Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 112 (2017) 316–322

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017, 6-8 September 2017, Marseille, France

Improving speech recognition using data augmentation and acoustic model fusion

Ilyes Rebai[a,*], Yessine BenAyed[a], Walid Mahdi[a], Jean-Pierre Lorré[b]

[a] *Multimedia InfoRmation system and Advanced Computing Laboratory*
University of Sfax, Sfax, Tunisia
[b] *Director of Innovation @LINAGORA, Toulouse - France*

Abstract

Deep learning based systems have greatly improved the performance in speech recognition tasks, and various deep architectures and learning methods have been developed in the last few years. Along with that, Data Augmentation (DA), which is a common strategy adopted to increase the quantity of training data, has been shown to be effective for neural network training to make invariant predictions. On the other hand, Ensemble Method (EM) approaches have received considerable attention in the machine learning community to increase the effectiveness of classifiers. Therefore, we propose in this work a new Deep Neural Network (DNN) speech recognition architecture which takes advantage from both DA and EM approaches in order to improve the prediction accuracy of the system. In this paper, we first explore an existing approach based on vocal tract length perturbation and we propose a different DA technique based on feature perturbation to create a modified training data sets. Finally, EM techniques are used to integrate the posterior probabilities produced by different DNN acoustic models trained on different data sets. Experimental results demonstrate an increase in the recognition performance of the proposed system.

Peer-review under responsibility of KES International

Keywords: speech recognition, deep learning, data augmentation, ensemble method, linear logistic regression

1. Introduction

Speech recognition system is mainly based on three models, an acoustic model, a language model and a pronunciation lexicon. Besides, the performance of these models relies greatly on the amount of data used during training. During the last years, several researchers focused their works on improving two key elements in speech recognition: speech data, and acoustic model.

During the past decade, DNN based speech recognition systems have been demonstrated to provide significantly higher accuracy in continuous phone and word recognition tasks than the earlier state-of-the-art GMM-based systems [1,2,3,4]. More recent advances in deep learning techniques, such as convolutional deep neural networks (CNN)

* Corresponding author.
E-mail address: rebai_ilyes@hotmail.fr

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of KES International
10.1016/j.procs.2017.08.003

and deep recurrent neural networks (RNN), have shown high performance in speech recognition[5,6,7,8]. While deep learning in speech recognition has claimed state-of-the-art performance over conventional GMM based systems, there are sometimes fewer differences in performance between the different deep techniques. Nonetheless, it could be more beneficial to combine these various systems into a single one. In fact, the resulting of the ensemble is generally more accurate than any of the individual model that composes the ensemble[9,10].

Besides, another research direction for speech recognition focused on the development of efficient DA methods. In practice, a large amount of transcribed training data is usually needed to enable accurate speech recognition which is not the case for several languages. Therefore, DA is proposed, where the speech data is artificially augmented by applying different types of transformations. Indeed, it is a common strategy adopted to increase the quantity of training data[11,12,13,14]. For instance, vocal tract length perturbation (VTLP) has shown gains on the TIMIT phoneme recognition task using DNN based acoustic modeling[11].

In this paper, a new DNN based speech recognition system is proposed in which we take advantage from the existing approaches in order to improve the recognition performance. Specifically, DA and EM approaches and are combined into a single system. Indeed, we exploit DA approaches with DNN acoustic modeling. We first explore the VTLP and we propose a different DA technique based on feature perturbation to augment the training data. Next, EM techniques are used to integrate the posterior probabilities produced by different DNN acoustic models trained on different data sets to give improved prediction accuracy. Three types of techniques are evaluated in this work: major voting scheme, average scheme, and Linear Logistic Regression (LLR) for Fusion and Calibration. The experimental results demonstrate the effectiveness of the proposed system.

The remainder of the paper is organized as follows. In Section 2, we give details of the VTLP approach and the proposed feature perturbation technique used to generate transformed input features for deep neural network training. Section 3 reviews the EM and LLR techniques used for acoustic model combination. Next, we present the proposed speech recognition system and the experimental setup used in this work in Section 4 and 5 respectively. Finally, experimental results are presented in Section 6, followed by a conclusion in Section 7.

2. Data Augmentation

Data augmentation is a common strategy adopted to increase the quantity of training data. It is a key ingredient of the state of the art systems for image recognition and speech recognition[15,11]. With the widespread adoption of neural networks in speech recognition systems which require a large speech database for training such a deep architecture, DA is very useful for small data sets. Indeed, it is possible to augment speech databases and to use the augmented database to achieve improved accuracy.

