Long Range Standoff Speaker Identification Using Laser Doppler Vibrometer

Yunbin Deng
BAE Systems
Burlington, MA 01803
Yunbin.deng@baesystems.com

Abstract

Existing studies on speaker identification are mostly performed on telephone and microphone speech data, which are collected with subjects close to the sensor. For the first time, this study reports long range standoff automatic speaker identification experiments using laser Doppler vibrometer (LDV) sensor. The LDV sensor modality has the potential to extend the speech acquisition standoff distance far beyond microphone arrays to enable new capabilities in automatic audio and speech intelligence, surveillance, and reconnaissance (ISR). Five LDV speech corpuses, each consists of 630 speakers, are collected from the vibrations of a glass window, a metal plate, a plastic box, a wood slate, and a concrete wall, using Polytec LDV model OFV-505. The distance from the LDV sensor to the vibration targets is 50 feet. State of the art i-vector speaker identification experiments on this LDV speech data show great promise of this LDV long range acoustic sensing modality.

1. Introduction

Acoustic ISR applications require high signal to noise ratio (SNR) speech signal acquisition from uncooperative individuals at long distance, in covert mode, and sometimes without directly line of sight to the subject. Existing standoff speech acquisition method uses microphone array to enhance SNR, but still has very limited range due to the quick drop of speech pressure as it propagates through the air. In addition, the speech signals acquired by the standoff microphone array are contaminated by acoustic noise near the signal source, or near the sensor, or along the path between the sensor and the source. Furthermore, when the speaker is inside a building or vehicle without direct line of sight, the microphone array often can't collect usable speech. As such, most reported studies on microphone array speaker identification had very limited range and accuracy. For example, McCowan used microphone array to speaker distance of 70cm in a conference room setting [1]. Wang studied microphone array to speaker distance up to 3 meters [2]. Mematollahi gave an overview of distant speaker recognition, where different datasets and technologies were reviewed for this interesting non-intrusive speaker identification application domain. Equal Error Rate (EER) around 10% were achieved at 6 feet microphone to source distance [3].

LDV provides an alternative means of distant speech acquisition based on Doppler effect. The voice source signal can cause vibrations of any surrounding objects. Such vibrations are usually at a micrometer or even nanometer scale. The displacement and speed of these small vibrations can be captured using a LDV, which is detailed in section 2.1. The captured vibrations thus represent the voice source, similar to electrical signals generated by microphone membrane vibration. In contrast to microphone speech, the LDV captured vibration signal can be very close to the voice source, thus the SNR is not impacted much by the noise close to the LDV sensor and only impacted by noise close to the source. The LDV speech sensing modality is thus much less sensitive to environmental noise and can acquire better quality data at a much longer standoff distance. In addition, the laser can be chosen at the infrared range, offering covert operation capability. Furthermore, the voice of a subject inside a building can still cause micro-vibrations on window, blinds, wall, etc, thus LDV is capable of 'listening' without direct line of sight to the voice source.

The idea of sound vibration measurement from a distance using infrared beam and laser is not new. In fact, LDV has found many applications in aerospace, acoustic, architecture, and automotive, to name just a few. Perhaps the most well-known use of laser microphone is in the spycraft, dating from early 1960s to the more recent capture Osama bin Laden [12][13]. In the field of biometrics, LDV was applied to short range standoff cardiac biometric identification [14].

However, there is no published work on using LDV for long range standoff automatic speaker identification. Existing preliminary studies using LDV for speech sensing has limited to audio event detection [4] and as an aid for speech detection [5]. In this work, we report a large scale speech vibration data acquisition using LDV from five different targets at 50 feet distance. Automatic speaker identification experiments are conducted to show the

promise of LDV long range standoff voice sensing modality for ISR applications.

The rest of the paper is organized as follows: Section 2 details the first large scale LDV speaker recognition data acquisition from five different vibration materials. Section 3 characterizes these LDV speech data. Section 4 describes different LDV speaker identification experimental setups and results. Section 5 discusses the current research and future directions. Section 6 concludes this work.

