J Sign Process Syst
https://doi.org/10.1007/s11265-017-1287-x

Auxiliary Features from Laser-Doppler Vibrometer Sensor for Deep Neural Network Based Robust Speech Recognition

Lei Sun[1] · Jun Du[1] · Zhipeng Xie[2] · Yong Xu[3]

Received: 4 February 2017 / Revised: 26 August 2017 / Accepted: 18 September 2017

Abstract Recently, the signals captured from a laser Doppler vibrometer (LDV) sensor have shown the noise robustness to automatic speech recognition (ASR) systems by enhancing the acoustic signal prior to feature extraction. In this study, an alternative approach, namely concatenating the auxiliary features extracted from the LDV signal with the conventional acoustic features, is proposed to further improve ASR performance based on the deep neural network (DNN) for acoustic modeling. The preliminary experiments on a small set of stereo-data including both LDV and acoustic signals demonstrate its effectiveness. Thus, to leverage more existing large-scale speech databases, a regression DNN is designed to map acoustic features to LDV features, which is well trained from a stereo-data set with a limited size and then used to generate pseudo-LDV features from a massive speech data set for parallel training of an ASR system. Our experiments verify that both the features from the limited scale LDV data set as well as the massive scale pseudo-LDV features can yield significant improvements of recognition performance over the system using purely acoustic features, in both quiet and noisy environments.

Keywords Laser Doppler vibrometer · Auxiliary features · Deep neural network · Regression model · Speech recognition

1 Introduction

As one of the most practical and valuable applications of machine learning and artificial intelligence, constructing a stable and easy-to-use automatic speech recognition (ASR) system is always a hot issue in the past decades. Previously, conventional hidden Markov model (HMM)-based [1] speech recognizers have been commonly used with each acoustic state modeling by a Gaussian mixture model (GMM), referred to as a GMM-HMM system. However, GMMs have a serious shortcoming that they are statistically inefficient for modeling data that lie on or near a nonlinear manifold in the data space [2]. On the contrary, artificial neural networks trained by back-propagating error derivatives have the potential to learn much better models of data than GMMs. Meanwhile, deep neural networks (DNNs) have demonstrated a great capacity to extract discriminative internal representations that are robust to the many sources of variability in speech signals. Over the last few years, efficient learning methods [3–5] and speed-up hardwares boost DNN-HMM structure to replace GMM-HMM in speech recognition. So far, with great mass of data, ASR system based on DNN-HMM outperforms previous methods by a large margin and becomes the mainstream method.

Besides those laboratorial success, an available ASR system should be able to handle not only clean conditions but also complex noisy scenes. In fact, most ASR systems

✉ Jun Du
jundu@ustc.edu.cn

Lei Sun
sunlei17@mail.ustc.edu.cn

Zhipeng Xie
xzp2013@mail.ustc.edu.cn

Yong Xu
yx0001@surrey.ac.uk

[1] University of Science and Technology of China, No. 96, JinZhai Road, Hefei, Anhui, People's Republic of China

[2] iFlytek Research, iFlytek Co., Ltd., Hefei, China

[3] University of Surrey, Guildford GU2 7XH, UK

Published online: 27 September 2017

are likely to suffer severe performance degradation when speech is mixed with some unavoidable interrupting factors like environment noise, room reverberation, disturbances from different microphones and recording non-linearities [6]. Hence, robust speech recognition under real conditions is still a knotty problem to be solved.

To solve these problems, many processing techniques were proposed and could be divided into several categories [7–9]. First, a batch of speech enhancement methods to improve speech signal quality [10] thus can help ASR system which is well trained in relatively clean conditions. But to attain the best enhancement performance, these enhancement algorithms should be customized under certain conditions, such as noisy or reverberation conditions, single-channel or multi-channel scenes. There are also investigations focusing on finding robust acoustic features [11, 12], which emphasize the temporal structure in speech. Some open challenges are held such as REVERB (REverberant Voice Enhancement and Recognition Benchmark) challenge [13], CHiME challenge [14], ASpIRE (Automatic Speech Recognition In Reverberant Environments) [15] in order to bring more efforts to improve robust speech recognition.

Recently, auxiliary information gathered from non-acoustic sensors like bone-, throat- and air- microphones is shown to be effective for ASR systems to make better decisions under noisy environments [16–18]. Photo-acoustic technique shows promising results on robust recognition due to their inherent immunity to acoustic noise as well as non-contact operation [19, 20]. Combining traditional acoustic features with speech information captured by these sensors, recognition performances are further improved [21].

