Deep Neural Network for Robust Speech Recognition With Auxiliary Features From Laser-Doppler Vibrometer Sensor

Zhipeng Xie[1], Jun Du[1], Ian McLoughlin[2], Yong Xu[3], Feng Ma[3], Haikun Wang[3]

[1] NELSLIP, The University of Science and Technology of China, Hefei, PRC
[2] School of Computing, University of Kent, Medway, UK
[3] iFlytek Research

xzp2013@mail.ustc.edu.cn, jundu@ustc.edu.cn, ivm@kent.ac.uk
{yongxu5, fengma, hkwang}@iflytek.com

Abstract

Recently, the signal captured from a laser Doppler vibrometer (LDV) sensor been used to improve the noise robustness automatic speech recognition (ASR) systems by enhancing the acoustic signal prior to feature extraction. This study proposes another approach in which auxiliary features extracted from the LDV signal are used alongside conventional acoustic features to further improve ASR performance based on the use of a deep neural network (DNN) as the acoustic model. While this approach is promising, the best training data sets for ASR do not include LDV data in parallel with the acoustic signal. Thus, to leverage such existing large-scale speech databases, a regression DNN is designed to map acoustic features to LDV features. This regression DNN is well trained from a limited size parallel signal data set, then used to form pseudo-LDV features from a massive speech data set for parallel training of an ASR system. Our experiments show that both the features from the limited scale LDV data set as well as the massive scale pseudo-LDV features are able to train an ASR system that significantly outperforms one using acoustic features alone, in both quiet and noisy environments.

Index Terms: laser Doppler vibrometer, auxiliary features, deep neural network, regression model, speech recognition

1. Introduction

Automatic speech recognition (ASR) has achieved tremendous progress during last few decades, and currently performs very well under clean conditions. However, modern recognition systems suffer from severe performance degradation in the presence of unavoidable interrupting factors like environment noise, room reverberation, disturbances from different microphones and recording non-linearities [1]. To solve these problems, many processing techniques [2, 3, 4], including speech enhancement algorithms [5] and new robust acoustic features [6][7], have been developed to improve recognition performance under low signal-to-noise ratio (SNR) conditions. However these existing approaches, while achieving some improvements, are far from being a comprehensive solution.

Recently, the results of new approaches using auxiliary information gathered from non-acoustic sensors like bone-, throat- and air- microphones show that such sensors can supply useful information to help ASR systems make correct decisions under noisy environments [8][9][10]. Photo-acoustic technique shows promising results on robust recognition due to their inherent immunity to acoustic noise as well as non-contact operation [11][12]. Combining traditional acoustic features with speech information captured by these sensors, recognition performances are further improved [13]. According to [14][16], the laser doppler vibrometer (LDV) sensor is a non-contact measurement device that is capable of measuring the vibration frequencies of moving targets. It is directed at a speaker's larynx, and captures useful speech information at certain frequency bands. So far, LDV has been used to detect the remote voice signal from surrounding vibrated objects [15]. And in [16][17], LDV sensors are presented as making accurate and reliable voice activity detection (VAD) decision, as well as improving the speech recognition results.

Conventional hidden Markov model (HMM)-based speech recognizers have been used in [17] with LDV data. Each acoustic state is modeled by Gaussian mixture models (GMMs), referred to as a GMM-HMM system. However, recent studies have shown that deep neural network (DNN)-based HMM systems (denoted as DNN-HMM) perform significantly better than GMM-HMM systems on large vocabulary speech recognition tasks [18][19]. DNNs, currently one of the most popular deep learning methods, are joint models combining nonlinear feature transformation and classification [20]. DNNs have demonstrated a great capacity to extract discriminative internal representations that are robust to the many sources of variability in speech signals.

