Discriminative Capacity and Phonetic Information of Bottleneck Features in Speech

Ana Montalvo(✉) and José Ramón Calvo

Advanced Technologies Application Center, 7th A Street, #21406 Havana, Cuba
{amontalvo,jcalvo}@cenatav.co.cu
http://www.cenatav.co.cu

Abstract. The impressive gain in performance obtained using deep neural network (DNN) for automatic speech recognition have motivated their application to other speech related tasks such as speaker recognition and language recognition, but there is still uncertainty about what is deep training strategy extracting, from the acoustic data, to make it such a powerful learning tool. This paper compares the discriminative capacity and the phonetic information conveyed by the feature-space maximum likelihood linear regression (fMLLR) before and after passing through a DNN trained to discriminate between tri-phone tied states. The proposed experimentation reflected the superiority of DNN bottleneck features regarding its information content.

Keywords: Entropy · Phonetic information · Bottleneck features · Deep neural network

1 Introduction

Five years ago, most speech recognition systems used hidden Markov models (HMMs) to deal with the temporal variability of speech and gaussian mixture models (GMMs) to determine how well each state of each HMM fits a sequence of feature vectors that represents the acoustic input.

In recent years, discriminative hierarchical models such as deep neural networks (DNNs) [1] became feasible and significantly reduced the word error rate.

When neural nets were first used, they were trained discriminatively. It was only recently that researchers showed that significant gains could be achieved by adding an initial stage of generative pretraining. The successes achieved using pretraining led to a resurgence of interest in DNNs for acoustic modeling.

There are mainly two different ways to incorporate deep learning techniques in speech recognition related tasks [1].

In the first configuration which is called DNN-HMM or simply hybrid, a DNN is used to compute the posterior probabilities of context-dependent HMM states based on observed feature vectors, then a Viterbi decoding is performed with these posteriors. The second configuration, which is called tandem, uses the DNN to perform a nonlinear discriminative feature transformation, which can

C. Beltrán-Castañón et al. (Eds.): CIARP 2016, LNCS 10125, pp. 134–141, 2017.
DOI: 10.1007/978-3-319-52277-7_17

be regarded as a bridge between low-level acoustic input and high-level phonetic information [2], hence exploiting the output from different layers of the DNN may lead to improved utterance representation.

Specifically in the bottleneck (BN) tandem approach proposed by [3], a DNN in which one of the internal layers has a small number of hidden units (relative to the size of the other layers) is trained. Whereas the linear output of this bottleneck layer is taken as output instead of the posteriors.

The impressive gains in performance obtained using BN features for automatic speech recognition have motivated their application to other speech related tasks such as speaker recognition and language recognition with their particular backends.

But, why are they so successful? What is deep training strategy extracting from the acoustic data? Which is the most appropriate setting to achieve success?

In the present work we assess the discriminative capacity and the phonetic information of bottleneck features applying GMMs and entropy based measures. As far as the authors know, this type of study has not been reported before. Taking into account these aspects would allow to have a priori information on the behavior of the BN features, which in turn could be very useful to adjust the layout of the DNN, in pursuit of maximizing information, without performing the entirely recognition process. This research is aimed to be applied in obtaining new representations for identifying languages on noisy and short duration signals.

We first briefly describe the BN framework in Sect. 2. In the following sections, we discuss about BN features discriminative capacity and phonetic information (3 and 4 respectively). Section 5 describes the corpus, the toolkit and the performed experiments. Section 6 presents some results and discussion, and Sect. 7 will conclude this work.

2 Bottleneck Features

BN features are conventionally generated from a neural network in which one of the hidden layers has a very small number of hidden units relative to the other layers. This layer of small size is described as bottleneck layer and can be viewed as an approach of nonlinear dimensionality reduction, since it constricts the neural network so that the information embedded in the input features will be forced into a low dimensional representation.

BN features have become a very powerful way to use learned representations from data, and a lot of different features have been tested in order to feed the neural network with the optimal acoustic representation toward the particular classification problem [4].

There is a complex relationship between acoustic and BN features, certain correlation can therefore exist between the two. BN features provide a different view of the same speech signals, complementary information characterizing the acoustic realization are also implicitly learnt by BN features.