2. Standoff LDV Speech Data Collection

As LDV is a relatively new modality for the speech community, a short introduction of its principle of operation is first given in section 2.1. This is followed by some details of the LDV system used in this study, described in section 2.2. As no public LDV data exists for speaker identifications study, a LDV speech corpus is developed based on TIMIT speech corpus, detailed in section 2.3.

2.1. LDV Fundamentals

The basic LDV principle of operation is illustrated in Figure 1. The LDV sensor head generates a laser source signal at frequency f_0, which first passes through a beam splitter and then a Bragg cell to have frequency shifted by f_b [6]. The laser beam coming out of the LDV sensor head at frequency $f_0 + f_b$ travels through the air, till it hits a vibrating surface. The wavelength of laser beam can be chosen depending on the application. For example, infrared laser can be used for covert operation. For ultra-long standoff range applications, a long range focal lens can be added to the sensor head, as shown in Figure 1.

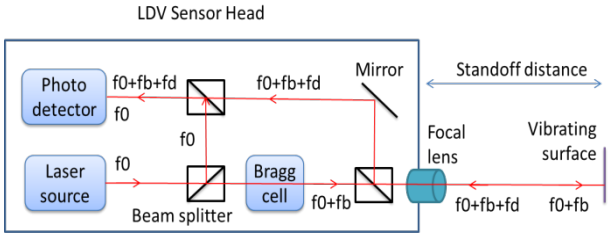

Figure 1: *LDV Principle of Operation.*

The motion of the vibration surface causes a Doppler shift, f_d, to the laser beam, given by

$$f_d = 2 * v(t) * \frac{\cos(\theta)}{\mu} \quad (1)$$

, where $v(t)$ is the vibrating velocity, θ is the angle between the laser beam and viberating surface, and μ is the laser wavelength. As such, the laser should be perpendicular to the surface to achieve the maximum dynamic range. Light scatters from the vibrating surface, but only some portion of it is reflected back into the sensor head, which is sensed by a photo detector and eventually is used to decode the f_d. The strength of the returned signal thus depends on the reflective property of the target. To achieve very good signal return, a retro-reflective tape can be put on the vibrating surface. The photo detector can sense the frequency-modulated signal at frequency f_b+f_d, with carrier frequency f_b and modulation frequency f_d. This modulated signal is further fed to a demodulator to extract signal of interest, i.e., the speed and displacement of the vibration. This vibration represents the speech source signal.

2.2. PolyTec LDV Data Acquisition System

This study uses a LDV sensor head, model *OFV-505*, from *Polytec*. The sensor head, with dimension of 4.7 x 3.1 x 13.6 (inch) and weight 3.4kg, offers a visible laser beam at wavelength 633 nm. The measurable vibration bandwidth is 0.05Hz to 1MHz and the best resolution can be down to 0.05 μm/s. The laser energy is less than 1 mW and is eye-safe [7]. The maximum achievable standoff distance is about 300 meters, if used with a super long range focal lens and a good reflective surface. In this pilot study, data collections are performed at 50 feet distance and no long range focal lens is used.

The output of the LDV sensor head is connected to a modular vibrometer controller, i.e., Polytec model *OFV-5000*. It has four internal slots for velocity and displacement decoders. In this experiment, the decoded analog velocity output from the decoder is directly fed into a laptop audio input through a BNC connector. The *OFV-5000* data collection sensitivity was set to 0.2 mm/s/v. The controller tracking speed is set to slow and the bandpass frequency range is set to 100Hz to 5 kHz.

2.3. LDV Speech Corpus Development

Although this paper reports standoff LDV speaker identification, the corpus was designed with both speaker ID and speech recognition applications in mind. As the TIMIT corpus offers a very balanced phoneme text and a large pool of 630 subjects [8], it was chosen to create the first LDV counterpart datasets. The conceptual setup of the data acquisition system is shown in Figure 2. The laptop plays the whole TIMIT corpus sentence by sentence using a Harman/kardon speaker. The played speech wave causes vibration of a surrounding target. The vibration of the surrounding target is captured by the standoff polytec LDV sensor system at 50 feet distance. The decoded speed of vibration analog signal is recorded by the laptop's sound card and saved as 16 Khz wave file.

