SPEECH MEASUREMENTS USING A LASER DOPPLER VIBROMETER SENSOR: APPLICATION TO SPEECH ENHANCEMENT

Yekutiel Avargel

AudioZoom Ltd
P.O. box 114
Midreshet BenGurion, Sde-Boker, Israel
kuti@audio-zoom.com

Israel Cohen

Department of Electrical Engineering
Technion – Israel Institute of Technology
Technion City, Haifa 32000, Israel
icohen@ee.technion.ac.il

ABSTRACT

In this paper, we present a remote speech-measurement system, which utilizes an auxiliary laser Doppler vibrometer (LDV) sensor. 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 immune to acoustical disturbances. For improved speech enhancement, we propose a new algorithm for efficiently combining the signals from the LDV-based sensor and a standard acoustic sensor. The algorithm includes a pre-filtering process, to suppress impulsive noises that severely degrade the LDV-measured speech, and a soft-decision voice activity detector (VAD) in the time-frequency domain. Experimental results demonstrate the performance of the proposed system in transient noise environments.

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

1. INTRODUCTION

Achieving high speech intelligibility in noisy environments is one of the most challenging and important problems for existing speech-enhancement and speech-recognition systems [1, 2]. Under low signal-to-noise ratio (SNR) conditions and highly non-stationary noise environments, the perceived speech quality is severely degraded, and existing voice communication systems fail to properly suppress interferences in such conditions.

Recently, several approaches have been proposed that make use of auxiliary nonacoustic sensors, such as bone- and throat- microphones (e.g., [3–7]). Such sensors typically measure vibrations of the speech-production anatomy (e.g., vocal-fold vibrations) and are relatively immune to acoustic interferences [3]. The speech information captured by these sensors can then be combined with the acoustic noisy signal to further improve speech intelligibility. In [4], air- and throat-microphones are combined by training features mapping from both sensors to improve noise robustness of automatic speech recognition (ASR) systems. In [5], a voice activity detector (VAD) is constructed from a throat sensor to improve speech recognition accuracy. A multisensory technique is demonstrated in [6] for improved speech enhancement, and a general electromagnetic motion sensor (GEMS) is utilized in [7] 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.

In this paper, we present an alternative approach that enables a remote measurement of speech, using an auxiliary laser Doppler vibrometer (LDV) sensor. An LDV is a non-contact measurement device which is capable of measuring vibration frequencies of moving targets [8]. 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. We propose a speech enhancement scheme for efficiently combining the LDV signal with an acoustic signal degraded by highly non-stationary noise. Since the LDV-measured signal is characterized by impulse-like noise (due to random constructive and destructive interferences of backscattering waves), we include a pre-filtering process to efficiently suppress impulsive noises. A soft-decision VAD in the time-frequency domain is derived and incorporated into the optimally-modified log-spectral amplitude (OM-LSA) algorithm [1] to further enhance its performance under highly non-stationary noise conditions. Experimental results demonstrate both noise robustness and improved speech intelligibility compared to using the acoustic sensor alone. It is worthwhile noting that the enhanced signal can be used as an input to existing ASR systems to improve recognition accuracies. A detailed ASR performance evaluation, however, is currently under research.

The paper is organized as follows. In Section 2, we describe the basic principles of LDV in measuring acoustic speech signals. In Section 3, we formulate the problem of speech enhancement using auxiliary LDV measurements. In Section 4, we propose a new enhancement approach using an LDV-based VAD in the time-frequency domain, and finally in

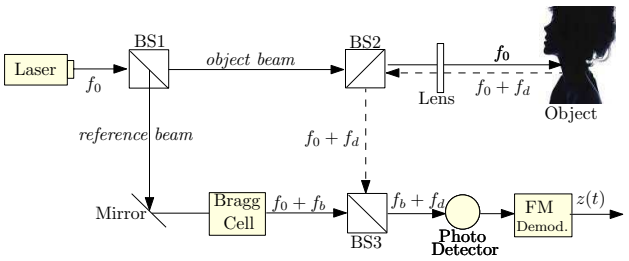

Fig. 1. Block diagram of a laser Doppler vibrometer (LDV).