2.1. Vocal Tract Length Perturbation

With DNN based acoustic modeling, vocal tract length perturbation (VTLP)[11] has shown gains on the TIMIT phoneme recognition task. VTLP was further extended to large vocabulary continuous speech recognition (LVCSR)[12]. It was reported that selecting VTLP warping factors from a limited set of perturbation factors was better[12].

In practice, for each utterance in the training set, a warping factor α is randomly chosen from $[0.9, 1.1]$ to warp the frequency axis. Therefore, the vocal tract length of the speaker is slightly perturbed to distort the original speech spectrum of the utterance to create a new replica of it. In this work, a set warping factors, $\{0.9, 1.0, 1.1\}$ is used to create three copies of the original features. Furthermore, the same warping factors are applied to all speakers in the training set.

2.2. Feature Perturbation

We propose in this work a different method for DA and compare it with the existing augmentation technique VTLP. Feature perturbation aims at modifying the extracted acoustic feature vectors by adding random values. This trans-

formation may degrade the speech quality, change the speaker's voice, etc., which is useful when creating a modified version of the original data. In fact, this is typically what is expected from such an audio perturbation technique.

Given speech signal $s(n)$, this can be represented as a combination of the original message signal with the noise signal: $s(n) = m(n) + e(n)$, where $m(n)$ is the message signal and $e(n)$ is the noise signal. Then, applying a feature extraction technique, e.g. MFCC or PLP, the resulting transformation is: $feat^s = feat^m + feat^e$.

Our idea consists on perturbing the extracted features $feat^s$ for each utterance by adding random values $feat^r$ chosen from a range (e.g., [0, 1]). Thus, the added random values could be viewed as an explicit noise features: $feat^p = feat^s + feat^r$ where $feat^r$ is the random features, and $feat^p$ is the final features, which we called perturbed features. A larger range may create unrealistic distortions. Thus, the size of the features are doubled: the warped features plus the original ones.

3. Model Fusion Techniques

3.1. Ensemble Method Techniques

Multiple classifier methods aim at combining the predictions of several classifiers, acoustic recognition models, to supply a single classification result. The resulting of the ensemble is generally more accurate than any of the individual classifiers that compose the ensemble. There are different methods to obtain and combine diverse classifiers.

The simplest and efficient methods aim at combining the log-likelihood outputs using either major voting scheme or average scheme. Given K acoustic recognition models R_k, major voting consists on combining these models using *argmax* function. This function returns the maximum value predicted by all models for the same class. Another way to combine multiple classifier is to compute the average. Accordingly, given an input x, the predictions of all K models are averaged:

$$y' = \frac{1}{K} \sum_{k=1}^{K} R_k(x) \tag{1}$$

3.2. Linear Logistic Regression for Fusion and Calibration

Linear logistic regression is a very common approach for converting a vector of features into likelihood ratios. The standard logistic regression model is based on the logistic function, defined as:

$$P(x; \alpha, \beta) = \frac{1}{1 + e^{-y(\alpha x + \beta}} \tag{2}$$

where x is the input feature vector, and y is the corresponding output class.

In the case of fusion and calibration, this standard is reformulated in order to take into account multiple recognizers. Indeed, using K recognition models, denoted as $R_k, k = 1..k$, in which each model predict the same N classes, and in the same log-likelihood format, a new recognition model R' is formed and parameterized by $\theta = \alpha_1, \alpha_2, ..., \alpha_K, \beta'$, where $\alpha_k \in \mathbb{R}$ is a scalar weight of the model k, and $\beta' \in \mathbb{R}^N$ is a log-likelihood vector[16]. For a given input x, the output of the new recognizer is formed as:

$$R' = \beta' + \sum_{k=1}^{K} \alpha_k R_k(x) \tag{3}$$

The weighted sum over all recognition models forms what is known as a *fusion* or *combination*. The vector b' is used as an affine calibration transformation on the fusion output[16].

4. Proposed System

In this paper, we propose to improve the conventional ASR systems with two key elements. Usually, several acoustic models are trained and the model that achieves the best performance is used in the recognition phase. However, it may be more efficient to combine all the trained models. Therefore, we adapt the idea of EM and score fusion for combining the predictions of multiple acoustic models. Figure 1 presents the proposed ASR architecture.