Laser doppler vibrometer (LDV) sensor [22, 23] is a kind of non-contact measurement device that is capable of measuring the vibration frequencies of moving targets. When it is directed to a speaker's larynx, it captures valuable speech information at certain frequency bands. Due to its insensitivity to environments, LDV signal is appropriate as a supplement to strengthen the robustness of ASR system. In [23, 24], LDV sensors are presented as making accurate and reliable voice activity detection (VAD) decision, as well as improving the speech recognition results.

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 [24], the main difference is we directly use LDV features for acoustic modeling while in [24] 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 [24]. In this paper, all experiments are conducted on two datasets, a small LDV dataset which contains both normal acoustic signals and LDV signals, the other is a large recorded dataset contains only normal acoustic signals. Log Mel-filter-bank (LMFB) features are extracted from all datasets and are used for DNN modeling. First, we train a baseline acoustic-only model (denoted as $\mathrm{DNN_N}$) on LDV dataset using only acoustic features. Then we concatenate the normal acoustic features with LDV features, and train another acoustic-LDV model (denoted as $\mathrm{DNN_C}$). The latter system yields a relative 10.97% reduction of word error rate (WER) which demonstrates the effectiveness of using LDV features.

Due to the limited size of existing LDV datasets, DNN models may not be trained sufficiently. Hence, we use the large dataset to pre-train the network as a better initialization for DNN fine-tuning. Because the large dataset lacks of LDV signals, to construct an acoustic-LDV system we consider obtaining LDV features by converting normal acoustical features from a large dataset into pseudo-LDV features. Accordingly, we create 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, yielding a very well trained DNN-based ASR system with dual input features. The promoting performance of recognition accuracies confirms the validity of both initialization DNN and feature-mapping network.

This work is an extension of the recently disclosed version [25] with the new contributions: (1) more comprehensive introduction of LDV signals, (2) more technical details of feature mapping DNN, (3) the clearer training procedure for each system, (4) more experiments and analysis.

The rest of the paper is organized as follows. Section 2 introduces the details of LDV signal. In Section 3, we propose the auxiliary LDV features for building an ASR system. Section 4 demonstrates the use of feature-mapping DNN to derive pseudo-LDV features from a large dataset, and then use them to jointly optimize the DNN network. Section 5 gives the experimental conditions, datasets, system operation and discusses results. Finally we conclude the paper in Section 6.

2 Details of LDV Signal

Unlike traditional sound pick-up equipments, an LDV sensor is a non-contact measurement device which can detect the vibration signal of a moving target based on the principle of interferometry, as illustrated in Fig. 1. It is composed of several major parts, including a laser, a Bragg cell, a photo detector, an optical lens, and many beam splitters (BS), etc. First a laser beam with frequency f_0 is emitted from the laser, then divided into a reference beam and an object beam by BS1. Along a straight line through BS2 and the lens, the object beam can arrive at the target vibrated object (speaker's larynx). The corresponding backscattered beam

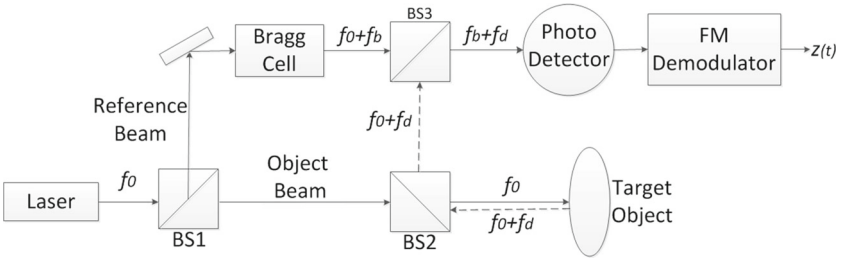

Figure 1 Basic components of a laser Doppler vibrometer.

with a Doppler shift f_d will be reflected to BS3. Meanwhile, the reference beam passes through a Bragg cell, which produces a frequency shift of f_b. After BS3, the two beams are mixed together to generate a signal with frequency shift of $f_b + f_d$, which is then converted to a voltage signal by a photo-detector. Finally, an FM-demodulator outputs the signal $z(t)$ with f_b and f_d respectively being its carrier and modulated frequencies. More details can be found in [23, 24]. Figure 2 shows the spectrograms of an acoustic signal and corresponding LDV signal of the same utterance at a sample rate of 16kHz.