In our experiments the BN features are extracted from a DNN consisting of stacked Restricted Boltzman Machines [1] pre-trained in an unsupervised manner [5].

The DNN configuration is $n\times40$-1024-1024-1024-42-1024-$d_{softmax}$, where the input is a concatenation of the current frame with the preceding and following $(n-1)/2$ neighboring frames. In our case $n = 9$ considers 9 staked adjacent frames of the 40-dimensional acoustic feature for each time instance. $d_{softmax}$ is the number of units in the output layer, in practice, $d_{softmax}$ is set to 2171 according to english tri-phone tied states present in our database.

The outputs of the bottleneck layer yield a compact representation of both acoustic and phonetic information for each frame independently.

As input of the DNN we used standard Mel-Frequency Cesptral Coefficients (MFCC) followed by a feature space transformation using feature-space maximum likelihood linear regression (fMLLR). This feature has shown to be critical dealing with speaker variability [6].

3 Discriminative Capacity of the BN Features

A DNN trained to discriminate between tri-phone tied states must attenuate the information related to other sources such as the channel, speaker, gender and sessions information, so that the final linear classifier (softmax layer) can effectively discriminate between tri-phone classes. BN features extracted from such a classifier should perform well as phonotactic feature for language recognition, and be more robust to the variabilities above mentioned [7].

In order to compare the phoneme discriminative capacity of the BN features with their predecessors fMLLR, we modeled both feature distributions using Gaussian Mixture Models. For each distribution (BN features and fMLLR) a Gaussian Mixture Model-Universal Background Model (GMM-UBM) was trained and MAP adapted on phonetically segmented and annotated corpora, to build a GMMs representing each phoneme. Then a phoneme classification all across the test set was performed and identification rate (IR) was used to measure the phoneme discriminative capacity of each representation.

4 Phonetic Information Conveyed by the BN Features

In this paper we propose to measure the phonetic information conveyed by the BN features and fMLLR, using two entropy based metrics.

For comprehensiveness of exposition, we offer a brief outline of these methods and a recall of the definition of the entropy measure.

If Y and X are discrete random variables, Eqs. 1 and 2 gives the entropy of Y before and after observing X:

$$H(Y) = -\sum_{y\in Y} p(y) log_2(p(y)), \tag{1}$$

$$H(Y|X) = -\sum_{x\in X} p(x) \sum_{y\in Y} p(y|x) log_2(p(y|x)). \tag{2}$$

Information Gain (IG)

The information found commonly in this two variables is defined as the mutual information or IG:

$$IG = H(Y) - H(Y|X) \tag{3}$$

If entropy $H(Y)$ is regarded as a measure of uncertainty about a random variable Y, then $H(Y|X)$ is a measure of what X does not say about Y. IG can be read then as the amount of uncertainty in Y, minus the amount of uncertainty that remains in Y after X is known, in other words IG is the amount of uncertainty in Y which is removed by knowing X.

This corroborates the intuitive meaning of IG as the amount of information (that is, reduction in uncertainty) that knowing either variable provides about the other.

Symmetrical uncertainty (SU)

Unfortunately, IG is biased in favor of features with more values, that is, attributes with greater numbers of values will appear to gain more information than those with fewer values even if they are actually no more informative. SU compensates for information gains bias toward attributes with more values and normalizes its value to the range $[0, 1]$.

$$SU = \frac{IG}{H(Y) + H(X)} \tag{4}$$

5 Experimental Setup

We performed two experiments on TIMIT. The first experiment compares BN and fMLLR features regarding the discriminative power of each of the 48 GMMs (one per phoneme). The second experiment uses the same GMM-UBM, this time to estimate global (Eq. 1) and conditional entropy (Eq. 2), and to evaluate the performance of entropy based measures.

5.1 Data Corpus

The TIMIT corpus consists of 4288 sentences (approximately 3.5 hours) spoken by 630 speakers of 8 major dialects of American English [8]. The training set contains 3,696 sentences from 462 speakers. The development set contains 400 sentences from 50 speakers and the test set 192 sentences from 24 speakers. We defined our test set as the merge of the TIMIT's development and test sets.

