WEIGHTED TRAINING FOR SPEECH UNDER LOMBARD EFFECT FOR SPEAKER RECOGNITION

Muhammad Muneeb Saleem, Gang Liu, John H.L. Hansen

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080, USA
{muneeb,gang.liu,john.hansen}@utdallas.edu

ABSTRACT

The presence of Lombard Effect in speech is proven to have severe effects on the performance of speech systems, especially speaker recognition. Varying kinds of Lombard speech are produced by speakers under influence of varying noise types [1]. This study proposes a high-accuracy classifier using deep neural networks for detecting various kinds of Lombard speech against neutral speech, independent of the noise levels causing the Lombard Effect. Lombard Effect detection accuracies as high as 95.7% are achieved using this novel model. The deep neural network based classification is further exploited by validation based weighted training of robust i-Vector based speaker identification systems. The proposed weighted training achieves a relative EER improvement of 28.4% over an i-Vector baseline system, confirming the effectiveness of deep neural networks in modeling Lombard Effect.

Index Terms— Lombard Effect, deep neural networks, speaker identification, robust, weighted training

1. INTRODUCTION

Lombard Effect is described as a type of stressed speech produced by a speaker when exposed to a noisy environment. This changes neutral speech production in terms of several reported parameters including duration, pitch, intensity, and spectral slope [1]. Lombard Effect in speech data has been shown to severely impact performance of speech systems [1, 2, 3]. Different compensation schemes have been proposed to counter this impact in speech recognition systems [4, 5, 6] and a few for speaker identification (SID) systems [1]. Deep neural networks (DNN) have been proven to work well for speech recognition tasks [7, 8] but have rarely been applied for stressed speech classification [9] or speaker recognition under Lombard Effect.

This study explores the capability of deep neural networks in extracting information from stressed speech under Lombard Effect. Furthermore, we explore the use of this information in building a robust SID system that is resilient towards the effects of background noise in human speech using metadata from the validation phase of DNN training.

2. UT-SCOPE DATABASE

The speech data utilized in this study was drawn from the UT-SCOPE (Speech under COgnitive and Physical stress and Emotion) database. Details about the database can be found in [10]. Speech was collected from speakers under nine different noisy environments. It must be noted that noise was played through open-air headphones so all data is noise-free clean speech. Three noise types were considered: large crowd noise (LCR) at 70, 80, and 90 dB-SPL, highway noise (HWY) in a car at 70, 80, and 90 dB-SPL, and pink noise (PNK) at 65, 75, and 85 dB-SPL. Neutral speech data was also collected from the same speakers for comparative analysis. The speech comprises of 20 phonetically balanced TIMIT sentences, five repetitions of 10 digits, and spontaneous speech. Speech files consist of an average of 3 seconds of data, which makes it challenging for speaker recognition. Subjects included 24 female and 6 male speakers. After randomizing, 75% of the data was used as training and validation, while the rest was used as test data; both for modeling the deep neural network and the speaker identification system.

3. LOMBARD FLAVOR CLASSIFICATION

It has been shown that speech under Lombard Effect severely deteriorates speaker identification (SID) systems [1]. This study will focus on a novel method to significantly reduce errors in a demanding application like SID.

3.1. DNN Architecture

For features, 39-dimensional Mel-frequency Cepstral Coefficients (MFCC) are extracted, which include static, delta, and delta-delta coefficients. A 25ms Hamming window with 10ms shift was applied. The feature vectors are normalized to zero mean and unit variance to enable learning via neural networks. This normalization is done for the training set only; the mean and variance on training data is then used to scale the validation and test data. This paper uses the effectiveness of deep neural networks in extracting deeper meanings from simple cepstral features.