The novelty of this work is to derive LDV features from LDV sensor information, combine these with the corresponding traditional acoustic features to improve recognition performance under both clean and noisy conditions. In comparison to the recent work on LDV sensor for speech recognition [17], the main difference is we directly use LDV features for acoustic modeling while in [17] LDV information is adopted to improve the VAD and indirectly help to boost ASR system. In this sense, our proposed approach can be perfectly incorporated with [17]. Furthermore, we will show that using well-trained DNN weights for initialization leads to even greater gains in recognition performance. Due to the limited size of existing LDV datasets, we additionally consider obtaining more LDV features in training data by converting normal acoustical features from a large dataset into pseudo-LDV features. To do this, we first create and train a regression DNN to learn a mapping relationship from normal acoustic features into LDV features. The trained feature-mapping network allows pseudo-LDV features to be generated in parallel with acoustic features from acoustic-only training data, allowing us to create a very well trained DNN-based dual feature ASR system.

The rest of the paper is organized as follows. Section 2

describes the DNN acoustic system which combines acoustical features with LDV features, then Section 3 demonstrates the use of another DNN to derive pseudo-LDV features from a large dataset. Section 4 introduces the experimental conditions, datasets, system operation and discusses results. Finally we conclude the paper in Section 5.

2. LDV Feature Combination

In this section, we exploit the availability of LDV features by combining them with traditional acoustic speech features. Figure 1 shows a comparison of two different DNN based acoustic systems. One uses normal acoustic features (we refer to this as DNN_N) while the other introduces LDV features to be concatenated with the acoustic features (we refer to this as DNN_C). Both will be evaluated later in Section 4.

Our LDV dataset (which will be described in detail in Section 4.1), contains parallel acoustic microphone and corresponding LDV data files for each sentence. The traditional approach is to obtain the log Mel-filter-bank (LMFB) features from normal speech and feed these into the DNN input layer with adjacent context frames. In our system we propose combining the LMFB features from normal speech with LMFB features extracted from the LDV signal. We ensure that each feature vector has the same dimension of n. Then we merge the two features by concatenating them together into a dimension of $2n$. Here, to avoid poor local optima, pre-training methods have been proposed to better initialize the parameters prior to back propagation (BP). We use the contrastive divergence (CD) criterion to train each pair of layers in the network as restricted Boltzmann machines (RBM) and grow the network layer-by-layer in an unsupervised way [19].

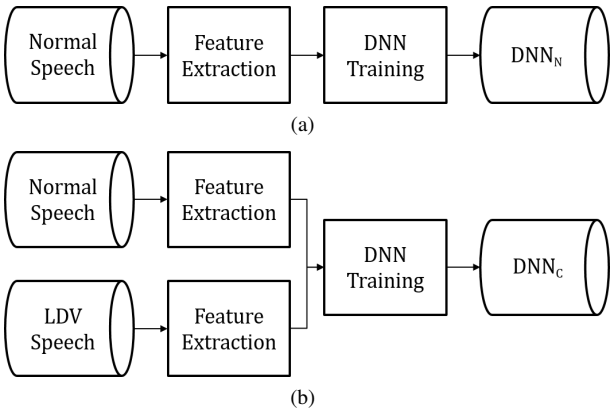

Figure 1: *(a) DNN_N is trained using traditional acoustic features from normal speech, (b) DNN_C is trained using a combination of traditional features and LDV features.*

3. LDV Feature Generation and Combination With a Large Dataset

While the incorporation of LDV features will be shown in Section 4 to improve recognition performance, overall accuracy of both DNN_N and DNN_C is restricted by the small size of the LDV database (i.e. the availability of data that contains parallel recordings of acoustic speech and LDV signals) such that the DNN-based ASR systems are not trained sufficiently well. We therefore aim to make use of much larger datasets, and will test this with **acoustic-only ASR** at first. In particular, we will use a large scale dataset of acoustic recordings of common conversations in moving vehicles gathered by the iFlytek company, which we named the CZ speech corpus (from the initials of the Mandarin phrase meaning 'in car'), which has the same spoken environment as the LDV dataset, although the style of conversations and content between the two datasets are totally different, described in detail in Section 4.1.

