.9
LMS	No	8.2	1.5	0.9	1.6	1.1	0.7	0.4
output	Yes	5.8	0.8	0.5	1.5	0.6	0.5	0.3

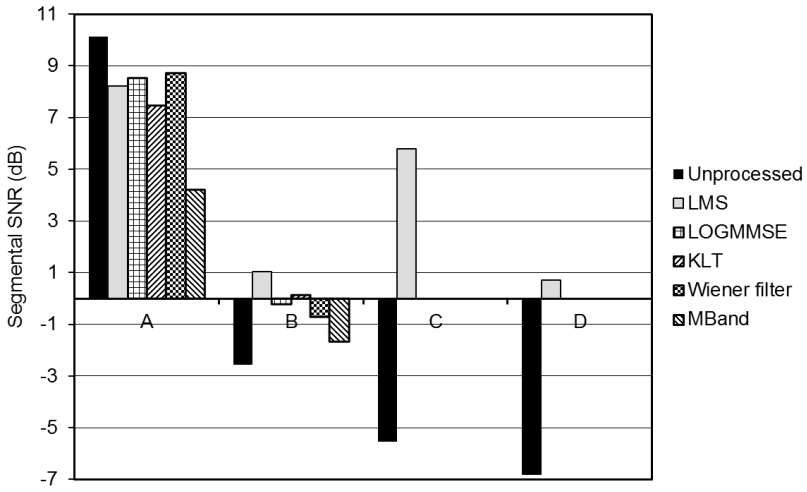

Figure 5: Segmental SNR ratings for the reconstructed speech by the LMS enhancement method compared to other speech enhancement methods, for test conditions: A – clean, B – with noise, C – with channel distortion, and D – with both noise and channel distortion.

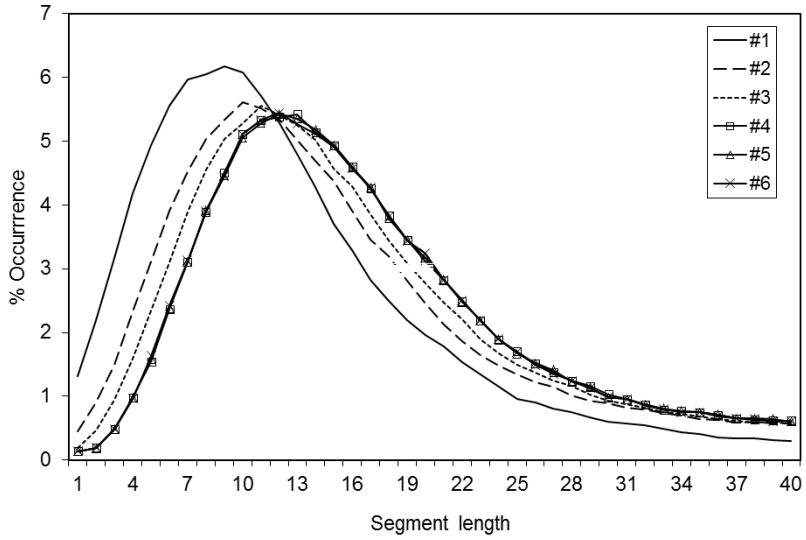

Figure 6: Histogram of the length (in number of frames) of the longest matched segments found by the LMS algorithm as a function of the iteration index, for the test speech with combined noise and channel distortion.

Table 6: WER of the baseline recognition system for the enhanced speech, comparing the LMS algorithm with other speech enhancement algorithms, for the Aurora 4 data with noise (test condition B).

Input	No processing	LMS	LOGMMSE	KLT	Wiener filtering	MBand
Clean	9.4	10.3	12.1	13.5	12.5	15.0
With noise	50.3	20.5	46.2	44.6	49.8	43.2