Obviously, the LDV signal is a sensitive signal to vibration of the object target. Simultaneously, to those acoustical disturbances which severely degrade ASR performance, LDV signals have instinctive robustness. Thus, LDV signals could be strongly complementary with the traditional acoustic signals. In Fig. 2, the LDV signal mainly contains the information at low-frequency (up to 3 kHz) and is shown to be robust to environmental disturbances in comparison the normal acoustic signal. In [23], speech enhancement algorithm based on LDV signals was proposed to attenuate

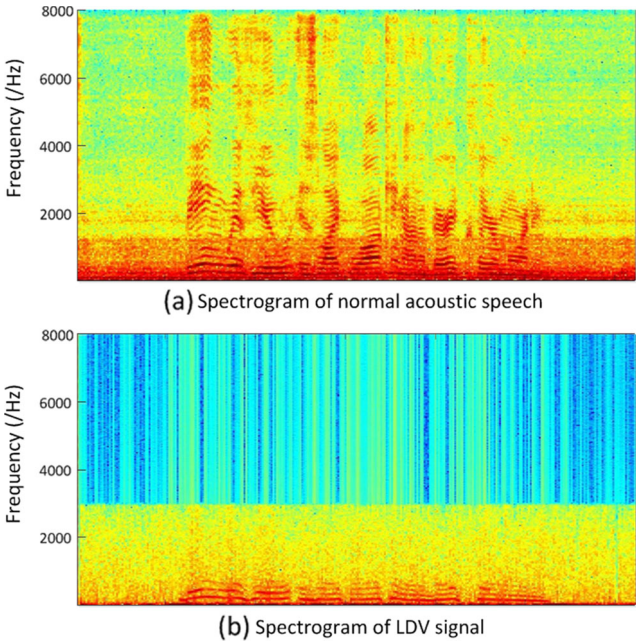

Figure 2 Comparison between two kinds of signals: **a** spectrogram of normal speech. **b** spectrogram of LDV signal.

a kind of random impluses named speckle noise in [26] which would limit the applicability of LDV-based measurement devices. Then a soft-decision VAD was derived in time-frequency domain and the gain function of the optimally-modified log-spectral amplitude (OM-LSA) was appropriately modified. Despite the effectiveness of z highly non-stationary noises, this method often fails to retain weak-speech components and may lead to the degradation of speech quality [24]. Rather than reducing the speckle noises, VAD strategy in [24] ignored them by detecting spectral harmonic patterns. Similarly, the resulting VAD information was used to improve OM-LSA performance under low SNR conditions. It's demonstrated that the proposed ASR system using LDV information can substantially improve recognition accuracies. Unlike this method, in this study we directly extract the features of LDV signals and combine with traditional acoustic features for acoustic modeling.

3 The Auxiliary LDV Features

In this section, we exploit the concatenation of auxiliary LDV features with acoustic features, in comparison to acoustic-only speech features for acoustic modeling. Both systems are built on our LDV dataset consisting of parallel acoustic microphone and LDV data for each sentence.

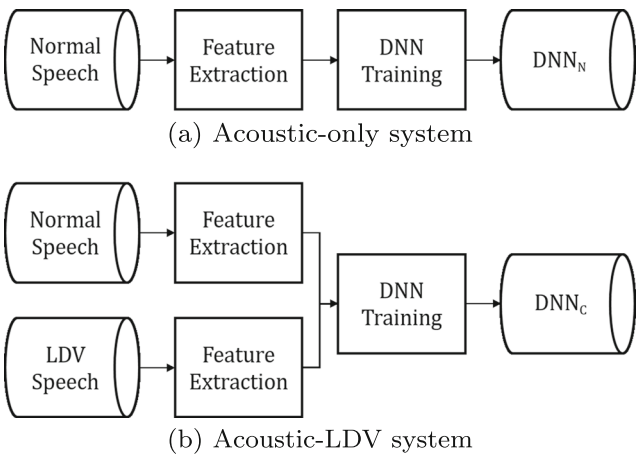

Figure 3 **a** DNN_N is trained using traditional acoustic features from normal speech, **b** DNN_C is trained using a combination of traditional acoustic features and LDV features. Both of them are trained on LDV dataset.