Considering the mismatch between CZ and the LDV datasets, instead of using RBM and CD algorithms to pre-train the DNN acoustic model as normal, we first train an acoustic-only DNN-based ASR system from the CZ database alone. This provides a good initialisation start point, i.e. a well-trained DNN. This DNN is then fine-tuned by using the acoustic-only data from the LDV dataset (LDV-acoustic). The resulting acoustic-only ASR system is named DNN_{LN} and is shown in Figure 2. We now extend this to **acoustic + LDV feature AS-**

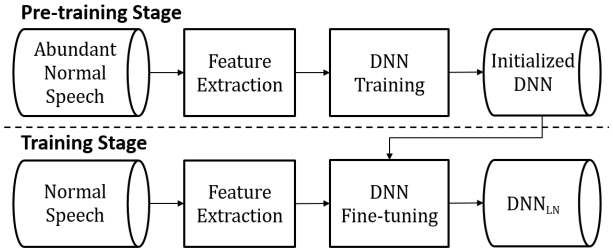

Figure 2: *Structure of pre-training the acoustic-only DNN_{LN} with a large dataset used for initialization.*

R which combines LDV and acoustic features together. Since the large CZ dataset only contains acoustic speech recordings, we obtain corresponding pseudo-LDV features by first training a mapping network to learn the relationships between the features from acoustic speech and the features from the LDV signal. For mapping, we use a regression DNN, shown in Figure 3(a), learning the relationship between normal acoustic features and LDV features. The training procedure of regression DNN is similar to that in [21].

Once the DNN mapping network is ready, we can obtain pseudo-LDV features by mapping from the normal acoustic features extracted from the CZ dataset, which is shown in Figure 3(b). The mapping is only required during pre-training stage. Then we merge these two features together and use them to train a DNN model in the normal way. Once the pre-training stage finished, the initialized model is transferred to the next stage for training. In the training stage, we use data from the LDV dataset, which includes the LDV signal and acoustic speech recordings in parallel, hence the mapping network is not required. In operation, the two types of features are merged just like in the DNN_C system described in Section 2. The resulting DNN, referred to DNN_{LC} ('L' for large scale, 'C' for combined features), will be evaluated with the other systems in the following section.

4. Experiments and Results

4.1. Corpus

In this paper we make use of two independent speech corpora: the LDV dataset gathered by VocalZoom company *, which includes speech recordings captured by LDV sensors along with corresponding acoustic recordings. The second database we use comes from the iFlytek company research group †, which

*http://vocalzoom.com
†http://www.iflytek.com

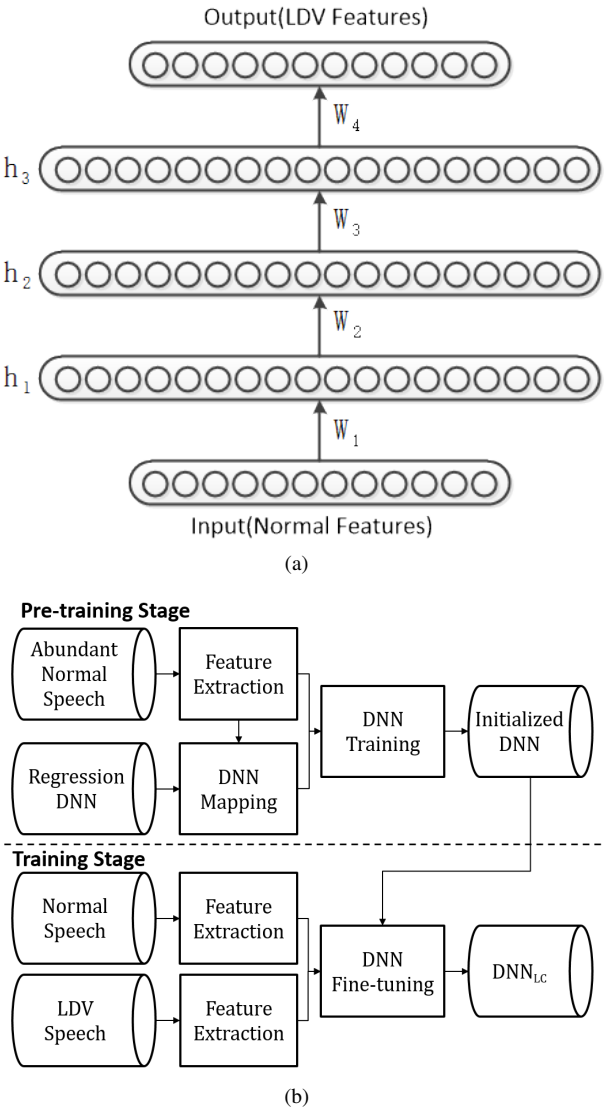

Figure 3: *(a) Training a DNN mapping network, (b) Training DNN_{LC} with a large dataset used for initialization and LDV dataset used for fine tuning.*

provides a large resource recordings of native English speakers. This is used to pre-train the DNN acoustic models.