To study the feasibility of speaker identification using LDV based on vibration from various targets, this data collection experiment considered five targets: a glass window, a metal plate, a large plastic box, a wood slate, and a concrete basement wall. The sound level of speaker was at normal 50~60 dB range. The distance from the speaker to vibrating source is about 3 to 6 feet. The distance varied for different targets to avoid LDV signal saturation. For all

target types, a retro-reflective tape is applied to the LDV targeting spot and the LDV sensor head to target distance is fixed at 50 feet. The use of retro-reflective tape is feasible for audio surveillance application of a known area. For ISR application to inaccessible area, more advanced LDV models that do not need retro-reflective tape should be used. This experiment was conducted in a basement and contains occasional background noise from heating system, water heater, and human walking on the floor above.

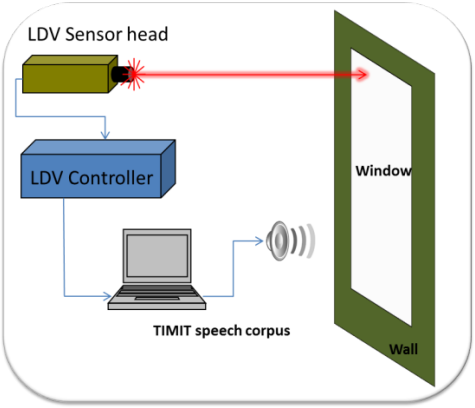

Figure 2: Automatic LDV dataset collection system.

3. LDV Speech Signal Characterization

3.1. LDV Speech Signal and Spectrum

The acquired LDV signal represents the underlying speech waveform. Each vibrating material functions as a new type of acoustic transducer, similar to the microphone membrane. However, each material type has very different vibrating characteristic and frequency response, causing very different LDV speech spectrum. The speech contents are intelligible in all cases. Figure 3 illustrate a comparison of the original TIMIT microphone speech and LDV speech collected from five different targets, representing the same utterance in TIMIT corpus (TRAIN/DR1/FCJF0/SA1.wav). The plots show that the original TIMIT microphone speech spectrum is impacted differently by each target material and often results in loss of spectrum contents.

3.2. LDV Speech Feature Extraction

As the LDV speech signal are intelligible and representing voice signature similar to microphone speech, the standard Mel-Frequency Cepstral Coefficient (MFCC) is used in this study as a baseline to extract feature for gender and speaker recognition.

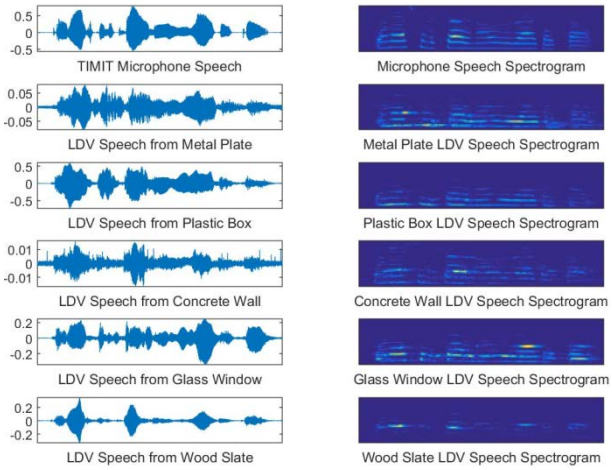

Figure 3: *Microphone and LDV speech signals and spectrogram plots show different transducer effect.*

4. Standoff Speaker Recognition Experiment

4.1. Same Material Recognition Experiment

As shown in session 3.1, the transducer effects of different target material are very different. The first experiment is to demonstrate speaker recognition feasibility using LDV speech, with training and testing data from the same type of material. The KALDI speech recognition toolkit is used to conduct i-vector based gender and speaker recognition experiments [9][10]. Specifically, for each material type, the collected LDV TIMIT dataset is used to build models in following steps:

1. MFCC features are extracted with delta feature added, mean normalized, and unvoiced frames are removed.
2. Use the original training and testing data partition in the TIMIT dataset for this experiment.
3. Build a gender-independent universal background model (UBM) with 1024 mixture of Gaussian using the training data. Start with a diagonal model and then train a full covariance model.
4. Partition the data into male and female subset for both training and testing set. Adapt the gender independent GMM to build gender-dependent full covariance UBM.
5. Gender identification experiment is run on the full test set, including both male and female subjects, consisting of 1680 tests, based on likelihood ratio on the two gender-dependent UBMs.
6. Build i-vector extractor using male subjects' training data. The i-vector dimension is kept at 400.
7. For each male subject in the test set, the first 8 utterances are used to create a single i-vector model,

representing that subject. The rest two utterances are used to create two test i-vectors.