A deep neural network is randomly initialized for classification purposes without generative pre-training. Pre-training using Restricted Boltzmann Machines [11] was found to result in suboptimal results for Lombard Effect classification. The architecture comprises of a Multi-layer Perceptron with sigmoid activation functions in the hidden layers. The visible layer consists of nodes for feature vector input. The number of hidden layers tested ranged from 1 to 11, with increased number for increased classification complexities. The output layer consists of logistic regression nodes employing the softmax function. This layer enables the DNN to output classification results as class probabilities which sum to 1. Target classes are expressed by Y, the weight matrix and bias vector by $\mathbf{W}$ and $\mathbf{b}$ respectively.

$$P(Y = i|\mathbf{x}, \mathbf{W}, \mathbf{b}) = \text{softmax}_i(\mathbf{W}\mathbf{x} + \mathbf{b}) = \frac{e^{\mathbf{W}_i\mathbf{x}+\mathbf{b}_i}}{\sum_j e^{\mathbf{W}_j\mathbf{x}+\mathbf{b}_j}} \quad (1)$$

The classification result is obtained by noting the index of the node with the maximum class probability:

$$y_{predict} = \text{argmax}_i P(Y = i|\mathbf{x}, \mathbf{W}, \mathbf{b}). \quad (2)$$

Minimization of cross-entropy error is set as the objective function, which maximizes target class membership probabilities on training data. The loss function is expressed as negative log-likelihood,

$$\ell(\theta = \{\mathbf{W}, \mathbf{b}\}, \mathcal{D}) = -\sum_{i=0}^{|\mathcal{D}|} \log(P(Y = y_i|\mathbf{x}_i, \mathbf{W}, \mathbf{b})). \quad (3)$$

Mini-batch Stochastic Gradient Descent [12] is used to train the DNN under the backpropagation algorithm. To introduce better regularization in the DNN model so that it performs better on test data, L2-norm regularization is applied. Also called 'weight decay', this regularization method prevents overfitting by preventing the weight parameters from becoming very large.

3.2. Deep Classifier Performance

3.2.1. Lombard and Neutral Speech Classification

The normalized acoustic features are submitted to the network in batches, and the network is trained to classify them into neutral speech or Lombard Effect. For binary classification between neutral speech and any of the Lombard Effect flavors, 95.7% accuracy was achieved raising existing benchmarks. A balanced accuracy of 94.9% per class (to take uneven priors into account) was achieved as mentioned in Table 1.

3.2.2. Lombard Noise-type and Neutral Speech

For the four-way classification task into neutral speech and three noise-dependent Lombard flavors (LCR, HWY, PNK), accuracy as high as 69.1% was achieved with DNN. Table 2 shows the confusion matrix for classification results. The resulting classifier is used for adaptation of a state-of-the-art SID system in the next section.

Table 1. Classification Accuracies for Neutral and Lombard speech types; *Unweighted* means raw accuracy on test data, while *Balanced* means adjusted/weighted accuracy per class.

Classification Type	Neutral/ Lombard	Neutral, Noise-type	Neutral, Noise -type/level
Classes	2	4	10
Unweighted	95.7	69.1	60.0
Balanced	94.9	66.0	49.4

Table 2. Confusion matrix for 4-way classification between neutral and Lombard speech (Classification rates are in %, figures in bold refer to matched train/test conditions).

Test Condition	NEU	LCR	HWY	PNK
NEU	**94.2**	0.7	1.4	3.7
LCR	5.2	**43.3**	21.7	29.8
HWY	4.6	17.3	**62.8**	15.3
PNK	6.2	18.1	12.5	**63.2**

3.2.3. Lombard Noise-type, Noise-level and Neutral Speech

Classification was also performed on the same data by further refining the classified Lombard Effect flavor into the 3 different noise levels behind each of them. An overall accuracy of 60% was achieved by a single DNN model in classifying all 9 Lombard flavors (3 noise levels against 3 noise types) and neutral speech.

3.2.4. Results and Analysis

Referring back to Table 1, it shows DNN classification performance over different combinations of Lombard Effect flavors. Unweighted accuracy is for all samples in testing data which contain unbalanced classes. Balanced accuracy is calculated to balance class biases. It is evident that the classifier performs well even with slightly biased training for neutral speech and all Lombard Effect flavors. The relatively larger gap in accuracy when additionally classifying the type of noise shows that noise-level is more sensitive to classification compared to noise-types.