Figure 3 shows the difference between two DNN-based systems. The acoustic-only baseline system DNN_N uses acoustic features extracted from normal speech. Correspondingly, the acoustic-LDV system DNN_C introduces LDV features to be concatenated with the acoustic features. The acoustic-only approach is to obtain the LMFB features from normal speech and feed them into the DNN input layer with adjacent context frames. However, our proposed acoustic-LDV approach combines the LMFB features of normal speech with the LMFB features extracted from the LDV signals. The two types of LMFB features have the same dimension. Suppose an acoustic-only feature vector is with the dimension of D. After concatenation, the merged feature vector is with the dimension of $2D$. During the training process, we introduce pre-training methods to better initialize the parameters prior to back propagation (BP), preventing being stuck in poor local optima. Contrastive divergence (CD) criterion is used to train each pair of layers in the networks as restricted Boltzmann machines (RBM) and grow the network layer-by-layer in an unsupervised way [27]. To better illustrate these procedures, we rewrite them in Algorithm 1.

Algorithm 1 Training procedure for acoustic-only and acoustic-LDV systems

Step1:**Extract two types of features**

1. Extract acoustic features of normal speech utterances in LDV dataset.
2. Extract features of the corresponding LDV signals in LDV dataset.
3. Concentrate normal acoustic features and LDV features in the frame-level.

Step2:**Train DNN-based acoustic models**

1. Use layer-by-layer generative pre-training algorithm to initialize parameters in each layer with normal acoustic features, we refer it as DNN_N.
2. Use layer-by-layer generative pre-training algorithm to initialize parameters in each layer with concatenated features, we refer is as DNN_C.
3. Fine-tune DNN_N and DNN_C with the corresponding features, respectively.

4 Extension to a Large Dataset

4.1 Acoustic-only ASR with a Large Dataset

While the use of LDV features is shown to improve recognition performance in Section 5, the overall accuracies of both DNN_N and DNN_C are still quite low as the DNN-based ASR systems are not trained sufficiently well with the small size of the LDV dataset (i.e. the availability of stereo-data containing both recordings of acoustic speech and LDV signals). We therefore aim to make use of much larger datasets. Specifically, a large scale dataset including the recordings of conversational speech in moving vehicles collected by the iFlytek company [28] is adopted, denoted as the CZ speech corpus (from the initials of the Mandarin phrase meaning 'in car'). Although both CZ and LDV datasets are recorded in similar car environments, the speaking styles and contents are totally different, which will be elaborated in Section 5.1.

Considering the mismatch between CZ and the LDV datasets, instead of using RBM and CD algorithms to pre-train the DNN acoustic model, we first train an acoustic-only DNN model from the CZ database alone, which is then fine-tuned by using the acoustic-only data from the LDV dataset. As shown in Fig. 4, the resulting acoustic-only ASR system is named as DNN_{LN}. To better illustrate the procedure, we rewrite them in Algorithm 2.

Algorithm 2 Training procedure for the acoustic-only system with a large dataset

Step1:**Use a large dataset for pre-training**

1. Extract acoustic features of normal speech utterances from the large dataset.
2. Use layer-by-layer generative pre-training algorithm to initialize parameters in each layer, and then fine-tune the DNN model.

Step2:**Use LDV dataset for fine-tuning**

1. Extract acoustic features of normal speech utterances from the LDV dataset.
2. Use the well-trained DNN model in **Step1** as initialization, fine-tune the whole network, we refer it as DNN_{LN}.

4.2 The Feature Mapping Network

Since the large CZ dataset only contains acoustic speech recordings, we obtain corresponding pseudo-LDV features by first training a mapping network. For mapping, we use a regression DNN, shown in Fig. 5, learning the relationship between normal acoustic features and LDV features from the LDV dataset. In detail, multiple context frames

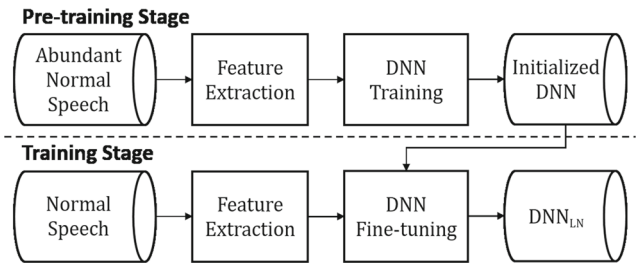

Figure 4 Pre-training the acoustic-only DNN_{LN} with a large dataset for initialization.