4.1.1. LDV dataset

The LDV dataset contains 13 thousand recordings in total at a sample rate of 16 kHz with 16-bit accuracy. Speakers use mainly United States English and Hebrew to utter a selection of common sentences from daily life, such as "I see, that is a problem". Some human-to-machine style sentences are also included, especially in cars, such as "FM ninety five point three". In practice, the LDV sensor is directed to a speaker's throat region at a certain distance and measures its vibration velocity, like vocal-fold vibrations. During recording, besides capturing in a clean environment, recordings were made where interfering acoustic noise was present. In those recordings an undesired speaker and background noises (from a moving vehicle) are present in addition to the desired speaker. Measurements by the LDV and acoustic sensors were recorded simultaneous. Detailed information about the data recordings can be found in [16]. For DNN training, the LDV corpus was partitioned into: *training set* consisting of data from 54 speakers for a total duration of 9.9 hours; *development set* consisting of data from 4 speakers for a total duration of 0.62 hours; *testing set* also consisting of data from 4 speakers for a total duration of 0.75 hours.

4.1.2. Large CZ dataset

The CZ corpus contains more than 66 thousand recorded sentences over a total duration of 620 hours, which is much larger than the LDV dataset. This imbalance might make the acoustic DNN model largely influenced by the CZ dataset when combing the two datasets. However, the CZ dataset has a better coverage of pronunciations and speakers as a supplement to the LDV dataset. Similarly, all files were also recorded at a sampling rate of 16 kHz with 16-bit accuracy, which is matched with the LDV data. Native speakers from USA (133 speakers), Canada (78 speakers) and England (26 speakers) were asked to speak some conversations in three common environments relating to; *cars*, including some commands to machines, some names and locations recorded in vehicles; *tourism*, including shopping-related utterances, numbers and the names of famous tourist attractions; *daily communications* involving education, catering and health-care conversations. These were recorded first into high-quality audio files, then replayed in three different vehicles, namely Toyota, Volkswagen and BMW cars, in 5 different scenarios, shown in Table 1. The dataset is named CZ after the initials of the phrase 'in-car' in Mandarin Chinese. The 'Outside' column details the environment that the car is parked in or moving through, while the 'AC' column indicates whether the air conditioner is operating, either on a medium setting or turned off.

Table 1: *Detailed information of* 5 *scenes used for recording within the CZ corpus.*

No.	Car Speed	Window	Outside	AC
1	stationary	closed	downtown	middle
2	stationary	open	car park	off
3	$\leq$ 40km/h	closed	downtown	off
4	41 $-$ 60km/h	closed	countryside	middle
5	80 $-$ 120km/h	closed	highway	middle

4.2. Experimental settings

The features we use for both DNN regression and acoustic modeling are 72-dimensional LMFB features (24-dimensional static LMFB features with Δ and $\Delta\Delta$) and include an input context of 10 neighbouring frames (± 5) yielding a final dimensionality of 792 (72×11). Furthermore, when combined the two LMFB feature vectors of normal speech and LDV speech, a merged acoustic feature vector is with dimensionality of 1584($72 \times 2 \times 11$).

To train the regression DNN, we use the 792-dimensional LMFB features of normal speech as input to learn the targeting LDV features with the same dimension. There are 2 hidden layers with 2048 hidden units in each layer and a final linear output layer, i.e. a structure of 792-2048-2048-792.

The DNN acoustic model uses a regular structure with 6 hidden layers having 2048 hidden units in each layer and a final softmax output layer with 9004 units, corresponding to the senones of the HMM system. For DNN_N and DNN_C systems, the networks were initialized using layer-by-layer generative pre-training using $6, 5, 5, 5, 5, 5$ iterations of the BP algorithm in each layer. As for DNN_{LN} and DNN_{LC}, they were initial-

ized from a well trained DNN using the large scale CZ dataset and combined LMFB features of two signals respectively. In all experiments, the decoding is performed by using a 3-gram language model (LM) with a dictionary consisting of more than 240 thousand words of native English.