Varying levels of depths were required to achieve effective classification. Increasing number of hidden layers were employed as classification complexity increased from binary to 10-way classification. The results show that after careful tuning of neural network parameters (learning rate, momentum,

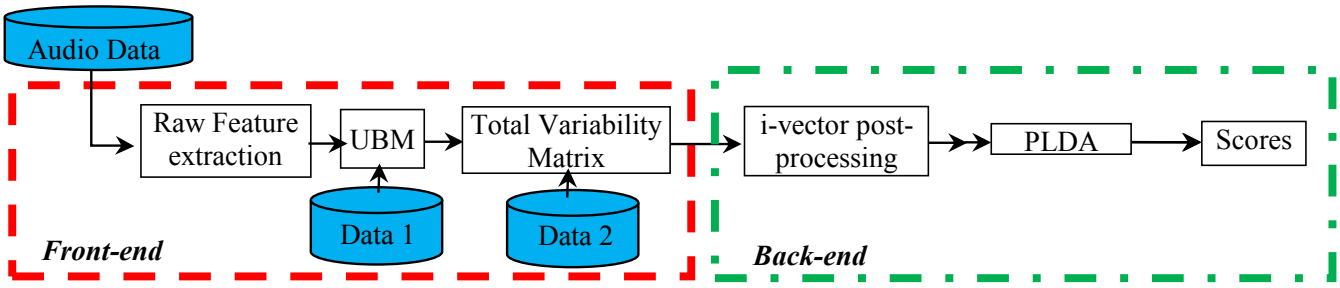

Fig. 1. System block diagram of i-Vector based SID for Lombard speech. Data 1 and 2 correspond to raw features for UBM and TV matrix, respectively. Here, both are the same as training data. 'Audio data' reflects all acoustic data used in verification.

regularization, nodes per layer, and number of hidden layers), even complex phenomena such as Lombard Effect can be effectively modeled.

4. DNN ASSISTED SPEAKER RECOGNITION

4.1. i-Vector based Speaker Identification (SID) System

The classification system includes feature extraction and back-end modeling, which is illustrated in Fig. 1. MFCC features are referred to as raw features, since they can be further processed into refined features such as i-Vectors. An i-Vector based system is the state-of-the-art platform for acoustic event identification, such as SID [13], and language ID [14, 15, 16, 17]. However, it has not been explored for Lombard speech. i-Vectors are extracted following factor analysis [13]. The i-Vector model is represented by:

$$M = m + T\omega \tag{4}$$

where T is the total variability matrix, ω is i-Vector, m is the universal background model (UBM) mean super-vector, and M is the super-vector derived from raw features. The extraction converts frame length-varied spectral features matrix into a fixed-dimension features vector for each speech utterance.

All available training data are employed to train both the UBM and total variability matrix using the EM algorithm. Next, the i-Vectors for both training and test sets are extracted with the total variability matrix. 100-dimensional i-Vectors are used for the purpose of this experiment which suits the relatively small database used. The extracted i-Vector of each speech utterance contains both inter-speaker and intra-speaker variabilities. Therefore, the PLDA classifier is employed in SID systems [18, 19]. PLDA is also adopted as back-end classifier here (Fig. 1).

4.2. Training Methods

Four separate SID systems are trained for each type of speech; under the three Lombard Effect flavors and one for neutral speech. Test data is classified by the DNN as belonging to either of these four categories. Based on results from classifier, the test data classified in each class is forwarded to the SID system trained with the respective class data. The overall system is illustrated in Fig. 2. Two kinds of approaches are analyzed in this paper.

4.2.1. Fixed Training

The first method forks speech training data and uses speech under only a single Lombard Effect or only neutral speech to train each of the four SID systems. It uses neutral speech to train one SID system, training data of Lombard Effect speech under large crowd noise in second SID system, and so on to train all four SID systems.