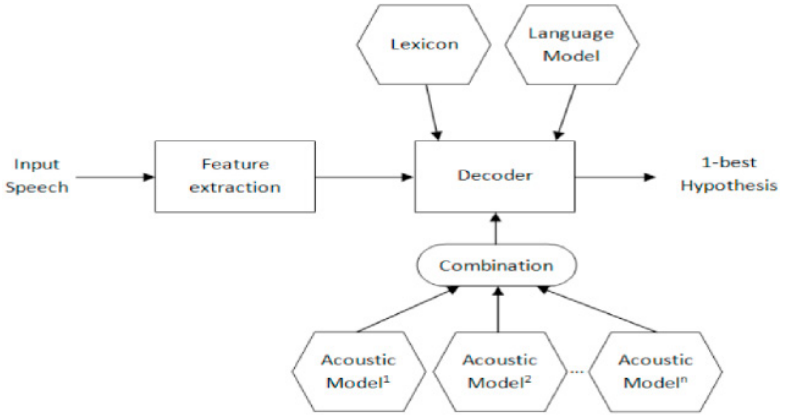

Fig. 1. The proposed ASR architecture.

The main idea is that each acoustic model produces different information for representing the input signal, which makes the models diverse enough for efficient use in an ensemble. One strategy is to learn different acoustic models with different data separately. After that, the log-posteriors are combined to generate the final log-posteriors. Accordingly, DA techniques are used in this work to build acoustic models using different data. Then, these models are combined and the result is fed to the decoder along with the language model and the lexicon to produce the final transcript.

5. Experimental Setup

We report results on automatic speech recognition task in French. The speech database is approximately 49 hours. We divided the data into two subsets: a train set and a test set. In order to guarantee that there was no speaker overlap between the training and test sets, we split the data by speakers to make sure that speakers appear in one set will not appear in the other sets. Finally, 47 hours are used for training and the remaining for the test set. Two tri-gram language models were trained on the training transcripts along with additional data collected from different source. Indeed, a first model was built using around 20M words, denoted as Medium Language Model (MLM), while a second model was built using approximately 35M words, denoted as Large Language Model (LLM).

The lexicon used in this work is issued from the CMU Sphinx Speech recognition Toolkit[1]. The French dictionary is composed of 62K unique words. In addition, we extend the lexicon with words in the training corpus that do not appear in it. The pronunciations for these words are generated using the Sequitur G2P toolkit.

5.1. Features

Mel-frequency cepstral coefficients (MFCCs) are used in this work. However, the extracted feature are not directly used for training. Indeed, a set of transformation is applied in order to improve the feature representation and extract the most relevant information. First, the static feature vectors are computed. Then, a feature dimensionality reduction,

[1] https://sourceforge.net/projects/cmusphinx/files/Acoustic and Language Models/French/fr.dict/download

Linear Discriminant Analysis (LDA)[17], is then used to reduce the dimensionality of the static vector stacked with a four spliced vectors (117 dimensional vector) to 40 dimensional vector. Next, a feature transformation, Maximum Likelihood Linear Transform (MLLT)[18], is applied to transform the obtained feature vector. Finally, a feature space Maximum Likelihood Linear Regression (fMLLR)[19] (also known as global CMLLR) is applied to normalize inter-speaker variability. The final vector is 40-dimensional features used to train the DNN models.

5.2. DNN configuration

DNN based acoustic models are used in our experiments. These models provide state of the art performance on various LVCSR tasks. Hence, they provide a strong baseline to verify the gains due to the DA techniques and the model fusion methods.

The neural networks had the following configuration. The input features are the 40-dimensional features as detailed in the previous section. The output layer is a softmax layer, and the outputs represent the log-posterior of the output labels, which correspond to context-dependent HMM states (there were about 3100 states in our experiments). The number of neurons in the hidden layer is the same for all hidden layers.

Three different kinds of DNN are evaluated in this work: deep maxout networks with p-norm nonlinearity function[20], denoted as DNN-pnorm, DNN with tangent hyperbolic activation function, denoted as DNN-tanh, and DBM-DNN that uses the generative weights of deep Boltzmann machines (DBMs) for initialization of DNNs.

6. Experimental Results

Various experiments have been conducted in this work using different configurations, while training and test sets have been kept separate. Word Error Rate (WER) was considered for computing the score of the recognition process. Additionally, Phone Error Rate (PER) was also used for performance evaluation. Table 1 presents the results using various training techniques.