When training the DNNs, we use 90 % of the training set's sentences (1,012,340 frames) as training data and the remaining 10 % (112,483 frames) as a validation set.

5.2 The DNN Toolkit

The DNN toolkit selected for our experiments was PDNN [9], together with Kaldi[1] [10] one of the most popular toolkits for constructing ASR systems.

PDNN is written in Python, and uses the Theano library [11] for all its computations. The initial model to force align the data and generate a label for each feature vector, was built using Kaldi (TIMIT “s5” recipe).

We used the *BNF Tandem* PDNN’s recipe [9] to train the DNN. The resulting BN features are mean and variance normalized.

5.3 Experiments

The 13-dimensional MFCC features are spliced in time taking a context of ±4 frames, followed by de-correlation and dimensionality reduction to 40 using linear discriminant analysis. The resulting features are further de-correlated using maximum likelihood linear transform [12]. This is followed by speaker normalization using fMLLR, also known as constrained MLLR [13].

To asses the discriminative capacity of both kind of features (BN and fMLLR), different GMM-UBMs were trained varying the number of gaussian components, all of them with shared diagonal covariance. 48 GMMs were obtained with MAP mean adaptation for all the GMM-UBM previously obtained. The relevance factor used was $\tau = 10$. The consecutive frames of each phoneme are grouped together and evaluated against the 48 models regarding the identification rate (IR).

To deeply understand the intrinsic information of the features and to evaluate how informative can the fMLLR become after passing through a DNN, two entropy based measures were used: Information Gain (IG) and Symmetrical Uncertainty (SU).

6 Results

The first experiments show that BN features are better than fMLLR discriminating phonemes, when modeled with GMM (Table 1).

Table 1. Analysis of the features discriminative capacity.

GMM dimension	BN features IR (%)	fMLLR IR (%)
4 components	**66.8**	53.2
8 components	66.4	54.0
16 components	66.7	56.4
32 components	64.9	**57**
64 components	62.8	56.8

[1] https://kaldi.svn.sourceforge.net/svnroot/kaldi/trunk.

It is worth noting how the best IR obtained with fMLLR is far from the worst obtained with BN features: BN features need less gaussian components to fully express its complexity. Using a GMM of 4 components yields the best discriminative capacity.

The fact that modeling BN features with gaussians allows the evaluation of phoneme discriminative capacity, could be useful to set the optimal DNN configuration without carrying out all the acoustic modeling and decoding part of the speech recognition process.

Analyzing the behavior of IG shown in Fig. 1, BN features outperforms fMLLR for all the gaussian models. For 4-components GMM is achieved the best performance and the IG distance between both features is wider.

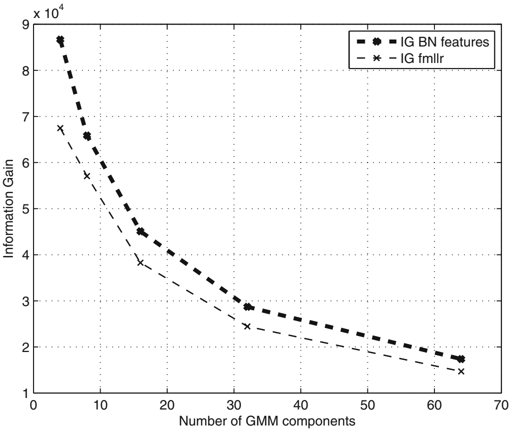

Fig. 1. Phonetic information: IG behavior for BN and fMLLR features.

Based on the previously computed entropy and conditional entropy, and looking at Fig. 2, one can conclude that the increase of the IG for BN features it is due to the intrinsic information carried by the features themselves, represented by the entropy.

In the other hand, the conditional entropy showed that both features are closely related with their respective phonetic models, so having small conditional entropy values, which is in complete correspondence with SU behavior.

The SU behaved pretty similar for all gaussian models (Table 2), showing a slightly higher performance in BN features.