4.2.2. Weighted Training

A second more innovative approach is proposed in our study. The weighted approach makes use of validation results from DNN to build SID models better adapted for each test dataset subsequently classified by the DNN. DNN classification result on validation data is monitored to observe the percentage of non-target class samples present in each of four classified sets of data. Since this validation data more closely resembles the practical results on test data by the classifier, this class distribution is used to add non-target Lombard Effect speech samples in the training data for each of the four SID systems. The additional data for training is added so that training set classes are probabilistically in the same proportion as the validation set. This makes each of the four speaker identification systems robust towards samples from another class, be it a Lombard flavor or neutral speech. This method outperforms the already high performing baseline trained on all classes as shown in next section.

5. SPEAKER IDENTIFICATION RESULTS

5.1. Baseline

The baseline SID system is trained with the full set of training data including neutral speech and all 3 noise-free Lombard

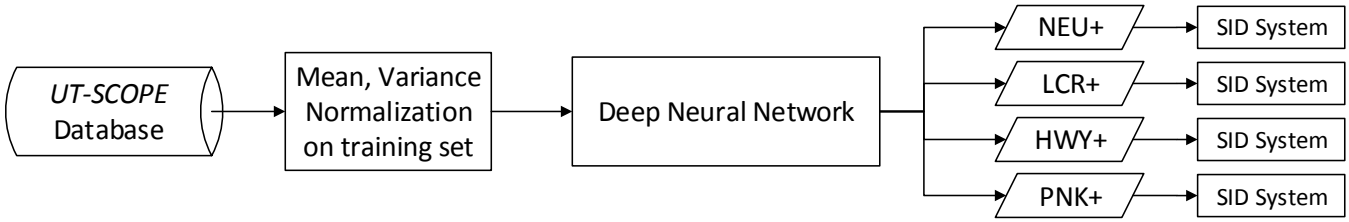

Fig. 2. System level block diagram of Deep Neural Network interface with SID systems ('+' indicates presence of non-target speech types in test data).

Table 3. SID performance with DNN classifier (%EER); 'All' error rates are for all 4 systems combined.

Training	NEU+	LCR+	HWY+	PNK+	All
Fixed	1.30	0.20	0.76	0.71	1.27
Weighted	0.75	0.17	0.47	0.47	0.53

Table 4. EER and Relative Improvement Comparison of different training methods.

DNN classifier	Absent	Present	
Training	Baseline	Fixed	Weighted
EER (%)	0.74	1.27	0.53
Rel. Imp. (%)	-	-71.6	+28.4

Effect flavors. The same i-Vector based SID system was used for all training methods. The baseline gives an upper bound on performance because Lombard Effect speech has also been included in the train data.

5.2. Fixed Training

Fixed training does not exploit all information from DNN classification. This is because of the presence of around 31% test samples belonging to other flavors or neutral speech. This method adversely affects neutral speech in particular because of the adverse impact of Lombard speech. Table 3 shows the error rates for each set of test data classified by the DNN as belonging to a particular type of Lombard or neutral speech. The test data in each column represents a majority of samples belonging to a speech type along with misclassified samples belonging to the other 3 classes.

5.3. Weighted Training

Each of the four SID systems are trained with one of the Lombard flavors or neutral speech, and a part of non-target class training data (for each of the three remaining classes) is included in proportion to the validation data classification results by DNN. This proportional inclusion prevents the SID system from being blind to other possible flavors, and thus avoids overfitting. Table 3 highlights the improvement over other systems, which is due to the inclusion of training data for DNN-misclassified speech in training the SID system. The proposed system outperforms the baseline system without DNN classification by +28.4%.

5.4. Discussion

Table 4 highlights overall error rates for the weighted and unweighted models in the presence of a DNN classifier, and the baseline. The probabilistically weighted training method exhibits an overall improvement of +28.4% in EER for the final task of speaker identification in presence of Lombard speech. Since the database contains three different noise levels behind each kind of Lombard speech produced, the proposed system performance also exhibits its resilience towards varying levels of background noise, induced Lombard Effect. Fixed, single-class training is unable to provide good results since it enhances the impact of Lombard Effect by narrowing training to a single type of speech, which leaves the system vulnerable to misclassified test data belonging to other speech types.

