ABSTRACT
In this paper, we propose a robust speech-recognition approach that utilizes an auxiliary laser Doppler vibrometer (LDV) sensor. The LDV-measured signal is used for enhancing the noisy acoustic signals before feeding the automatic speech recognition (ASR) engine. The enhancement algorithm includes a time-frequency voice activity detector (VAD), which is derived from the LDV signal by using a two-stage algorithm. The first stage consists of a correlation-based rough detector, and the second stage introduces a robust harmonic-structure tracking. Both noise robustness and improved speech intelligibility are attained by the proposed enhancement algorithm. ASR experimental results demonstrate a substantial improvement in recognition-accuracy performance in parked- and moving-vehicle environments under low signal-to-noise ratio conditions.

Index Terms— speech recognition, speech enhancement, nonacoustic sensors, laser vibrometry.

1. INTRODUCTION
Achieving high recognition accuracy in noisy environments is one of the most challenging and important problems for existing automatic speech recognition (ASR) systems. Under relatively low signal-to-noise ratio (SNR) conditions and highly non-stationary noise environments, the perceived speech quality is severely degraded. This may cause a mismatch between the received signal and the ASR training data and worsen the recognition-accuracy performance.

Recently, several approaches for improved speech recognition have been proposed, which make use of auxiliary nonacoustic sensors, such as bone- and throat- microphones (e.g., [1–5]). Such sensors typically measure vibrations of the speech-production anatomy (e.g., vocal-fold vibrations) and are relatively immune to acoustic interferences [1]. Speech intelligibility can then be improved by combining the the acoustic noisy signal with the speech information captured by these auxiliary sensors. In [2], air- and throat- microphones are combined by training features mapping from both sensors to improve noise robustness of ASR systems. In [3], a voice activity detector (VAD) is constructed from a throat sensor to improve speech recognition accuracy, and the general electromagnetic motion sensor (GEMS) is utilized in [5] for speech coding. A major drawback of most existing sensors is the requirement for a physical contact between the sensor and the speaker. Contact-based auxiliary sensors must be strapped or taped on facial locations to measure speech vibrations.

Alternatively, the use of an auxiliary non-contact laser Doppler vibrometer (LDV) sensor has been recently introduced for improved speech enhancement [6]. When focusing on the larynx, this sensor captures useful speech information at low-frequency regions (up to $1.5 - 2$ kHz), and is shown to be isolated from acoustical disturbances. The algorithm proposed in [6], however, often fails to retain weak-speech components and may severely degrade speech quality, especially for increased energy of impulsive (speckle) noises, which often degrade the LDV signal. As such, it cannot be efficiently used for speech recognition.

In this paper, we propose a robust speech-recognition system that utilizes remote speech measurements from an auxiliary LDV sensor. These measurements are used for enhancing the noisy acoustic signals before feeding the ASR engine. The LDV-measured signal is first used to derive an accurate time-frequency VAD with a two-stage algorithm, consisting of a correlation-based rough detector followed by a robust harmonic-tracking algorithm. Contrary to [6], the proposed VAD does not attempt to reduce the speckle noises, but rather ignore them by detecting spectral harmonic patterns. The resulting VAD is then incorporated into the optimally-modified log-spectral amplitude (OM-LSA) algorithm [7] to further enhance its performance under low SNR conditions. Both noise robustness and improved speech intelligibility are attained by the proposed enhancement algorithm. An ASR experiment, including parked- and moving-vehicle scenarios in low SNR conditions, is conducted. The results demonstrate the effectiveness of the proposed approach in substantially improving
acoustic speech signal [Fig. 2(a)] and an LDV-measured signal [Fig. 2(b)], as recorded in a moving-car environment with a sampling rate of 16 kHz. The acoustic sensor corresponds to a laptop (T60) located 40 cm from the speaker; whereas the LDV sensor is located 1 m from the speaker. Clearly, the LDV signal is immune to acoustic interferences. On the other hand, it captures useful speech information only at low-frequency regions (up to 1 kHz). We further observe that the measured laser signal is degraded by an interference, characterized by random impulses. This impulse-like noise is generally referred to as speckle noise [10] and arises from random constructive and destructive interferences of waves that backscatter from a relatively rough surface.

3. TIME-FREQUENCY VAD DERIVATION

In this section, we exploit the immunity of the LDV sensor to acoustic disturbances in order to derive a reliable VAD in the time-frequency domain. We propose a two-stage algorithm. The first stage consists of a correlation-based rough detector, and the second stage introduces a robust harmonic-tracking algorithm.