Without loss of generality, speaker recognition experiments are only conducted on male subjects' data. The full 112 male subjects in the test set are use in all the speaker authentication experiments. Two types of tests are performed:

1. Conduct speaker authentication experiment by computing the cosine distance of each test i-vector against all speaker models in the test set. For the total 112 male subjects in the test set, 25,088 (=112*2*112) total cosine similarity scores are generated to compute final EER.
2. Linear Discriminative Analysis (LDA) is further applied to the i-vectors to reduce the i-vector dimension to 50. This step further enhances speaker authentication accuracy. The cosine distance is then applied to the dimension-reduced i-vectors and the EER are computed the same way.

The gender recognition and speaker authentication EERs for each target material are reported in Table 1. The accuracies using TIMIT microphone speech under the same parameter settings are also included in last row of Table 1 for comparison. Although the standoff LDV recognition results are not in par with close-talk microphone data, these results show great promise of using LDV for long range standoff speaker recognition applications. It also shows that LDA on i-vector helps with the recognition accuracy in most cases.

Table 1: *Gender Identification and Speaker Authentication EER (%) using LDV Speech Collected from Five Different Targets and the Original TIMIT Microphone Speech Data (last row).*

Vibrating Material	Gender ID	Speaker Authentication	
		i-vector	LDA
Glass Window	2.7	12.1	9.4
Metal Plate	16.1	13.4	9.8
Plastic Box	4.6	7.6	6.3
Wood Slate	5.4	9.8	8.5
Concrete Wall	12.4	10.7	10.7
LDV Average	**8.2**	**10.7**	**8.9**
Microphone	1.3	1.8	0.9

4.2. Mixed Data Recognition Experiment

Previous same material recognition experiment shows that the LDV speech data does not perform as good as the original TIMIT microphone speech data. This could due to a few factors: 1) The TIMIT data was collected in a well-controlled noise free environment, while these LDV datasets are collected in a low noise environment. 2) The targets selected are common real-life objects and are not good vibrators compared with microphone membrane. This results in much lower signal dynamic range, especially for the concrete wall case. 3) The transducer effects of these targets results in information loss at various frequency range, as shown in Figure 1.

The goal of this experiment is to leverage LDV data collected from all targets to enhance gender identification and speaker authentication accuracy for testing data from a specific target. The experimental set up is similar to section 4.1. The major difference is that the training data from five material are mixed together to train one i-vector model and one LDA model. The trained model is then applied separately to the same five test sets defined in section 4.1.

In addition, we applied Probabilistic Linear Discriminative Analysis (PLDA) algorithm to learn a speaker discriminative transfer while coping with noise and five different transducer effects from these five different materials. The idea is to gain performance by learning the speaker information embedding dimensions in the i-vector space [11].

The results of this mixed data training are shown in Table 2. Compared with results in Table 1, the gender identification EER was reduced by 18.3%, relatively. Note, 1024 mixtures of Gaussian was used in both experiments. Increasing the number of Gaussian mixture may further help accuracy as there are richer statistics to be modeled in this mixed data case. On average, the mixing of training data helps the i-vector and i-vector+LDA speaker authentication accuracy.

The PLDA on i-vector further improves the LDA based EER by 20%, relatively. The PLDA system achieves an average EER of 6.4% on the standoff LDV speech data. These results brings LDV based long range standoff gender and speaker recognition EER closer to a practical useable range for intelligence, surveillance, and reconnaissance applications.