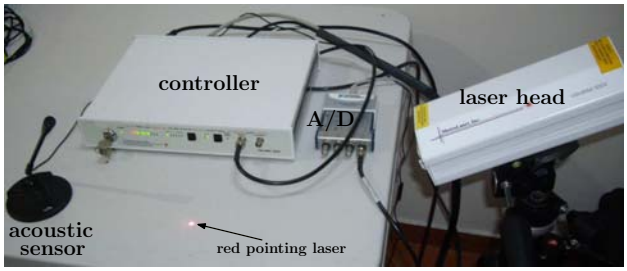

Fig. 2. Experimental setup.

Section 5, we present experimental results that demonstrate the effectiveness of the proposed approach.

2. ACOUSTIC SPEECH MEASUREMENTS WITH LDV

In this section, we briefly review the basic principles of LDV in measuring acoustic speech signals and describe our measurement setup.

2.1. Principles of LDV

An LDV is a non-contact measurement device which measures, based on the principle of interferometry, the Doppler frequency shift of a laser beam reflected from a moving (vibrating) target. In our case, the LDV sensor is directed to a speaker's throat and measures its vibration velocity (e.g., vocal-fold vibrations), as illustrated in Fig. 1. A coherent beam from the laser, with frequency f_0, is divided into a reference beam and an object beam using a beam-splitter BS1. The object beam, which passes through a beam-splitter BS2, is directed to the vibrated object (speaker's throat) by an optical lens, and backscattered to a beam-splitter BS3 with a Doppler shift f_d. This frequency shift is related to the instantaneous throat-vibrational velocity $\nu(t)$ via $f_d(t) = 2\nu(t)\cos(\alpha)/\lambda$, where α is the angle between the object beam and the velocity vector, and λ is the laser wavelength. Simultaneously, the reference beam passes through a Bragg cell, which produces a frequency shift of f_b. The resulting beam-shifted beams (object and reference) are mixed together at the beam-splitter BS3 to generate a signal with frequency $f_b + f_d$, which is then converted to a voltage signal by a photo-detector (e.g., a photodiode). Clearly, the resulting signal is a frequency-modulated (FM) signal with f_b and f_d being its carrier and modulated frequencies, respectively. For a vibration frequency f_v with amplitude A_v, for instance, the LDV-output signal after an FM-demodulator is

$$z(t) = f_b + [2A_v \cos(\alpha)/\lambda] \cdot \cos(2\pi f_v t)\,. \qquad (1)$$

2.2. Measurement Setup

The experiments presented in this paper are conducted by employing the VibroMet^{TM} 500V LDV from MetroLaser [9] that consists of a remote laser-sensor head and an electronic controller (see Fig. 2). The device operates at 780 nm wavelength and may detect vibration frequencies from DC to over 40 kHz; thus being suitable for measuring voice vibrations. Its operational working distance ranges from 1 cm to 5 m. Note that the MetroLaser LDV is presented here only to demonstrate a remote speech measurement with laser-based sensors. Its practical use in real voice communication systems is somehow limited due to its relatively heavy equipment. A new practical laser-based sensor, which is small and does not require heavy equipment, is currently under development.

In our experimental setup, a speaker is located at a distance of 75 cm from the LDV and 1 m from the acoustic sensor. Figure 3 shows the spectrogram and waveform of the speech signal, measured by the LDV with a sampling rate of 8 kHz, in a relatively noise-free environment. It should be noted, though, that the LDV speech measurements are relatively immune to acoustic interferences and insensitive to facial movements (i.e., vertical or horizontal head movements). Nonetheless, when a speaker moves outside the laser-beam direction, the beam should be re-focused on the speaker's throat. Figure 3 shows that when focusing on the larynx, the LDV sensor captures useful speech information only at low-frequency regions (up to 1.5 kHz). In addition, we 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 may severely limit the applicability of LDV-based measurement devices. Speckle noise arises from random constructive and destructive interferences of waves that backscatter from a relatively rough surface. An algorithm for attenuating this noise is presented in Section 4.