4.3. LDV feature combination

The recognition performance is evaluated by word error rate (WER in %) and the sentence error rate(SER in %). Table 2 lists a performance comparison of the two systems with or without using the combined auxiliary features from the LDV sensors. The only difference of DNN_N and DNN_C is the input feature dimension, namely 72 versus 144 for one frame. Both the WER and SER of feature combination DNN_C system can be reduced by about 6% over the DNN_N system using normal speech, which verifies the effectiveness of the auxiliary LDV features.

Table 2: *Results of LDV feature combination*

System	Feature_dim	SER	WER
DNN_N	72	89.71%	58.88%
DNN_C	144	84.23%	52.42%

To further explore the effectiveness of using LDV information in different environments, we test those two systems on two subsets of utterances recorded in clean and noisy environments, as shown in Table 3. From the results, we can make an observation that the auxiliary LDV features can improve the recognition performances for both clean and noisy environments, with relative word error rate (WER) reductions of 11.5% and 12.5%, respectively.

Table 3: *Results of LDV feature combination in different environment conditions*

System		Feature_dim	SER	WER
DNN_N	clean	72	89.64%	56.44%
	noisy	72	93.43%	71.96%
DNN_C	clean	144	81.07%	49.96%
	noisy	144	88.89%	62.93%

All the above results indicate that the LDV signal can provide more useful discriminative information in addition to the normal speech, which can boost the ASR system in all environments.

4.4. LDV feature combination with a large dataset

The results of the systems initialized by the large CZ dataset are shown in Table 4. With more training data, the DNN_{LN} system using acoustic-only features significantly outperforms DNN_N system in Table 2, with the WERs from 58.88% to 32.93%. The DNN systems initialized from the large CZ dataset in the pre-training stage always perform better, irrespective of whether the LDV features are used. Moreover, by the comparison of DNN_{LN} with DNN_{LC}, the use of LDV features achieves a relative WER reduction of 20.6%, which is even more significant than that under the smaller LDV dataset with all real LDV features in Table 2. This implies that the LDV features are potentially more powerful with larger training data even with the pseudo-LDV features generated from the regression DNN with the relationship learned on a small stereo data set of both the normal speech and LDV data.

The system in Table 4, denoted as joint-DNN_{LC}, is a modified version of DNN_{LC} where the training data used for DNN initialization in the pre-training stage includes both the LDV and CZ datasets. A remarkable performance gain is achieved by joint-DNN_{LC} over DNN_{LC}, which indicates that more diversified data in the pre-training stage is always helpful. However, this gain is not significant as the proportion of LDV dataset is too small compared with the large CZ dataset.

Table 4: *Results of the systems with the large CZ dataset for DNN initialization.*

System	Feature_dim	WER
DNN_{LN}	72	32.93%
DNN_{LC}	144	26.13%
joint-DNN_{LC}	144	25.22%

Finally, to give the reader a better understanding of the differences between the LDV and CZ datasets, two more experiments are designed. First, if the test set of LDV-acoustic data is directly evaluated by the pre-trained model using CZ dataset as in Figure 2, the recognition performance is extremely poor, which confirms that those two datasets are quite different in speaker styles, speech contents, etc. Second, when the pre-trained model of joint-DNN_{LC} system is adopted for testing, WER is 37.04%, which performs much better than DNN_C with the WER of 52.42%. From the two experiments, we can make an interesting observation that the recognition performance is not satisfactory when the model is trained on each dataset (LDV or CZ) separately while the model trained with two datasets merged can yield a very significant improvement of recognition accuracy, which implies the two datasets are strongly complementary in terms of the coverage of speaker styles and speech contents.

5. Conclusions

In this paper, we have investigated the use of auxiliary information derived from an LDV sensor for improving ASR performance. Due to the properties of LDV data which make it immune to acoustic interference, we combine LDV features with normal acoustic speech features to train a DNN acoustic model. Experimental results show significant improvements of recognition accuracy under both clean and noisy conditions. Furthermore, after pre-training the DNN model with pseudo-LDV features combined with acoustic features extracted from a large data set, ASR system achieves much better performance than that trained with smaller LDV datasets alone.