Let $Z(k, \ell)$ denote the short-time Fourier transform (STFT) of the LDV signal $z(n)$, where $\ell = 0, 1, \ldots$ is the frame index and $k = 0, 1, \ldots, N - 1$ is the frequency-bin index. We use overlapping frames of $N$ samples with a framing-step of $M$ samples. Let $z(\ell) = \begin{bmatrix} Z(k_1, \ell) & Z(k_1 + 1, \ell) & \cdots & Z(k_2, \ell) \end{bmatrix}^T$, where $k_1$ and $k_2$ define the frequency range that contains useful speech information in the LDV signal. We define the (normalized) correlation between consecutive speech frames as

$$\rho(\ell) = \frac{z^H(\ell)z(\ell - 1)}{\|z(\ell)\|^2 \|z(\ell - 1)\|^2}. \quad (1)$$

To decrease estimation variance, $\rho(\ell)$ is smoothed by a first-order recursive averaging

$$\hat{\rho}(\ell) = \alpha_\rho \hat{\rho}(\ell - 1) + (1 - \alpha_\rho) \rho(\ell), \quad (2)$$

where $\alpha_\rho (0 < \alpha_\rho < 1)$ denotes a smoothing parameter. Then, motivated by the relatively-high correlation between speech frames, we define the following rough decision about speech presence

$$I_1(\ell) = \begin{cases} 1, & \text{if } \hat{\rho}(\ell) \geq T \quad \text{(speech is present)} \\ 0, & \text{otherwise} \quad \text{(speech is absent)} \end{cases} \quad (3)$$

where the threshold $T$ is set to satisfy a certain false-alarm probability $P(\hat{\rho}(\ell) \geq T | H_0(\ell)) = \epsilon$, and $H_0(\ell)$ indicates speech absence hypothesis. This probability can be numerically evaluated by assuming the background noise in the mea-
In the second stage, the rough decision of the detector in (3) is smoothed by detecting harmonic locations. Specifically, let \( \hat{\lambda}_{\text{global}}(k, \ell) \) denote the global background-noise spectrum estimate, derived by using the improved minima-controlled recursive averaging (IMCRA) algorithm [11], and let \( \hat{\lambda}_{\text{local}}(\ell) \) denote an estimator for the instantaneous noise spectrum in the \( \ell \)-th frame. The latter can be derived by discarding the low frequency region (due to possible speech information)

\[
\hat{\lambda}_{\text{local}}(\ell) = \frac{1}{N - k_2 - 1} \sum_{k = k_2 + 1}^{N-1} |Z(k, \ell)|^2.
\]

(4)

Accordingly, global and local SNRs are defined, respectively, by \( \gamma_g(k, \ell) \triangleq \frac{|Z(k, \ell)|^2}{\hat{\lambda}_{\text{global}}(k, \ell)} \) and \( \gamma_l(k, \ell) \triangleq \frac{|Z(k, \ell)|^2}{\hat{\lambda}_{\text{local}}(\ell)} \). Then, we define the following indicator for speech presence

\[
I_1(k, \ell) = \begin{cases} 
1, & \text{if } I_1(\ell) = 1, \gamma_g(k, \ell) \geq T_g \\
0, & \text{otherwise}
\end{cases}
\]

(5)

where \( k_1 \leq k \leq k_2 \) and the thresholds \( T_g \) and \( T_l \) are set to satisfy a certain false-alarm probability. Under a Gaussian assumption, we typically use \( T_g = T_l = 9.2 \). The high thresholds are attributable to the relatively high SNR associated with the LDV signal. We note that the local-SNR condition is introduced here in order reduce false spectral detections due to bursts of high-energy speckle noise. Moreover, harmonic false-detections can be further reduced by excluding those high-energy spectral components that are not local maxima. That is, a positive VAD decision \( I_1(k_0, \ell_0) = 1 \) is updated and set to 0 if

\[
|Z(k_0, \ell_0)|^2 < \max \left\{ |Z(k_0 - 1, \ell_0)|^2, |Z(k_0 + 1, \ell_0)|^2 \right\}.
\]

(6)

Finally, we exploit the correlation between speech frames to entail continuous positive VAD decisions. Specifically, let \( c(k, \ell) \) represent the number of positive VAD decisions at an \((2K_c + 1 \times 2L_c + 1)\)-vicinity of \((k, \ell)\), i.e.,

\[
c(k, \ell) = \sum_{k' = k-K_c}^{k+K_c} \sum_{\ell' = \ell-L_c}^{\ell+L_c} I_1(k', \ell').
\]

(7)