3. PROBLEM FORMULATION

In this section, we formulate the problem of speech enhancement, assuming an auxiliary LDV measurement of the speech signal is available. Let $x(n)$ and $d(n)$ denote speech and un-

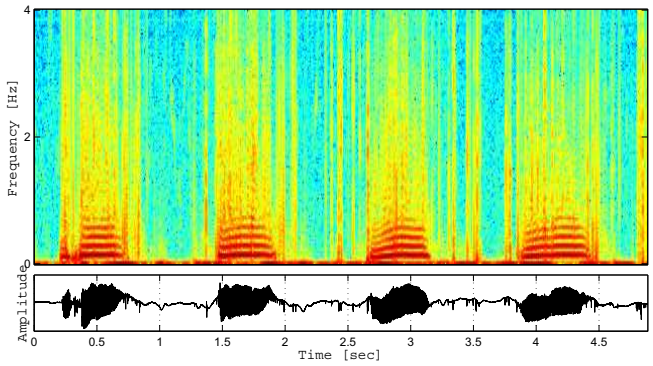

Fig. 3. Spectogram and waveform of a speech signal measured by LDV.

correlated additive noise signals, respectively, and let $y(n) = x(n) + d(n)$ be the observed signal in the acoustic sensor. In the STFT domain, we have $Y_{\ell k} = X_{\ell k} + D_{\ell k}$, 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 $H_0^{\ell k}$ and $H_1^{\ell k}$ indicate, respectively, speech absence and presence hypotheses in the time-frequency bin (ℓ, k), i.e.,

$$\begin{aligned} H_0^{\ell k} &: Y_{\ell k} = D_{\ell k} \\ H_1^{\ell k} &: Y_{\ell k} = X_{\ell k} + D_{\ell k}\,. \end{aligned} \tag{2}$$

An estimator for the clean speech STFT signal $X_{\ell k}$ is traditionally obtained by applying a gain function to each time-frequency bin, i.e., $\hat{X}_{\ell k} = G_{\ell k} Y_{\ell k}$. The OM-LSA estimator [1] minimizes the log-spectral amplitude under signal presence uncertainty, resulting in

$$G_{\ell k} = \{G_{H_1;\ell k}\}^{p_{\ell k}}\, G_{\min}^{1-p_{\ell k}}\,, \tag{3}$$

where $G_{H_1;\ell k}$ is a conditional gain function given $H_1^{\ell k}$, $G_{\min} \ll 1$ is a constant attenuation factor, and $p_{\ell k}$ is the conditional speech presence probability. Denoting by $\xi_{\ell k}$ and $\gamma_{\ell k}$ the *a priori* and *a posteriori* SNRs, respectively, we get [1]

$$p_{\ell k}^{-1} = 1 + (1 + \xi_{\ell k})\, e^{\upsilon_{\ell k}} q_{\ell k} / (1 - q_{\ell k})\,, \tag{4}$$

where $q_{\ell k} = \mathcal{P}\left(H_0^{\ell k}\right)$ is the *a priori* probability for speech absence, and $\upsilon_{\ell k} \triangleq \gamma_{\ell k}\xi_{\ell k}/(1+\gamma_{\ell k})$. In highly non-stationary noise environments, it is difficult to determine $q_{\ell k}$, and therefore the estimator (3) does not yield satisfactory results. To further attenuate noise transients, while not compromising for higher speech-components degradation, a reliable estimator for the speech presence probability is required.

4. SPEECH ENHANCEMENT ALGORITHM

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. This VAD is then used as an estimator for the speech presence probability and incorporated into the OM-LSA algorithm to enhance its performance in highly non-stationary noise environments. The LDV signal is first pre-filtered with a high-pass filter (at approximately 50 Hz), in order to reduce its relatively large DC energy. The resulting filtered signal is denoted by $z(n)$.

4.1. Speckle-Noise Suppression

Motivated by the impulsive nature of speckle noise, we propose a decision rule based on the signal kurtosis. The use of kurtosis for detecting speckle noise was first introduced in [10] for LDV-based mechanical fault diagnostic, and is extended here to speech signals.