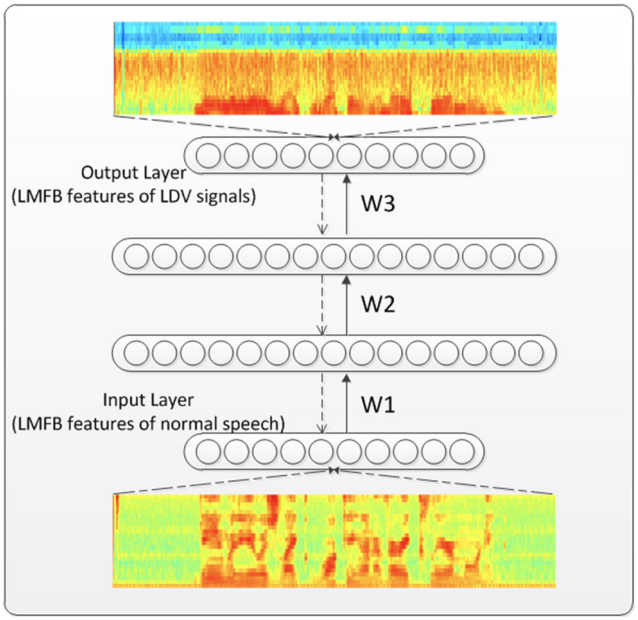

Figure 5 Structure of feature-mapping DNN converting acoustic features to LDV features.

of normal acoustic features are combined together as the DNN input. Mean squared error (MSE) is used as the training criterion to minimize the difference between output pseudo-LDV features and reference LDV features:

$$E = \frac{1}{N}\sum_{n=1}^{N} \|\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}(\boldsymbol{y}_{n-\tau}^{n+\tau}, \boldsymbol{W}, \boldsymbol{b}) - \boldsymbol{x}_{n-\tau}^{n+\tau}\|_2^2 + \lambda\|\boldsymbol{W}\|_2^2 \quad (1)$$

where $\hat{\boldsymbol{x}}_{n-\tau}^{n+\tau}$ and $\boldsymbol{x}_{n-\tau}^{n+\tau}$ are the n^{th} $D(2\tau+1)$-dimensional vectors of estimated and reference LDV features, respectively. $\boldsymbol{y}_{n-\tau}^{n+\tau}$ is a $D(2\tau+1)$-dimensional vector of input acoustic features with neighboring left and right τ frames as the acoustic context. $\boldsymbol{W}$ and $\boldsymbol{b}$ denote all the weight and bias parameters. λ is the regularization weighting coefficient to avoid overfitting. The objective function is optimized using back-propagation with a stochastic gradient descent method in mini-batch mode of N sample frames. The training procedure of regression DNN is similar to that in [29].

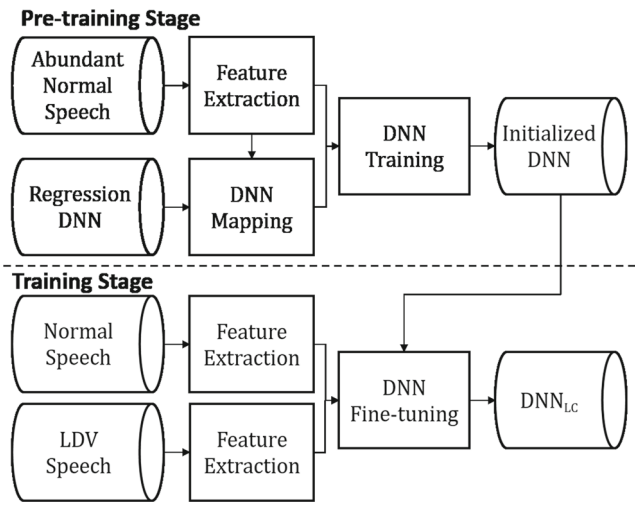

Figure 6 Training DNN_{LC} with a large dataset used for initialization and LDV dataset used for fine-tuning.

