ROBUST SPEECH RECOGNITION USING MULTIVARIATE COPULA MODELS

Alireza Bayestehtashk[1]*, Izhak Shafran*[2] *and Amir Babaeian*[3]

[1]Oregon Health & Science University, [2] Google Inc [3] University of California San Diego

{bayesteh@ohsu.edu, izhak@google.com and ababaeian@ucsd.edu}

ABSTRACT

In this paper, we continue our investigation into copula models for real-valued multivariate features with the goal of compensating for the mismatch in the training and the testing conditions. Previously, we reported results on UCI classification tasks where our method consistently outperformed other competing classifiers [1]. Here, we extend this work from classification to recognition and elaborate further on the mathematical properties of our models in the form of lemmas. We report results on the Aurora 4 automatic speech recognition (ASR) task which contains utterances with wide range of background noise that are not well represented in the training data. Our results show that the proposed copula-based models improve the accuracy by about 7% (11.6 vs 12.4) over a comparable baseline.

Index Terms— Copula model, Robust speech recognition, Deep neural network, Aurora 4

1. INTRODUCTION

The performance of current ASR systems degrades severely when there is a mismatch between training and testing condition, for example, in the presence of background noise of the type not represented in the training data or in the presence of significant amount of reverberations or channel distortions. These variations are currently modeled implicitly by the ASR acoustic models, specifically by Gaussian mixture models (GMMs) and more recently by deep neural networks (DNNs).

The deep neural networks (DNNs) are particularly effective in large vocabulary tasks with large amounts of training data. HMMs, on the other hand, are simpler and faster to train. As such, they are still employed in small tasks with limited training data. Both these models are capable of representing real-valued multivariate stochastic processes and have relatively simple estimation algorithms for learning the optimal parameters for a recognition task from labeled training data. With enough parameters, the models have the capacity to easily overfit the training data. Therefore, for good generalization, practitioners cautiously choose the optimal model size by empirically evaluating the performance on a held-out data set.

Digging a bit deeper, the features used to extract the necessary information to model speech do not explicitly factor the observed signal into the additive and convolutional components present in the input. The input features presented to the DNNs are typically the logarithm of the mel-warped frequencies and GMMs with mel-warped cepstral coefficient. Both these features have homomorphic properties where convolutional noise becomes additive but the additive noise and the speech signal interact in non-linear manner.

The strategies adopted to disentangle the additive and convolutional noise can be broadly categorized into model-based and feature-based methods. Feature-based method transform the features to a representation where the effect of additive and convolutional noise are minimized. The simplest version of such a normalization is the well-known cepstral mean-variance normalization (CMVN) that removes the convolutional channel noise in the homomorphic cepstral domain. The method assumes that the channel noise varies slowly, a mild assumption that is often true. The key advantage of this feature-based method is that it generalizes remarkably well to test utterances with channels distortions that have never been seen before. Many other feature-based transformations have been developed and investigated, but with limited success. Among them, one of them is notable in that they share the same motivation as our work [2]. They learn a coarse transformation so that the histogram of their test features matches those of their training features. Gaussianization also shares a similar motivation although in their case the transformed features are more constrained, to have Gaussians marginals. Both these approaches are *ad hoc* in that they do not take into effect the influence of the transformation into the computation of the likelihood of the input signal. In contrast, our method as described in Section 4, provides a principled mechanism to account for the transformation.

Model-based approach such as parallel model combination attempt to model the stochastic processes associated with speech and noise [3]. They disentangle the additive noise – components that are multiplicative with speech in the cepstral domain – using vector Taylor series approximation. The parameters of the model are learned in a supervised manner using training data containing representative noises. They have shown performance gains in certain tasks for GMM-based HMM acoustic models, but have not been found to be ef-

fective in large vocabulary speech recognition. For a more comprehensive review of different noise reduction techniques explored so far in the literature, see [3].