Then, we propose the following decision for speech harmonic locations

\[
I(k, \ell) = \begin{cases} 
1, & \text{if } I_1(k, \ell) = 1 \text{ and } c(k, \ell) \geq N_c \\
0, & \text{otherwise}
\end{cases}
\]

(8)

Based on the resulting decision \( I(k, \ell) \), we define a frame-based decision about speech presence

\[
I(\ell) = \begin{cases} 
1, & \text{if } \sum_{k = k_1}^{k_2} I(k, \ell) \geq 1 \text{ (speech is present)} \\
0, & \text{otherwise (speech is absent)}
\end{cases}
\]

(9)

Figure 3 shows the VAD results as applied to the LDV signal in Fig. 2(b), where results are zoomed-in to show only those frequencies that consists of useful speech information. The VAD decisions \( I(\ell) \) and \( I(k, \ell) \) are depicted in the upper and middle Figures, respectively. Clearly, the algorithm successfully detects the harmonic patterns without compromising for high false-detections rate due to speckle noise. The lowest Figure shows the actual pitch frequency, which is estimated based on the VAD decision by using a simple parabolic-fitting procedure.

Recall that a major drawback of the LDV sensor is its inability to measure unvoiced phonemes. Had these phonemes...
been located in the middle of a speech segment, they could still be identified as speech by imposing a minimal pause duration between detected segments. However, to successfully cope with unvoiced phonemes at the beginning and ending of words, we propose to extend each detected speech segment by classifying frames before and after this segment also as speech. Finally, false-detected speech segments with relatively short duration are discarded by requiring a minimal duration for a speech segment.

4. SPEECH ENHANCEMENT ALGORITHM

In this section, we introduce a speech-enhancement algorithm that employs the LDV-based time-frequency VAD $I(k, \ell)$, derived in the previous section. Let $x(n)$ and $d(n)$ denote speech and uncorrelated additive noise signals, respectively, and let $y(n) = x(n) + d(n)$ be the observed signal measured in an acoustic sensor. In the STFT domain, we have $Y(k, \ell) = X(k, \ell) + D(k, \ell)$. Let $H_0(k, \ell)$ and $H_1(k, \ell)$ indicate, respectively, speech absence and presence hypotheses in the time-frequency bin $(k, \ell)$. An estimator for the clean speech STFT signal $X(k, \ell)$ is traditionally obtained by applying a gain function to each time-frequency bin, i.e., $\hat{X}(k, \ell) = G(k, \ell)Y(k, \ell)$. In the following, we use the OM-LSA estimator [7], which minimizes the log-spectral amplitude under signal presence uncertainty, resulting in

$$ G(k, \ell) = \left(G_{H_1}(k, \ell)\right)^{p(k, \ell)} G_{\min}^{1-p(k, \ell)}, \quad (10) $$

where $G_{H_1}(k, \ell)$ is a conditional gain function given $H_1(k, \ell)$, $G_{\min} \ll 1$ is a constant attenuation factor, and $p(k, \ell)$ is the conditional speech presence probability. Denoting by $\xi(k, \ell)$ and $\gamma(k, \ell)$ the a priori and a posteriori SNRs, respectively, we get [7]

$$ p^{-1}(k, \ell) = 1 + [1 + \xi(k, \ell)] e^{\nu(k, \ell)} q(k, \ell) / [1 - q(k, \ell)], \quad (11) $$

where $q(k, \ell) = \mathcal{P}(H_0(k, \ell))$ is the a priori probability for speech absence, and $\nu \triangleq 7\xi/(1+\gamma)$. In highly non-stationary noise environments and low SNR conditions, it is difficult to determine $q(k, \ell)$, and therefore the estimator (10) does not yield satisfactory results.

Nonetheless, a reliable estimator for the speech presence probability can be attained by using the LDV-based VAD decision. Specifically, for each speech frame $\ell_0$ [i.e., $I(\ell_0) = 1$], and for every frequency bin $k$ ($k_1 \leq k \leq k_2$), we denote by $\hat{k}_0$ the nearest frequency bin that contains speech, i.e., $\hat{k}_0 = \arg\min_{k_0 \in \mathcal{K}} \{k|k - k_0|\}$ where $\mathcal{K} = \{k|I(k, \ell_0) = 1\}$, and define the following estimator for the speech-presence probability:

$$ \hat{p}(k, \ell_0) = \frac{|Z(k, \ell_0)|^2}{|Z(\hat{k}_0, \ell_0)|^2}. \quad (12) $$