4.3 Acoustic-LDV ASR with a Large Dataset

Once the DNN mapping network is generated, we can obtain pseudo-LDV features by mapping the normal acoustic features extracted from the CZ dataset, which is shown in Fig. 6. The feature mapping is only conducted during the pre-training stage. Then we merge the acoustic and pseudo-LDV features to pre-train a DNN model in a similar way to Fig.4. Next, the pre-trained model is transferred to the next stage as the initialization. In the training stage, we use data from the LDV dataset, including the parallel LDV signals and acoustic speech recordings, hence the mapping network is not required. The two types of features are concatenated just like the DNN_{C} system described in Section 3. The resulting DNN, referred to DNN_{LC} ('L' for large scale, 'C' for combined features), will be evaluated in the recognition stage. To better illustrate these procedures, we rewrite them in Algorithm 3.

Algorithm 3 Training procedure for the acoustic-LDV system with a large dataset

Step1:**Use the LDV dataset for mapping DNN**

1. Extract acoustic features of normal speech utterances in LDV dataset.
2. Extract features of corresponding LDV signals in LDV dataset.
3. Train a feature mapping network with acoustic-LDV feature pairs with randomly initialized parameters under MSE criterion.

Step2:**Use the large CZ dataset for pre-training**

1. Extract normal acoustic features from the large CZ dataset.
2. Generate pseudo-LDV features by the feature-mapping DNN trained in **Step1**.
3. Concentrate both normal acoustic features and pseudo-LDV features as DNN inputs to train a DNN acoustic model with generative layer-by-layer pre-training algorithm.

Step3:**Use the LDV dataset for fine-tuning**

1. Extract acoustic features of normal speech utterances in LDV dataset.
2. Extract features of corresponding LDV signals in LDV dataset.
3. Use the well-trained DNN model in **Step2** as the initialization, fine-tune the whole network with concatenated features, we refer it as DNN_{LC}.

5 Experiments and Results

5.1 Corpus

Two independent speech corpora are adopted for the experiments. The first one is the LDV dataset collected by VocalZoom company [30], which includes speech recordings captured by LDV sensors along with corresponding acoustic recordings. The second one comes from the iFlytek company [28], which provides a much larger resource of recordings of native English speakers to pre-train the DNN acoustic models.

5.1.1 LDV Dataset

The LDV dataset contains 13 thousand recordings in total at a sample rate of 16 kHz. 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. Besides capturing in a quiet environment, recordings were also made where interfering acoustic noises were present. In those recordings undesired speakers and background noises (from a moving vehicle) were presented in addition to the desired speaker. Measurements by the LDV and acoustic sensors were recorded simultaneously. More details can be found in [23]. For system training and evaluation, the LDV corpus was partitioned into: *training set* consisting of data from 54 speakers with a total duration of 9.9 hours; *development set* consisting of data from 4 speakers with a total duration of 0.62 hours; *testing set* also consisting of data from 4 speakers with a total duration of 0.75 hours.

5.1.2 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. Similarly, all files were recorded at a sample rate of 16 kHz. Native speakers from USA (133 speakers), Canada (78 speakers) and England (26 speakers) were asked to conduct conversations in three common environments relating to: *cars*, including the commands to machines, the 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, with 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.

5.2 Experimental Settings

The features we use for both DNN-based feature mapping 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 neighboring frames (± 5) yielding a final dimensionality of 792 (72×11). Furthermore, when combining the two LMFB feature vectors of normal speech and LDV signal, a merged acoustic feature vector is with dimensionality of 1584($72 \times 2 \times 11$).

To train the regression DNN, we use 792-dimensional LMFB features of normal speech as input to learn the target 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 soft-max 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 with 6,5,5,5,5,5 iterations of the BP algorithm in each layer. As for DNN_LN and DNN_LC, they were initialized 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.

5.3 The Effectiveness of Using LDV Features

The recognition performance is evaluated by word error rate (WER in %) and 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

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

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%

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 improved by about 6% over the DNN_N system using normal speech, which verifies the effectiveness of the auxiliary LDV features.

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 WER reductions of 11.5% and 12.5%, respectively.

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.

5.4 Extension to a Large Dataset

5.4.1 The Validity of Feature Mapping

Figure 7 is an illustration which is slightly modified in color gamut to better exhibit the validity of feature mapping network. A sentence from LDV dataset is selected. As we can see, the features of normal acoustic speech and LDV signal are totally different not only in intensity but also in distribution. After feature mapping, the generated pseudo LDV features in Fig. 7c can roughly imitate the profile of real LDV features in Fig. 7b. Limited by the small size of LDV dataset, the feature mapping network can not be perfectly trained. From another perspective of preventing overfitting, the data availability in this study is expedient and a larger stereo dataset should be adopted in the future.