One brute force approach that has been remarkably effective and has gained much popularity recently attempts to increase the diversity of the training data by artificially distorting the input signal. This technique, often referred to as multi-style training (MTR) in the literature [4], has been particularly effective in deep neural networks where the network has sufficient representational power to model them implicitly. The effectiveness of the method depends entirely on the diversity of the simulated distortions of the input and it is a non-trivial to task to generate all combinations of potential sources of input distortions. The two common distortions employed for this purpose are reverberations and additive background noise. In the case of reverberation, the distortion is computed by convolving the input with the impulse response of a room whose dimensions are specified along with the location of the source and the noise, in addition to the type of the noise. The resultant signal is distorted further with appropriately amplified or diminished background noise of a specified type. Thus the distortion has a number of parameters, each of which belongs to an open set, making it impractical to represent all potential distortions that may be present in real-world utterance. In contrast to such a brute force method, in this paper, we focus on explicitly addressing the distributional mismatch between training and testing conditions, especially on the marginals of the input features.

Copula models provide a principled approach for decoupling the marginal distributions from the component that models the interaction between the random variables. As such, they are well-suited to address the effect of the mismatch between the train and test set. This is described in details in the following section. In fact, it can be shown that the CMVN and histogram equalization are two special cases of copula-based models. In section 3, we discuss the optimal transform that minimize the distance between the train and test set. Our experiments and the results on Aurora 4 data set are reported in Section 4. Finally,we conclude with summary of our key results.

2. COPULA MODEL

Estimating multivariate distribution is still a challenging task in probability theory and statistics. The standard approach is to focus the attention entirely on choosing a parametric form for the joint distribution of the variables. The choice of joint distribution automatically dictates a specific form for marginal distributions, which may not be appropriate for a given application or data. There is no flexibility in picking a different form of distribution for the marginals even when such a misfit is known a priori. Except for the mathematical convenience, there is no real reason why the choice of the joint and the marginals have to be tightly coupled. For example, though the marginal distributions are the same in the two distributions illustrated in the figure 1, their joint distribution are markedly different. It would be convenient if the choice of suitable marginal distribution is decoupled from that of the joint distribution. Sklar's theorem provides the necessary theoretical foundation to decouple these choices [5]. He showed that any joint distribution can be uniquely factorized into its univariate marginal distributions and a Copula distribution. The Copula distribution is a joint distribution with uniform marginal distributions on the interval $[0, 1]$. More formally, Sklar's theorem states that any continuous Cumulative Distribution Function (CDF) can be uniquely represented by a Copula CDF:

$$F(x_1, x_2, \ldots, x_n) = C(F_1(x_1), F_2(x_2), ..., F_2(x_n)) \quad (1)$$

where F is an n-dimensional CDF with the marginal CDFs $F_1(x_1), \ldots, F_n(x_n)$ and C is a CDF from the unit hyper-cube $[0, 1]^n$ to the unit interval $[0, 1]$ called Copula CDF. If joint CDF is differentiable the density function can be computed by taking the n-th derivative of Equation(1):

$$f(X) = \frac{\partial^n C(F_1(x_1), \ldots, F_n(x_n))}{\partial x_1 \cdots \partial x_n} \quad (2)$$

where $X = [x_1, x_2, \ldots, x_n]^T$. By applying the chain rule to (2),:

$$\begin{aligned} f(X) &= \frac{\partial^n C(F_1(x_1), \ldots, F_n(x_n))}{\partial F_1(x_1) \cdots \partial F_n(x_n)} \\ &\quad \times \Pi_{i=1}^n \frac{dF_{x_i}(x_i)}{dx_i} \\ &= c(F_1(x_1), F_2(x_2), \ldots, F_n(n)) \Pi_{i=1}^n f_i(x_i) \end{aligned} \quad (3)$$

where $f_1(x_1), \ldots, f_n(x_n)$ are the marginal densities of f and $c(\cdot)$ is the Copula density function.

Equation (3) shows that any continuous density function can be constructed by combining a Copula function and a set of marginal distributions. Furthermore, the Copula function can be chosen independent of the marginal distribution. Equation (3) suggests a method for estimating the multivariate density. Since the estimation of the marginal densities are straightforward, the problem of density estimation can be reduced to the estimation of the Copula density function.

3. GAUSSIAN COPULA MODEL

3.1. Definition

Gaussian Copula density is the most common multivariate Copula function and it can be obtained by applying the method of inversion to standard multivariate Gaussian [6]:

$$\begin{aligned} c_{gaus}(U; R) &= \frac{1}{|R|^{\frac{1}{2}}} \exp\{-\frac{1}{2} U^T (R^{-1} - I) U\} \quad (4) \\ R_{ij} &= \frac{cov(x_i, x_j)}{\sqrt{var(x_i) var(x_j)}} \end{aligned}$$

where R is the correlation matrix.