The signal $z(n)$ is divided into overlapping frames by the application of a length-N window function $h(n)$: $z_\ell(n) = z(n + \ell M)h(n)$ for $0 \le n \le N-1$. Let $\mathcal{K}_\ell = E\left\{[z_\ell(n) - E\{z_\ell(n)\}]^2\right\}/\sigma_{z_\ell}^4$ denote the kurtosis on the ℓth frame, where $\sigma_{z_\ell}^2$ is its variance. The larger the amount of speckle noise in a given frame, the higher is the kurtosis on that frame. The kurtosis is smoothed in time using a first-order recursive averaging with a time constant α_s:

$$\mathcal{K}_{\text{av},\ell} = \alpha_s \mathcal{K}_{\text{av},\ell-1} + (1-\alpha_s)\mathcal{K}_\ell\,. \tag{5}$$

Moreover, in order to avoid false speckle-noise detection at the beginnings and endings of voiced phonemes, we consider the kurtosis of $\{z_\ell(n)\}_{n=0}^{N-M-1}$ and $\{z_\ell(n)\}_{n=M}^{N-1}$ (denoted by $\mathcal{K}_{\text{b};\ell}$ and $\mathcal{K}_{\text{e};\ell}$, respectively) and propose the following rough decision about speckle-noise presence:

$$I_\ell = \begin{cases} 1\,, & \text{if } \mathcal{K}_{\text{av},\ell}, \mathcal{K}_{\text{b},\ell}, \text{and } \mathcal{K}_{\text{e},\ell} > \mathcal{K}_0 \\ 0\,, & \text{otherwise} \end{cases}\,, \tag{6}$$

where $\mathcal{K}_0$ is a kurtosis threshold. At a beginning (or ending) of a phoneme, the value of either $\mathcal{K}_{\text{b};\ell}$ or $\mathcal{K}_{\text{e};\ell}$ decreases; thus reducing the probability of falsely detecting speckle noise in that frame. The output of the speckle-noise detector is then defined by

$$w_\ell(n) = G_\ell z_\ell(n)\,, \tag{7}$$

where $G_\ell = G_{\text{s;min}} \ll 1$ for $I_\ell = 1$ (speckle-noise is present), and $G_\ell = 1$ otherwise. Figure 4 shows the resulting signal achieved by applying the proposed speckle-reduction algorithm to the measured signal of Fig. 3. Clearly, the speech quality is improved and the speckle noise is substantially suppressed.

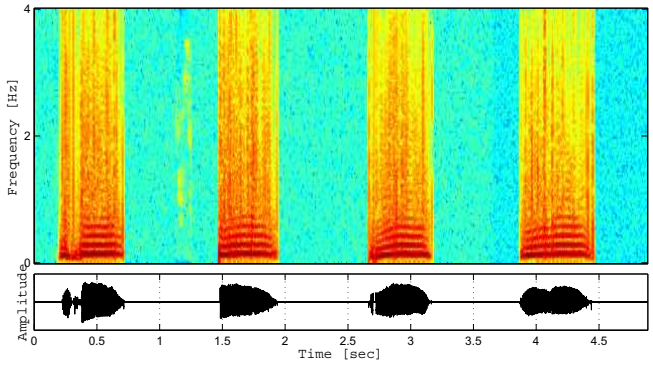

Fig. 4. Spectogram and waveform of an enhanced LDV speech signal achieved by applying the algorithm presented in Section 4.1 to the signal of Fig. 3.

4.2. LDV-Based Time-Frequency VAD

A soft-decision VAD is derived in the time-frequency domain based on the signal $w_\ell(n)$ and the minima-controlled-estimation algorithm [2]. Specifically, we define $S_{\ell k}$ to be a smoothed-version of the power spectrum $|W_{\ell k}|$, where $W_{\ell k}$ is the Fourier transform of $w_\ell(n)$. The smoothing is performed in both time and frequency domains. Let $S_{\min}^{\ell k}$ denote the minimum value of $S_{\ell k}$ within a finite window of length D, and let $\bar{\gamma}_{\ell k} \triangleq |W_{\ell k}|^2 \,/\, \left(B_{\min} S_{\min}^{\ell k}\right)$, where $B_{\min}$ represents the noise-estimate bias [2]. Then, we propose the following soft-decision VAD:

$$p'_{\ell k} = \begin{cases} 1\,, & \text{if } \bar{\gamma}_{\ell k} > \bar{\gamma}_1 \\ 0\,, & \text{if } \bar{\gamma}_{\ell k} < \bar{\gamma}_0 \\ \frac{\log(\bar{\gamma}_{\ell k}) - \log(\bar{\gamma}_0)}{\log(\bar{\gamma}_1) - \log(\bar{\gamma}_0)}\,, & \text{otherwise.} \end{cases} \tag{8}$$