SU values are close to one for both representations, indicating that the knowledge of features effectively predicts the labels. They are also quiet similar for both features because SU was defined to play a role in favor of variables with fewer values, and with such smalls values of conditional entropy this barely comes up.

In other words, small values of conditional entropy leave out the IG's bias weakness, and makes US less illustrative for the concerned phenomenon.

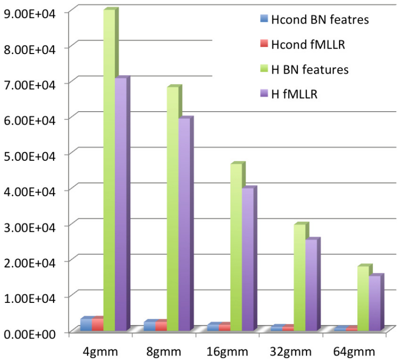

Fig. 2. Entropy contribution to IG.

Table 2. Symmetrical uncertainty.

Feature	4 components	8 components	16 components	32 components	64 components
BN	0.9632	0.9641	0.9642	0.9640	0.9631
fMLLR	0.9525	0.9586	0.9584	0.9585	0.9572

7 Conclusions

The main conclusion of this research is that BN features have more discriminative capacity than fMLLR and this in turn is a consequence of higher values of phonetic information and entropy.

These results show that the information content of BN features, is closely linked to its discriminative power. Then to calculate the entropy is a very easy way to evaluate features performance, avoiding conducting the recognition process as a whole, looking for an optimal network configuration.

As the best IR was obtained for a 4 components GMMs, we can conclude that BN features need less components to fully express its complexity than fMLLR.

It was also interesting to confirm that, the close dependence between features and phonetic labels, implies that the IG comes determined by the uncertainty contained in the features alone.

Future experiments will be conducted moving the BN layer, to asses its impact over language identification and to observ IG and US behavior.

References

1. 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.: Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Sig. Process. Mag. **29**(6), 82–97 (2012)

2. Song, Y., Cui, R., Hong, X., Mcloughlin, I., Shi, J., Dai, L.: Improved language identification using deep bottleneck network. Proc. ICASSP **2015**, 1695–1699 (2015)
3. Gr'ezl, F., Karafiát, M., Kontar, S., Cernocký, J.: Probabilistic and bottle-neck features for LVCSR of meetings. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 757–760 (2007)
4. Rath, S., Povey, D., Vesely, K., Cernocky, J.: Improved feature processing for deep neural networks. In: Proceedings of Interspeech, pp. 109–113 (2013)
5. Albornoz, E.M., Sánchez-Gutiérrez, M., Martinez-Licona, F., Rufiner, H.L., Goddard, J.: Spoken emotion recognition using deep learning. In: Bayro-Corrochano, E., Hancock, E. (eds.) CIARP 2014. LNCS, vol. 8827, pp. 104–111. Springer, Heidelberg (2014). doi:10.1007/978-3-319-12568-8_13
6. Yu, D., Deng, L.: Automatic Speech Recognition a Deep Learning Approach. Springer, London (2015)
7. Jiang, B., Song, Y., Wei, S., Liu, J.-H., McLoughlin, I., et al.: Deep bottleneck features for spoken language identification. PLoS ONE **9**(7), e100795 (2014). doi:10.1371/journal.pone.0100795
8. Lopes, C., Perdigao, F.: Phone recognition on the TIMIT database. Speech Technol. **1**, 285–302 (2011)
9. Miao, Y.: Kaldi+PDNN: Building DNN-based ASR systems with Kaldi and PDNN. Computing Research Repository, abs/1401.6984 (2014)
10. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K.: The Kaldi speech recognition toolkit. In: IEEE Workshop on Automatic Speech Recognition and Understanding. Signal Processing Society (2011)
11. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a CPU and GPU math expression compiler. In: Proceedings of the Python for Scientific Computing Conference SciPy (2010)
12. Gopinath, R.: Maximum likelihood modeling with Gaussian distributions for classification. Proc. IEEE ICASSP **2**, 661–664 (1998)
13. Gales, M.J.F.: Maximum likelihood linear transformations for HMM-based speech recognition. Comput. Speech Lang. **12**(2), 75–98 (1998)