The Gaussian Copula model can be constructed by substituting the Gaussian Copula density function into Equation (4):

$$f(X; R, \Lambda) = c_{gaus}(U; R) \prod_{i=1}^{n} f_i(x_i; \lambda_i) \quad (5)$$

where $u_i = \Phi^{-1}(F_i(x_i))$ and Φ^{-1} is the quantile function of standard normal distribution.

The main difference between the Gaussian Copula model in Equation (5), and standard Gaussian distribution is that the marginal density functions in the Gaussian distribution are necessarily Gaussian while the marginal density functions of the Gaussian Copula model can by any continuous density and this capability makes the Gaussian Copula model more flexible than the Gaussian distribution.

4. PROPOSED MODEL

In this section, we propose a systematic way based on the copula model to convert a multivariate distribution $f(X)$ into another multivariate distribution $g(Y)$ where X and Y are two random vectors of size n . For this transformation, we assume that each distribution is modeled by the its own Gaussian copula density function with a correlation matrix R_f and a set of the marginal densities :

$$f(X) = c_f(u_1, .., u_n; R_f) \prod_{i=1}^{n} f_i(x_i) \quad (6)$$

$$c_f(U_f; R_f) = \frac{1}{|R_f|^{\frac{1}{2}}} \exp\{-\frac{1}{2} U_f^T (R_f^{-1} - I) U_f\} \quad (7)$$

The support of every element of U_f is $[0, 1]$.

4.1. Finding an optimal transformation

Given $X \sim f(X)$ and $Y \sim g(Y)$ which are the distribution of test and training data respectively we are looking for an optimal transformation to convert the distribution $f(X)$ into $g(Y)$ to alleviate the mismatch between them. We use the assumption in (6) and consider three different cases including the general case where both test and train data follow a distribution of form (6) and we show the transformation shown in figure 1 is the optimal one for mapping the test data to a new space. An optimal transform is a transform that minimize a Kullback-Leibler divergence between f and g. The Kullback-Leibler divergence is always non-negative, $D_{KL}(f||g) \geq 0$, which is also known as the Gibbs' inequality, with $D_{KL}(f||g) = 0$ if and only if $f = g$ every where.

Lemma 1: Joint Multivariate Gaussian Distributions
For joint multivariate Gaussian distributions, $f(X) = N(0, \Sigma_x)$ and $g(Y) = N(0, \Sigma_y)$, the mapping $Y = \Sigma_y^{1/2} \Sigma_x^{-1/2} X$ is optimal since it reduces the Kullback-Leibler divergence between f and g to zero. For this special case F_i and G_i in Eq. (6) would be identity functions such that $F_i(x_i) = x_i$ and $G_1^{-1}(y_i) = y_i$ and the optimal mapping can be represented as a linear weight layer in a neural network as illustrated in Figure 1.

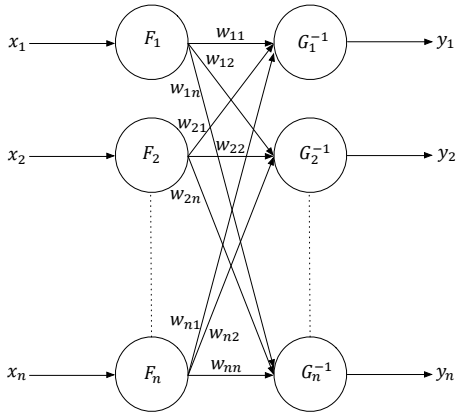

Fig. 1. The three components of the transformation necessary to reduce the KL divergence between the distribution of the training data x and the test data y.

Lemma 2: Independent Multivariates
For independent multivariate random variables, $f(X) = \prod_{i=1}^{n} f_i(x_i)$ and $g(Y) = \prod_{i=1}^{n} g_i(y_i)$ with cumulative distributions F_i and G_i respectively, the optimal mapping is $h_i = G_i^{-1}(F_i) : x_i \rightarrow y_i$, assuming G_i is invertible. In this case, the copula term c_f in Eq. (6) is identity and the resulting KL divergence is the minimum attainable value of zero. In the Figure 1, this corresponds to weights with values $w_{ii} = 1$ and $w_{ij} = 0$ for $i \neq j$.