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%

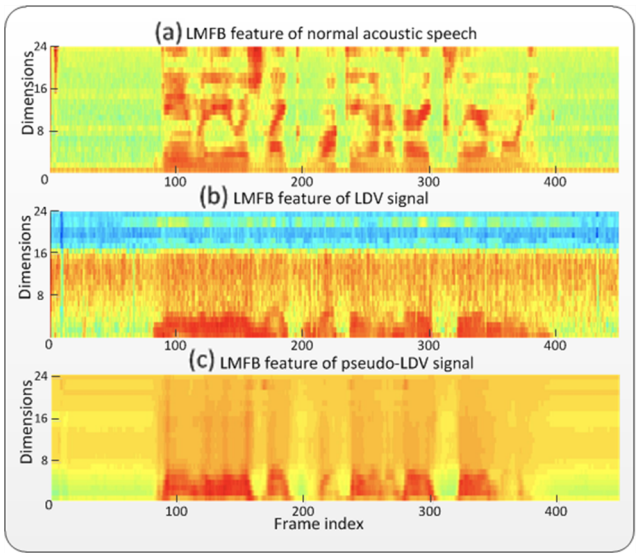

Figure 7 *An example of feature mapping network.*

5.4.2 Acoustic-LDV ASR 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 learned on a small stereo dataset 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

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%

helpful. However, this gain is not quite significant as the proportion of LDV dataset is too small compared with the large CZ dataset.

Finally, to give the reader a better understanding of the differences between the LDV and CZ datasets, more experiments are designed. First, as shown in Fig. 2, the auxiliary LDV features extracted from LDV signals are quite different from the conventional acoustic features. If we only use the LDV features to construct an ASR system, the performance is extremely poor with a WER of 93.54%. So in the current ASR framework, LDV features can only be used as auxiliary features. Second, if the test set of LDV-acoustic data is directly evaluated by the pre-trained model using CZ dataset as in Fig. 6, the recognition performance is much worse, which confirms that those two datasets are quite different in speaking styles, speech contents, etc. Third, 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 a WER of 52.42%. From these experiments, we can make an interesting observation that the recognition performance is not satisfactory if 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 data coverage for speaking styles and speech contents.

6 Conclusion

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. In the future, we will find the method to collect more LDV data and try to accelerate the practical progress.

Acknowledgments This work was supported in part by the National Natural Science Foundation of China under Grants 61671422 and U1613211, in part by the National Key Research and Development Program of China under Grant 2017YFB1002200, in part by the MOE-Microsoft Key Laboratory of USTC.