Table 1. Experimental results on the test set.

	PER		WER	
Training & decoding	MLM	LLM	MLM	LLM
GMM-HMM	16.10	15.88	21.18	20.56
SGMM	15.46	15.31	19.34	19.16
DNN-pnorm 3-layers	15.57	15.30	17.76	17.27
DNN-pnorm 5-layers	15.40	15.14	**17.38**	**16.87**
DNN-pnorm 7-layers	15.41	15.29	17.22	16.95
DNN-tanh 3-layers	15.76	15.39	18,54	18,15
DNN-tanh 5-layers	15.43	15.15	17,98	17,64
DNN-tanh 7-layers	15.40	15.11	17,77	17,50
DBM-DNN 3-layers	16.25	15.95	17.30	17.10
DBM-DNN 5-layers	16.03	15.76	17.14	16.91
DBM-DNN 7-layers	16.12	15.82	17.25	17.02

It could be seen that DNN outperforms all other non-DNN configurations. These results confirm the superiority of the deep learning techniques overs conventional ones: GMM-HMM, SGMM[21]. Furthermore, the DNN with 7 hidden layers achieves slightly worse results than the DNN with 5 hidden layers. This could be explained by the limited size of training set that could not perfectly trained using a very deep network. When comparing the different DNN configurations, we can see that deep maxout networks perform better than DNN-tanh and DBM-DNN ones. Specifically, DNN-pnorm produces 16.78% of WER while DNN-tanh and DBM-DNN achieve a WER equal to 17.50% and 16.91% respectively. Finally, the ASR system generally performs better using the large language model than the medium language model.

Table 2. Data augmentation results

	PER		WER	
Training & decoding	MFCC	VTLP	MFCC	VTLP
SGMM2 (original data)	14.31		19.16	
SGMM2 (generated data)	14.27	13.96	18.91	18.85
DNN (original data)	15.14		16.87	
DNN (generated data)	15.03	14.35	16.84	16.83

In Table 2, we present the experimental results using VTLP and feature perturbation. DNN-pnorm is used in this set of experiments. When comparing the DNN using generated data with the baseline DNN, a relative improvement of 0.3% WER was observed on the test set, when using VTLP and feature perturbation training data using SGMM model. However, no improvement was achieved using DNN models for both methods. When comparing the DA techniques, we could observe that using a simple feature perturbation is as good as VTLP. Thus, using feature perturbation was beneficial compared with VTLP, even when less feature size is used.

Finally, the last experiments were conducted to evaluate the performance of the propose model fusion method. In these experiments, the best DNN model using the original features and using perturbed features and using VTLP features are combined using three different methods: average, argmax, and LLR fusion. Figure 2 shows the objective function value during training for LLR fusion and calibration.

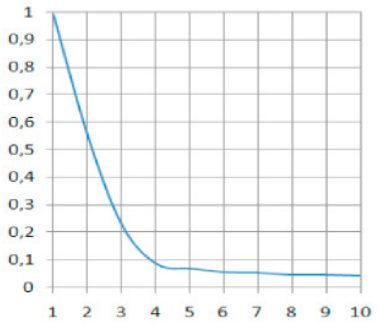

Fig. 2. Linear logistic regression training result.

Table 3. Multiple acoustic model fusion

	PER	WER
Average	14.16	16.40
Argmax	**13.58**	**16.36**
LLR	14.08	16.39

It could be seen that the model converges faster during training. Hence we trained the LLR system for few iterations. Table 3 compares the performance improvement from model fusion techniques on the test set. It can be seen that model fusion was efficient compared to the DNN baseline system. In fact, DNN-original features, DNN-VTLP features, and DNN-feature perturbation perform almost the same performance while an improvement of 0.5% in the WER is obtained when combining these three DNN models. In addition, LLR and average methods achieved the same performance, 16.39% WER, while argmax technique performs the best WER result, 16.36% WER, on the test set. A similar improvement can be observed using PER metric. These results prove the assumption that combining the trained models may be more efficient and demonstrate the effectiveness of EM techniques in improving the accuracy of the system.

7. Conclusion

In this work, we introduce an adaptation of a fusion architecture to the speech recognition tasks. The proposed architecture is based on the combination of model fusion and EM. The study reported in this paper gives a new technique for improving the speech recognition performance. Indeed, our system takes advantage from two key elements: model fusion methods which have been shown to be very efficient on several tasks, and DA techniques which have been demonstrated to be useful for improving the ASR performance.