6. CONCLUSION

Deep neural networks are shown to outperform traditional classifiers in distinguishing between neutral speech and speech under different kinds of Lombard Effect under various background noise-types and noise-levels. The resulting accuracies are 95.7% for 2-way, 69.1% for 4-way and 60.0% for 10-way classification using cepstral features. The proposed probabilistically weighted system uses validation data classification as *a priori* information and this results in a more robust training of SID system. Appending this system to background noise reduction algorithms can result in improved robustness for other corpora. The additional validation based information can also be used as metadata for calibration [20]. Future research can focus on using this system to counter environment and channel mismatch for speaker recognition.

7. REFERENCES

[1] J.H.L. Hansen and V. Varadarajan, "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 2, pp. 366–378, Feb 2009.

[2] V. Varadarajan and J.H.L. Hansen, "Analysis of Lombard effect under different types and levels of noise with application to in-set speaker ID systems.," in *INTERSPEECH*, 2006.

[3] C. Yu, G. Liu, S. Hahm, and J.H.L. Hansen, "Uncertainty propagation in front end factor analysis for noise robust speaker recognition," in *ICASSP 2014*, May 2014, pp. 4017–4021.

[4] J.H.L. Hansen, "Morphological constrained feature enhancement with adaptive cepstral compensation (mce-acc) for speech recognition in noise and Lombard effect," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 598–614, Oct. 1994.

[5] S. Chi and Y. Oh, "Lombard effect compensation and noise suppression for noisy Lombard speech recognition," in *ICSLP 1996*, Oct 1996, vol. 4, pp. 2013–2016.

[6] H. Boril and J.H.L. Hansen, "Unsupervised equalization of Lombard effect for speech recognition in noisy adverse environments," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1379–1393, Aug 2010.

[7] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, Jan 2012.

[8] G. Hinton, L. Deng, D. Yu, G.E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[9] A. Stuhlsatz, C. Meyer, F. Eyben, T. ZieIke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *ICASSP 2011*, May 2011, pp. 5688–5691.

[10] A. Ikeno, V. Varadarajan, S. Patil, and J.H.L. Hansen, "UT-scope: Speech under Lombard effect and cognitive stress," in *Aerospace Conference, 2007 IEEE*, March 2007, pp. 1–7.

[11] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, July 2006.

[12] T. Zhang, "Solving large scale linear prediction problems using stochastic gradient descent algorithms," in *Proceedings of the Twenty-first International Conference on Machine Learning*. 2004, ICML, ACM.

[13] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[14] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Interspeech*, pp. 861–864, 2011.

[15] Q. Zhang, G. Liu, and J.H.L. Hansen, "Robust language recognition based on hybrid fusion," in *Odyssey*, June 2014, pp. 152–157.

[16] G. Liu, C. Zhang, and J.H.L. Hansen, "A linguistic data acquisition front-end for language recognition evaluation," *Odyssey*, pp. 224–228, June 2012.

[17] G. Liu, S.O. Sadjadi, T. Hasan, J. Suh, C. Zhang, M. Mehrabani, H. Boril, A. Sangwan, and J.H.L. Hansen, "UTD-CRSS systems for NIST Language Recognition Evaluation 2011," *NIST 2011 LRE Workshop*, Dec. 2011.

[18] G. Liu, T. Hasan, H. Boril, and J.H.L. Hansen, "An investigation on back-end for speaker recognition in multi-session enrollment," in *ICASSP 2013*, May 2013, pp. 7755–7759.

[19] P. Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey*, June 2010.

[20] L. Ferrer, L. Burget, O. Plchot, and N. Scheffer, "A unified approach for audio characterization and its application to speaker recognition," in *Proc. of Odyssey Workshop*, 2012.