Note that the ratio between the thresholds $\bar{\gamma}_0$ and $\bar{\gamma}_1$ should be sufficiently large, since the noise level in $w_\ell(n)$ may be significantly low [see (7)]. Finally, to retain weak speech components, $p'_{\ell k}$ is smoothed in time, yielding

$$\tilde{p}_{\ell k} = \alpha_p \tilde{p}_{\ell-1 k} + (1-\alpha_p) p'_{\ell k}\,. \tag{9}$$

4.3. Spectral Gain Modification

In the following, we incorporate (9) into the OM-LSA spectral gain (3). Initially, the likelihood of speech in a given frame is defined by

$$P_\ell = \text{mean}\left\{\tilde{p}_{\ell k} | k_1 \le k \le k_2\right\}\,, \tag{10}$$

where the values of k_1 and k_2 are imposed by the frequency range of the LDV signal that contains useful speech information (see Section 2.2). The modification of the OM-LSA gain is then determined by comparing P_ℓ to a given threshold $P_{\min}$, as follows.

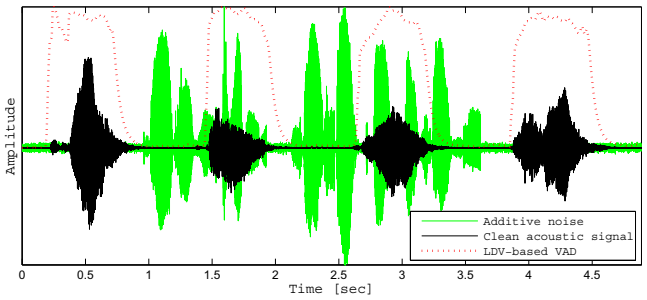

Fig. 5. Waveforms of the clean and noise signals (4 dB segmental SNR). The frame-based VAD decision (10) is depicted by a dotted line.

For any frame ℓ that satisfies $P_\ell \ge P_{\min}$, speech is assumed present. Accordingly, an estimate for $p_{\ell k}$ from (4) is achieved by substituting the smoothed VAD decision $\tilde{p}_{\ell k}$ from (9) for $q_{\ell k}$, the *a priori* probability, where $k_1 \le k \le k_2$. To further enhance the time-frequency bins that are probable to contain speech, we set $p_{\ell k} = 1$ whenever $\tilde{p}_{\ell k} > p_h$ and set $p_{\ell k} = 0$ for $\tilde{p}_{\ell k} < p_l$, where p_h and p_l are predefined parameters. On the other hand, for frames where $P_\ell \le P_{\min}$, speech is assumed absent, and $p_{\ell k}$ is set to 0 for $0 \le k \le N-1$. We further attenuate high-energy transient components to the level of the stationary background noise by updating the gain floor in (3) to $\tilde{G}_{\min} = G_{\min}\sqrt{\hat{\lambda}_{s,\ell k}/S_{y,\ell k}}$, where $\hat{\lambda}_{s,\ell k}$ is the stationary noise-spectrum estimate and $S_{y,\ell k} = \mu S_{y,\ell-1 k} + (1-\mu)\,|Y_{\ell k}|^2$ is the smoothed noisy spectrum ($0 < \mu < 1$).

5. EXPERIMENTAL RESULTS

In this section, we demonstrate the performance of the proposed approach in enhancing speech signals in highly non-stationary noise environments. The experimental setup is described in Section 2.2 (see Fig. 2). The desired speaker is degraded by an additional undesired speaker and a stationary background noise, and measured simultaneously by the LDV and the acoustic sensor with a sampling rate of 8 kHz. For the STFT, we use a Hamming analysis window of 32 ms length with 75% overlap between consecutive windows. For all the considered algorithms, the background-noise spectrum is estimated by using the improved minima-controlled recursive averaging (IMCRA) algorithm [2]. The values of the parameters used in the implementation of the proposed algorithm are: $\alpha_s = 0.9$, $\mathcal{K}_0 = 9$, $G_{\text{s;min}} = 0.01$ (Section 4.1); $\bar{\gamma}_0 = 1.5$ dB,$\bar{\gamma}_1 = 40$ dB, $\alpha_p = 0.85$ (Section 4.2); $P_{\min} = 0.1$, $p_h = 0.7$, $p_l = 0.1$, and $\mu = 0.8$ (Section 4.3). The OMLSA gain floor is set to $G_{\min} = 0.1$.