The good performances showed above promise the LDV sensor a bright future with much protential applications. Company like VocalZoom has already deliverd some solutions including Voice-Controlled Driver Assistance, Voice-Controlled Smartglasses and even Voice Authentication for personal securities. We also note that researchers are on the way aiming to develop a much smaller and more convenient laser-based sensor, which can make it get rid of the inconvenience of heavy equipments and will be more suitable for practical use.

6. Acknowledgements

This work was partially funded by the National Nature Science Foundation of China under Grant No. 61305002, National Key Technology Support Program under Grants No. 2014BAK15B05, the Strategic Priority Research Program of the Chinese Academy of Sciences under Grant No. XDB02070006, and MOE-Microsoft Key Laboratory of USTC.

7. References

[1] K. Han, Y. He, D. Bagchi, E. Fosler-Lussier, and D. Wang, "Deep neural network based spectral feature mapping for robust speech recognition," in *Proc. Interspeech*, 2015, pp. 2484–2488.

[2] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Springer, 1993, vol. 201.

[3] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.

[4] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–33, 2014. [Online]. Available: http://research.microsoft.com/apps/pubs/default.aspx?id=208124

[5] J. Du, Q. Wang, T. Gao, Y. Xu, L.-R. Dai, and C.-H. Lee, "Robust speech recognition with speech enhanced deep neural networks." in *Proc. Interspeech*, 2014, pp. 616–620.

[6] D. Bagchi, M. I. Mandel, Z. Wang, Y. He, A. Plummer, and E. Fosler-Lussier, "Combining spectral feature mapping and multi-channel model-based source separation for noise-robust automatic speech recognition," in *Proc. IEEE ASRU*, 2015.

[7] B. E. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech communication*, vol. 25, no. 1, pp. 117–132, 1998.

[8] 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 *Signal Processing Conference, 2010 18th European*. IEEE, 2010, pp. 1978–1982.

[9] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and X. Huang, "Direct filtering for air-and bone-conductive microphones," in *Multimedia Signal Processing, 2004 IEEE 6th Workshop on*. IEEE, 2004, pp. 363–366.

[10] N. Radha, A. Shahina, G. Vinoth, and A. N. Khan, "Improving recognition of syallabic units of hindi languagae using combined features of throat microphone and normal microphone speech," in *Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014 International Conference on*. IEEE, 2014, pp. 1343–1348.

[11] J. Breguet, J.-P. Pellaux, and N. Gisin, "Photoacoustical detection of trace gases with an optical microphone," in *10th Optical Fibre Sensors Conference*. International Society for Optics and Photonics, 1994, pp. 457–460.

[12] M. De Paula, A. De Carvalho, C. Vinha, N. Cella, and H. Vargas, "Optical microphone for photoacoustic spectroscopy," *Journal of applied physics*, vol. 64, no. 7, pp. 3722–3724, 1988.

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

[14] R. L. Goode, G. Ball, S. Nishihara, and K. Nakamura, "Laser Doppler vibrometer (LDV)." *Otology & Neurotology*, vol. 17, no. 6, pp. 813–822, 1996.

[15] W. Li, M. Liu, Z. Zhu, and T. S. Huang, "Ldv remote voice acquisition and enhancement," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 4. IEEE, 2006, pp. 262–265.

[16] Y. Avargel and I. Cohen, "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*. IEEE, 2011, pp. 109–114.

[17] Y. Avargel, T. Bakish, A. Dekel, G. Horovitz, Y. Kurtz, and A. Moyal, "Robust speech recognition using an auxiliary laser-doppler vibrometer sensor," in *Proc. Speech Process, Conf., Tel-Aviv, Israel*, 2011.

[18] G. E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[19] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[20] D. Yu, M. L. Seltzer, J. Li, J.-T. Huang, and F. Seide, "Feature learning in deep neural networks-studies on speech recognition tasks," *arXiv preprint arXiv:1301.3605*, 2013.

[21] T. Gao, J. Du, L. R. Dai, and C. H. Lee, "Joint training of front-end and back-end deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.