References

1. Baker, J.M., Deng, L., Glass, J., Khudanpur, S., Lee, C.H., Morgan, N., & O'Shaughnessy, D. (2009). *IEEE Signal Processing Magazine, 26(3)*.
2. Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al. (2012). *IEEE Signal Processing Magazine, 29*(6), 82.
3. Hinton, G.E. (2002). *Neural Computation*, *14*(8), 1771.
4. Hinton, G.E., Osindero, S., & Teh, Y.W. (2006). *Neural Computation*, *18*(7), 1527.
5. Hinton, G. (2010). *Momentum*, *9*(1), 926.
6. Han, K., He, Y., Bagchi, D., Fosler-Lussier, E., & Wang, D. (2015). In *Proceedings of Interspeech* (pp. 2484–2488).
7. Acero, A. (1993). *Acoustical and environmental robustness in automatic speech recognition,* Vol. 201. Berlin: Springer.
8. Gong, Y. (1995). *Speech Communication*, *16*(3), 261.
9. Li, J., Deng, L., Gong, Y., & Haeb-Umbach, R. (2014). *IEEE/ACM Transactions on Audio Speech, and Language Processing*, *22*(4), 745.
10. Du, J., Wang, Q., Gao, T., Xu, Y., Dai, L.R., & Lee, C.H. (2014). In *Proceedings of interspeech* (pp. 616–620).
11. Bagchi, D., Mandel, M.I., Wang, Z., He, Y., Plummer, A., & Fosler-Lussier, E. (2015). In *Proceedings IEEE ASRU*.
12. Kingsbury, B.E., Morgan, N., & Greenberg, S. (1998). *Speech Communication*, *25*(1), 117.
13. Kinoshita, K., Delcroix, M., Yoshioka, T., Nakatani, T., Sehr, A., Kellermann, W., & Maas, R. (2013). In *2013 IEEE workshop on applications of signal processing to audio and acoustics (WASPAA)* (pp. 1–4).
14. Vincent, E., Barker, J., Watanabe, S., Le Roux, J., Nesta, F., & Matassoni, M. (2013). In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 126–130).
15. Harper, M. (2015). In *2015 IEEE workshop on automatic speech recognition and understanding (ASRU)* (pp. 547–554).
16. Dekens, T., Verhelst, W., Capman, F., & Beaugendre, F. (2010). In *2010 18th European signal processing conference* (pp. 1978–1982): IEEE.
17. Liu, Z., Zhang, Z., Acero, A., Droppo, J., & Huang, X. (2004). In *2004 IEEE 6th workshop on multimedia signal processing* (pp. 363–366). IEEE.
18. Radha, N., Shahina, A., Vinoth, G., & Khan, A.N. (2014). In *2014 international conference on control instrumentation, communication and computational technologies (ICCICCT)* (pp. 1343–1348). IEEE.
19. Breguet, J., Pellaux, J.P., & Gisin, N. (1994). In *10th optical fibre sensors conference* (pp. 457–460). International Society for Optics and Photonics.
20. De Paula, M., De Carvalho, A., Vinha, C., Cella, N., & Vargas, H. (1988). *Journal of Applied Physics*, *64*(7), 3722.
21. Graciarena, M., Franco, H., Sonmez, K., & Bratt, H. (2003). *IEEE Signal Processing Letters*, *10*(3), 72.
22. Goode, R.L., Ball, G., Nishihara, S., & Nakamura, K. (1996). *Otology & Neurotology*, *17*(6), 813.
23. Avargel, Y., & Cohen, I. (2011). In *2011 joint workshop on hands-free speech communication and microphone arrays (HSCMA)* (pp. 109–114). IEEE.
24. Avargel, Y., Bakish, T., Dekel, A., Horovitz, G., Kurtz, Y., & Moyal, A. (2011). In *Proceedings speech process, conference*. Israel: Tel-Aviv.

25. Xie, Z., Du, J., McLoughlin, I., Xu, Y., Ma, F., & Wang, H. (2016). In *Proceedings of ISCSLP*.
26. Vass, J., Śmíd, R., Randall, R., Sovka, P., Cristalli, C., & Torcianti, B. (2008). *Mechanical Systems and Signal Processing*, *22*(3), 647.
27. Seltzer, M.L., Yu, D., & Wang, Y. (2013). In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7398–7402). IEEE.
28. [online]. iflytek. http://www.iflytek.com/.
29. Gao, T., Du, J., Dai, L.R., & Lee, C.H. (2015). In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE.
30. [online]. vocalzoom. http://vocalzoom.com/.

Lei Sun received the B.S. degree from the School of Computer and Communication Engineering, Northeastern University at Qinhuangdao in 2015. He is now a Ph.D. candidate at University of Science and Technology of China (USTC), Hefei, China. His current research interests include speech enhancement and robust speech recognition.

Jun Du received the B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC), Hefei, China, in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above period, he worked as an Intern twice for nine months at Microsoft Research Asia (MSRA), Beijing, China. In 2007, he also worked as a Research Assistant for six months in the Department of Computer Science, The University of Hong Kong. From July 2009 to June 2010, he worked with iFlytek Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing of USTC.

Zhipeng Xie is presently a Researcher at iFlytek Corporation, China. He received the B.E degree in communications engineering from Anhui University, China, in 2013. And he received M.Sc degree from the Department of Electronic Engineering and Information Science at the University of Science and Technology of China in 2016. His current research interests include automatic speech recognition and Multimedia information processing.

Yong Xu was born in 1988. He received the Ph.D. degree from the University of Science and Technology of China (USTC), Hefei, China, in 2015, on the topic of DNN-based speech enhancement and recognition. He currently works at the University of Surrey, Guildford, U.K. as a Research Fellow. He once visited Prof. Chin-Hui Lee's lab in Georgia Institute of Technology, USA from September 2014 to May 2015. Prior to his current work, he once also worked in IFLYTEK company from April 2015 to April 2016 to develop far-field ASR technologies.His research interests include deep learning, speech enhancement and recognition, audio and scene classification, etc.