Figure 5 shows the waveforms of the clean and additive noise signals as well as the frame-based VAD decision de-

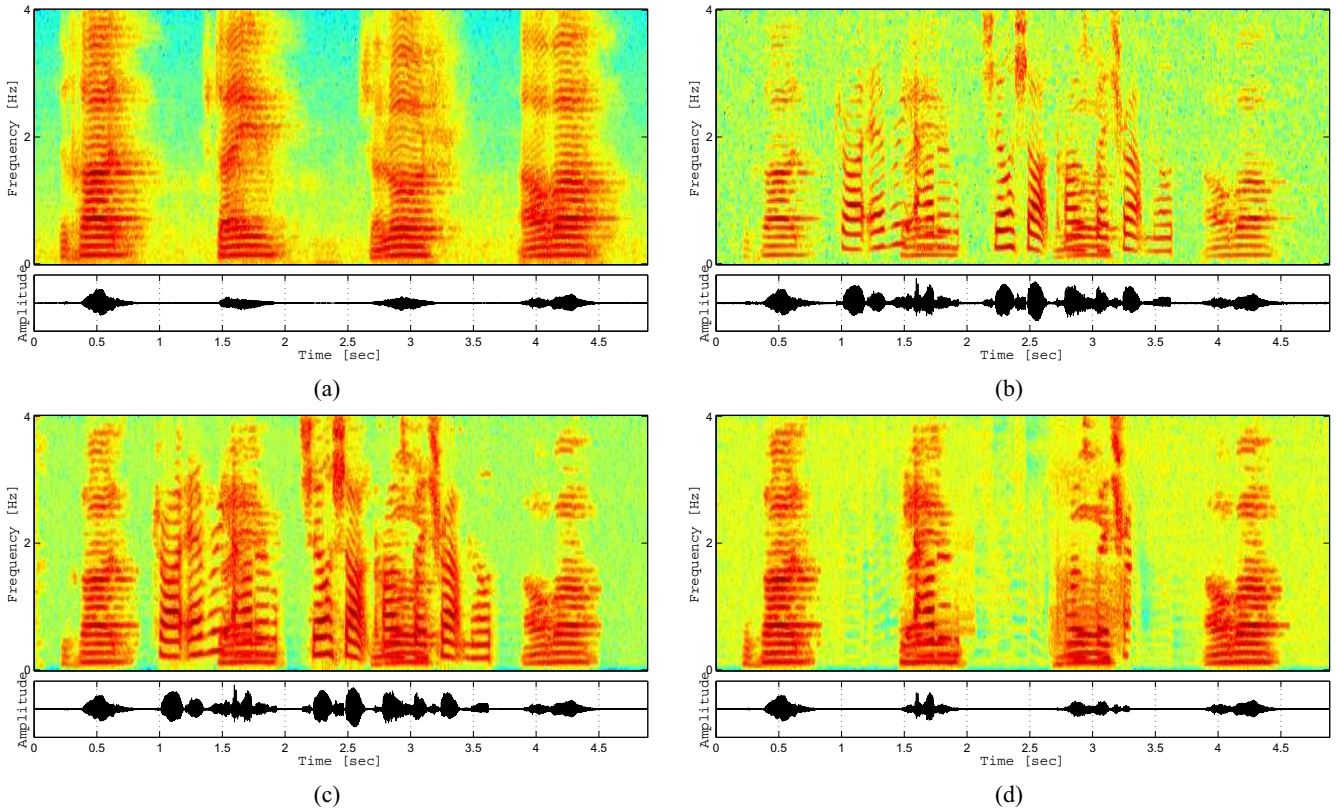

Fig. 6. Speech Spectrograms and waveforms. (a) Clean speech signal measured by the acoustic sensor. (b) Noisy signal (additional speaker and stationary noise; 4 dB segmental SNR). (c) Speech enhanced using the OMLSA algorithm. (d) Speech enhanced using the proposed algorithm.