Table 2: *Gender Identification and Speaker Authentication EER (%) by Mixing LDV Speech Collected from Five Different Targets. The last row show relative average EER reduction compared with Table 1.*

Vibrating Material	Gender ID	Speaker Authentication		
		i-vector	LDA	PLDA
Glass Window	4.9	9.8	6.7	5.4
Metal Plate	12.5	11.2	12.5	9.8
Plastic Box	3.2	4.9	4.9	3.6
Wood Slate	3.4	8.5	5.4	4.0
Concrete Wall	9.4	9.8	10.7	9.4
LDV Average	**6.7**	**8.8**	**8.0**	**6.4**
EER Reduction	18.3	17.8	10.1	N/A

The corresponding Detection Error Tradeoff (DET) curves for these five datasets are plotted in Figure 4, showing the tradeoff between the False Rejection Rate (FRR) and False

Acceptance Rate (FAR). The results show that the metal plate LDV data performs worse than the concrete wall LDV data. This appears to be counter intuitive. Further analysis on the LDV audio from metal target shows echo noise in some cases. This could due to the fact that the small metal plate was attached a water pipe, which introduced acoustic echoes not shown in other target cases.

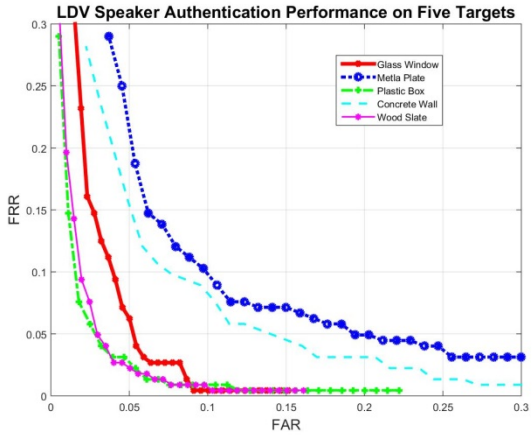

Figure 4: *Standoff speaker authentication DET curve using LDV speech from five different targets. All results are based on PLDA algorithm with mixed training data collected from 50 feet distance.*

4.3. Cross Material Recognition Experiment

Although the results of previous sections showed great promise of using LDV for long range standoff gender and speaker recognition, these studies assume a pre-defined vibrating targets. In real-world open field applications, the LDV system may not encounter the exact same vibrating target, and even the same type of material may not be available to acquire the data. The real application would require the system to work with arbitrary vibrating target available surrounding the subject, which may have a very different transducer effect on the LDV speech. Instead of collecting data from an exhaustive list of common targets, it will be ideal to have the algorithm provides reasonable good performance for a completely new object. The purpose of this across material recognition experiment is to provide such a baseline for further research in this direction.

In this setting, the gender and speaker recognition are conducted on each type of targets, but the models are trained on the rest four other types of material. This represents an extreme case where the testing target is a new material, thus provides a lower bound on the performance achievable on a new target. All parameter settings are the same as the previous two experiments.

The cross material recognition experimental results are summarized in Table 3. The last row show the absolute ERR degrades in all case, compared with mixed-trained results shown in Table 2. It is worth to mention that the PLDA algorithm significantly helps the LDV speaker recognition even under this unforeseen transducer effect testing condition. The final 10% EER serves as a performance lower bound, giving us some confidence to apply this LDV technology in an uncontrolled standoff speaker recognition settings. Note that the gender recognition is still based on the simple GMM likelihood ratio test and could poetically perform much better with i-vector, i-vector with LDA, and i-vector with PLDA technologies.

Table 3: *Gender Identification and Speaker Authentication EER (%) when Models Are Trained with LDV Speech Collected from Other Four Different Targets. The last row show absolute average EER increase compared with Table 2.*

Vibrating Material	Gender ID	Speaker Authentication		
		i-vector	LDA	PLDA
Glass Window	22.5	20.5	15.6	8.0
Metal Plate	24.9	19.2	17.0	15.6
Plastic Box	4.1	13.4	8.9	7.1
Wood Slate	11.5	18.3	10.3	8.9
Concrete Wall	12.1	21.9	19.2	10.3
LDV Average	**15.0**	**18.6**	**14.2**	**10.0**
EER Increase	8.3	9.8	6.2	3.6

5. Discussions and Future Research

This research provides the first comprehensive LDV speech corpus for long range standoff audio ISR research. The proposed algorithmic approaches show that LDV speaker recognition accuracy is close to an applicable range for ISR applications. Future LDV data collection effort should further extend the standoff range to 100 feet, 200 feet, and beyond. In addition, data should be collected in a complete uncontrolled outdoor environments under all weather conditions.