Then, an estimate for $p(k, \ell)$ from (12) is achieved by substituting $1 - \hat{p}(k, \ell_0)$ for $q(k, \ell)$, the a priori probability, where $k_1 \leq k \leq k_2$. To further enhance time-frequency bins that are probable to contain speech, we set $p(k, \ell) = 1$ whenever $\hat{p}(k, \ell) > p_h$, where $p_h$ is a pre-defined parameter. It should be noted that for $k > k_2$, the estimated speech-presence probability from the OM-LSA algorithm is utilized [7]. For noise-only frames $\ell_0$ [i.e., $I(\ell_0) = 0$], $p(k, \ell)$ is set to 0 for $0 \leq k \leq N - 1$. In this case, we further attenuate high-energy transient components to the level of the stationary background noise by updating the gain floor in (10) as $\hat{G}_{\min} = G_{\min} \sqrt{\hat{\lambda}_u(k, \ell)/S_y(k, \ell)}$, where $\hat{\lambda}_u(k, \ell)$ is the stationary noise-spectrum estimate and $S_y(k, \ell) = \mu S_y(k, \ell - 1) + (1 - \mu) |Y(k, \ell)|^2$ is the smoothed noisy spectrum ($0 < \mu < 1$).

The proposed enhancement algorithm is applied to the noisy signal of Fig. 2(a), using the following parameters: $N = 512, M = 128, k_1 = 3, k_2 = 21, \alpha_\rho = 0.85, K_s = 1, L_c = 2, N_c = 3, p_h = 0.6$, and $\mu = 0.8$. The spectrogram and waveform of the resulting signal are shown in Fig. 4. We observe that a significant suppression of the background noise is achieved, while still retaining weak speech components. Subjective listening tests confirm that the proposed approach substantially improves speech intelligibility.

5. EXPERIMENTAL RESULTS

In this section, we present ASR experimental results that demonstrate the effectiveness of the proposed enhancement algorithm in improving speech recognition accuracy. The speech recognition engine used is the HTK large-vocabulary
decoder (HVite). The acoustic features include 12 mel-frequency cepstral coefficients (MFCCs) and energy, extracted from the speech signals using their first- and second-order derivatives (a total number of 39 features per frame). We employ a 32 ms frame size, with a 16 ms overlap between consecutive frames. Regarding the acoustic models, we use 39 phonemes (ARPA phoneme-set), each modeled using a 3-state HMM with left-to-right topology, and 6 additional models for background noises. The HMM states are clustered into 3383 context-dependent triphone states, where the state’s output probabilities are modeled using 16 Gaussian mixture. The training data for the engine corresponds to the telephony speech of the Macrophone database, including 4005 speakers with a total duration of 44 hours. In addition, we use a closed-grammar dialing language model of the form: 
"[call | dial | phone] <name> [at (home | the office | work) | on (mobile | cellular)]", where <name> may be one of the 32 names recorded.

The noisy database for testing was recorded in parked- and moving-vehicle scenarios, using 50 utterance per session (total number of 200 utterances). The speakers were recoded by the LDV sensor, located 1 m from the speaker, and by additional two acoustic sensors: a laptop (T60) microphone and an omni-directional SM63 microphone by Shure; both located 40 cm from the speaker. The laptop sensor attains an averaged SNR of 2.12 dB in the parked-vehicle scenario, and 0.47 dB in the moving vehicle; whereas the Shure sensor yields 4.76 dB and 1.32 dB SNR values, respectively. The ASR performance was examined for both the original noisy signals and the enhanced signals using the proposed approach. Tables 1 and 2 summarizes the recognition results (in [%]) for both parked- and moving-vehicle scenarios, respectively, including deletion (Del.), substitution (Sub.), insertion (Ins.), and total word-error rate (WER). We observe that the speech-enhancement algorithm achieves a substantial improvement of approximately 60% in recognition accuracy, compared to using the noisy signals without enhancement. The most significant improvement is attained in deletion, which may be attributable to using a reliable VAD from the LDV signal.

### 6. CONCLUSIONS

We have presented a robust speech-recognition approach that utilizes an auxiliary LDV sensor. A time-frequency VAD was derived from the LDV-measured signal using a correlation-based detector, followed by a robust harmonic-tracking algorithm. The resulting VAD was then used to modify the gain function of the OM-LSA algorithm to further improve its performance. The enhancement algorithm was applied to the noisy acoustic signals, which were then inserted as input files for the ASR engine. A substantial improvement in recognition accuracy under low SNR conditions is achieved by the proposed approach, when compared to using the noisy signals without enhancement.

We note that an effort is currently underway to develop a small laser-based sensor, which does not require heavy equipment and may be more suitable for practical use in real voice communication systems.

### 7. REFERENCES