fined in (10). Clearly, the LDV-based VAD accurately tracks the clean acoustic speech even under non-stationary noise conditions. The corresponding spectrograms and waveforms are shown in Fig. 6, including the speech-signal estimate as obtained by applying the OMLSA to the acoustic sensor [Fig. 6(c)] and the proposed approach [Fig. 6(d)]. The signal measured by the LDV and its enhanced version are depicted, respectively, in Figs. 3 and 4. Table 1 summarizes three objective quality measures: segmental SNR (segSNR), log-spectral distortion (LSD) and noise reduction (NR). We observe that when the desired speaker is inactive, a substantial suppression of the non-stationary interference is achieved by the proposed approach (-31 dB noise reduction); whereas without the LDV sensor, the OMLSA algorithm expectedly fails to eliminate the undesired speaker. Moreover, during desired-speech frames, an improvement in speech quality is attained by the proposed approach, compared to applying the standard OMLSA algorithm to the acoustic sensor. Specifically, an improvement of 1.3 dB in SegSNR and 4 dB in LSD is evident.

Table 1. Segmental SNR, Log-Spectral Distortion and Noise Reduction Obtained Using the Acoustic Sensor Only (Without LDV) and the Proposed Approach (With LDV).

Method	SegSNR [dB]	LSD [dB]	NR [dB]
Noisy speech	4.01	9.2	0
Without LDV	6.35	7.01	-8.03
With LDV	7.64	3.01	-31.2

6. CONCLUSIONS

We have presented a remote speech-measurement system that utilizes an auxiliary LDV sensor, and proposed a speech-enhancement algorithm based on these measurements. Speckle noise was successfully attenuated from the LDV-measured signal using a kurtosis-based decision rule. A soft-decision VAD was derived in the time-frequency domain and the gain function of the OM-LSA algorithm was appropriately modified. The effectiveness of the proposed approach in suppressing highly non-stationary noise components was demonstrated.

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. Future research will concentrate on evaluating a detailed ASR performance using the proposed speech-enhancement approach.

7. REFERENCES

[1] I. Cohen and B. Berdugo, "Speech enhancement for nonstationary noise environment," *Signal Process.*, vol. 81, pp. 2403–2418, Nov. 2001.

[2] I. Cohen, "Noise spectrum estimation in adverse environments: Imroved minima controlled recursive averaging," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 5, pp. 466–475, Sep. 2003.

[3] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood, "Exploiting nonacoustic sensors for speech encoding," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 2, pp. 533–544, Mar. 2006.

[4] M. Graciarena, H. Franco, K. Sonmez, and H. Bratt, "Combining standard and throat microphones for robust speech recognition," *IEEE Signal Process. Lett.*, vol. 10, no. 3, pp. 72–74, Mar. 2003.

[5] T. Dekens, W. Verhelst, F. Capman, and F. Beaugendre, "Improved speech recognition in noisy environments by using a throat microphone for accurate voicing detection," in *18th European Signal Processing Conf. (EUSIPCO)*, Aallborg, Denmark, Aug. 2010, pp. 23–27.

[6] Z. Zhang, Z. Liu, M. Sinclair, A. Acero, L. Deng, J. Droppo, X. Huang, and Y. Zheng, "Multisensory microphones for robust speech detection, enhancement and recognition," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, Montreal, Canada, May 2004, pp. 781–784.

[7] C. Demiroglu, S. Kamath, D. V. Anderson, M. Clements, and T. Barnwell, "Segmentation-based noise suppression for speech coders using auxiliary sensors," in *Conf. Rec. Thirty-Eighth Asilomar Conf. on Signals, Systems and Computers*, Nov. 2004, pp. 2320 – 2323.

[8] M. Johansmann, G. Siegmund, and M. Pineda, "Targeting the limits of laser doppler vibrometry," in *Proc. IDEMA*, 2005, pp. 1–12.

[9] [Online]. Available: http://www.metrolaserinc.com

[10] J. Vass, R. Smid, R. Randall, P. Sovka, C. Cristalli, and B.Torcianti, "Avoidance of speckle noise in laser vibrometry by the use of kurtosis ratio: Application to mechanical fault diagnostics," *Mechanical Systems and Signal Process.*, vol. 22, pp. 647–671, 2008.