This paper presents an audio augmentation technique based on feature perturbation. Our technique has low implementation cost is proved to be as good as VTLP technique. Finally, we adapt the idea of EM and fusion to combine all model into a single ASR system. The intuitive idea is to use average and argmax technique for models combination. Along with that, LLR for Fusion and Calibration is also evaluated. Experimental results on ASR task in French showed the effectiveness and efficiency of the proposed system.

Acknowledgements

This work is conducted as a part of the project OpenPass::NG.

References

1. Dahl, G.E., Yu, D., Deng, L., Acero, A.. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on Audio, Speech, and Language Processing* 2012;**20**(1):30–42.
2. Deng, L., Li, J., Huang, J.T., Yao, K., Yu, D., Seide, F., et al. Recent advances in deep learning for speech research at microsoft. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE; 2013, p. 8604–8608.
3. Jaitly, N., Nguyen, P., Senior, A.W., Vanhoucke, V.. Application of pretrained deep neural networks to large vocabulary speech recognition. In: *Interspeech*. 2012, p. 2578–2581.
4. Mohamed, A.r., Dahl, G.E., Hinton, G.. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing* 2012;**20**(1):14–22.
5. Deng, L., Abdel-Hamid, O., Yu, D.. A deep convolutional neural network using heterogeneous pooling for trading acoustic invariance with phonetic confusion. In: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE; 2013, p. 6669–6673.
6. Sainath, T.N., Mohamed, A.r., Kingsbury, B., Ramabhadran, B.. Deep convolutional neural networks for lvcsr. In: *Acoustics, speech and signal processing (ICASSP), 2013 IEEE international conference on*. IEEE; 2013, p. 8614–8618.
7. Graves, A., Mohamed, A.r., Hinton, G.. Speech recognition with deep recurrent neural networks. In: *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE; 2013, p. 6645–6649.
8. Graves, A., Jaitly, N., Mohamed, A.r.. Hybrid speech recognition with deep bidirectional lstm. In: *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE; 2013, p. 273–278.
9. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.. A comparison of decision tree ensemble creation techniques. *IEEE transactions on pattern analysis and machine intelligence* 2007;**29**(1).
10. Ceamanos, X., Waske, B., Benediktsson, J.A., Chanussot, J., Fauvel, M., Sveinsson, J.R.. A classifier ensemble based on fusion of support vector machines for classifying hyperspectral data. *International Journal of Image and Data Fusion* 2010;**1**(4):293–307.
11. Jaitly, N., Hinton, G.E.. Vocal tract length perturbation (vtlp) improves speech recognition. In: *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*. 2013, .
12. Cui, X., Goel, V., Kingsbury, B.. Data augmentation for deep neural network acoustic modeling. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)* 2015;**23**(9):1469–1477.
13. Ragni, A., Knill, K.M., Rath, S.P., Gales, M.J.. Data augmentation for low resource languages. In: *INTERSPEECH*. 2014, p. 810–814.
14. Ko, T., Peddinti, V., Povey, D., Khudanpur, S.. Audio augmentation for speech recognition. In: *INTERSPEECH*. 2015, p. 3586–3589.
15. Cireşan, D.C., Meier, U., Masci, J., Gambardella, L.M., Schmidhuber, J.. High-performance neural networks for visual object classification. *arXiv preprint arXiv:11020183* 2011;.
16. Brümmer, N.. *Measuring, refining and calibrating speaker and language information extracted from speech*. Ph.D. thesis; Citeseer; 2010.
17. Duda, R.O., Hart, P.E., Stork, D.G., et al. *Pattern classification*; vol. 2. Wiley New York; 1973.
18. Gales, M.J.. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer speech & language* 1998;**12**(2):75–98.
19. Gopinath, R.A.. Maximum likelihood modeling with gaussian distributions for classification. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*; vol. 2. IEEE; 1998, p. 661–664.
20. Zhang, X., Trmal, J., Povey, D., Khudanpur, S.. Improving deep neural network acoustic models using generalized maxout networks. In: *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE; 2014, p. 215–219.
21. Povey, D., Burget, L., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., et al. Subspace gaussian mixture models for speech recognition. In: *IEEE International Conference on Acoustics Speech and Signal Processing*. IEEE; 2010, p. 4330–4333.