Algorithmically, future researches are needed to further address LDV specific signal processing and feature extraction approaches to improve this MFCC feature baseline. In the LDV speaker model domain, advanced approach in deep neural network may further improve performance. In addition, domain transfer and cross view approaches may be applied to better leverage correlation between the LDV and microphone speech data corpus.

From an application domain perspective, the developed technologies in speaker recognition can be applied to other domains, such as long range standoff automatic speech recognition, language identification, and vehicle classification based on vibration signals [15]. These have

important applications in civil surveillance, military ISR, and forensics.

From a system deployment perspective, a multi-model system is needed to integrate many other sub-system components, including automatic human subject localization, fast laser auto-focusing, best vibrating surface selection, real-time moving subjects tracking, environmental acoustic noise and air turbulence mitigation. These challenges are all need to be addressed for a fully functional long range acoustic LDV ISR system.

6. Conclusions

For the first time, this study proves the feasibility of automatic gender and speaker identification using LDV at a standoff range of 50 feet. Algorithmic advances show applicable performance under a loosely controlled setting from various vibrating surfaces. Further research are warranted for a fully functional long rang acoustic ISR system.

References

[1] McCowan, I. A., Pelecanos, J., & Sridharan, S. (2001). Robust speaker recognition using microphone arrays. In 2001: A Speaker Odyssey-The Speaker Recognition Workshop.

[2] Wang, L., Kitaoka, N., & Nakagawa, S. (2007). Robust distant speaker recognition based on position-dependent CMN by combining speaker-specific GMM with speaker-adapted HMM. Speech communication, 49(6), 501-513.

[3] Nematollahi, M. A., & Al-Haddad, S. A. R. (2015). Distant speaker recognition: an overview. International Journal of Humanoid Robotics, 1550032

[4] Wang, T., Zhu, Z., & Divakaran, A. (2010, April). Long range audio and audio-visual event detection using a laser Doppler vibrometer. In *SPIE Defense, Security, and Sensing* (pp. 77040J-77040J). International Society for Optics and Photonics.

[5] Avargel, Y., & Cohen, I. (2011, May). Speech measurements using a laser Doppler vibrometer sensor: Application to speech enhancement. In *Hands-free Speech Communication and Microphone Arrays (HSCMA), 2011 Joint Workshop on* (pp. 109-114). IEEE.

[6] https://en.wikipedia.org/wiki/Laser_Doppler_vibrometer

[7] Polytec OFV-505 Sensor Head Datasheet.

[8] Garofolo, John, et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. Web Download. Philadelphia: Linguistic Data Consortium, 1993.

[9] Povey, D., Ghoshal, A, Boulianne, G., Burget, L., Glembek, O., Goel, N, Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J. Stemmer, G., Vesely, K., "The Kaldi speech recognition toolkit", in Proc. IEEE ASRU, December 2011.

[10] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. Audio, Speech, and Language Processing, IEEE Transactions on, 19(4), 788-798.

[11] Ioffe, S. (2006). Probabilistic linear discriminant analysis. In Computer Vision–ECCV 2006 (pp. 531-542). Springer Berlin Heidelberg.

[12] Wallace, R., Melton, H. K., & Schlesinger, H. R. (2008). Spycraft: the secret history of the CIA's spytechs, from communism to Al-Qaeda. Penguin.

[13] "CIA Used Satellites To Prep For Bin Laden Raid". National Public Radio. 2011-05-04. Retrieved 10 June 2012. Mr. PIKE: ... shine a laser beam on those windows, you can detect those vibrations

[14] Chen, M., Sullivan, J. A., Singla, N., Sirevaag, E. J., Kristjansson, S. D., Lai, P. H., ... & Rohrbaugh, J. W. (2010). Laser doppler vibrometry measures of physiological function: evaluation of biometric capabilities. Information Forensics and Security, IEEE Transactions on, 5(3), 449-460.

[15] Deng, Y., Wang, T., Snyder, W., Fay, D., Richman, M., Standoff Vehicle Identification Using Laser Doppler Vibrometer, Military Sensing Symposia, June 2016.