Lemma 3: Joint Distributions with Gaussian Copulas
For two distributions, $f(X) = c_f(u_1, ..., u_n; R_f) \prod_{i=1}^{n} f_i(x_i)$ and $g(Y) = c_g(v_1, ..., v_n; R_g) \prod_{i=1}^{n} g_i(y_i)$, with Gaussian Copula model

$$c_j(U_j; R_j) = \frac{1}{|R_j|^{\frac{1}{2}}} exp\{-\frac{1}{2} U_j^T (R_j^{-1} - I) U_j\}$$

for $j = f, g$, the optimal transformation consists of $h_i = G_i^{-1}(WF) : X \rightarrow y_i$ and a linear transformation $V = R_g^{1/2} R_f^{-1/2} U$ which results in KL divergence of zero.

Proof: Consider, the Kullback-Leibler divergence between c_f and c_g, $D_{KL}(c_f(U; R_f), c_g(U; R_g))$.

$$= \int_{-\infty}^{\infty} c_f(U; R_f) ln \frac{c_f(U; R_f)}{c_g(U; R_g)} \frac{\exp^{-U^T I U}}{\sqrt{2\pi}^n} dU$$

$$= \int_{-\infty}^{\infty} \mathbb{N}_n(U; 0, R_f) ln \frac{c_f(U; R_f)}{c_g(U; R_g)} dU$$

$$= \int_{-\infty}^{\infty} \mathbb{N}_n(0, R_f) ln \frac{\frac{1}{|R_f|^{\frac{1}{2}}} exp\{-\frac{1}{2} U^T (R_f^{-1} - I) U\}}{\frac{1}{|R_g|^{\frac{1}{2}}} exp\{-\frac{1}{2} U^T (R_g^{-1} - I) U\}}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} \mathbb{N}_n(0, R_f)((\frac{1}{2} ln \frac{|R_g|}{|R_f|}) \\
&+ (-\frac{1}{2} U^T R_f^{-1} U) + (\frac{1}{2} U^T R_g^{-1} U)) dU \\
&= \frac{1}{2}(tr(R_g^{-1} R_f) + ln \frac{|R_g|}{|R_f|} - n)
\end{aligned}
$$

where n is the dimension of distributions, $u_i = \Phi^{-1}(F(x_i))$, $F(x_i) \in [0, 1]$ is the support of Gaussian copula model. Applying the linear transformation $V_g = R_g^{1/2} R_f^{-1/2} U_f$ and computing the distribution of c_g with respect to the new random vector U_f, $c_g(U_f; R_f)$ will follow the same distribution as $c_f(U_f; R_f)$ with correlation parameter R_f, thus the KL divergence between two Gaussian copula distributions c_f and c_g will be zero since $ln \frac{|R_f|}{|R_f|} = 0$ and $tr(R_f^{-1} R_f) = n$. With KL divergence of copula reduced to zero, the remaining marginals are independent random variables, as in Lemma 2, for which the optimal transformations once again are given by $h_i = G_i^{-1}(WF) : X \rightarrow y_i$. QED. In the Figure 1, this case corresponds to the choice of $W = R_g^{1/2} R_f^{-1/2}$.

5. EXPERIMENTAL RESULTS

We evaluate our proposed method on medium vocabulary speech recognition task using Aurora 4 [7]that was created by manually adding different types of noise (street traffic, train station, car, babble, restaurant, airport) to standard WSJ0. The training set we used in this paper consists of 7137 utterances from 83 speakers sampled at 16KHz. The level of the noise is changing from 5 to 15 DB in the training set. The evaluation set contains 4620 utterances from 8 different speakers with noise level ranging from 10 to 20 DB.

In Figure 2, we plot the marginals of the first two filter bank features computed on the training set (clean and multi-condition) and the test set. The plots clearly illustrate the mismatch between the training and test sets even with multi-conditional data in the training set.

We estimate a Gaussian copula model for all utterances in the training set. Then, we transform each utterance in the training set to have a similar distribution that of the entire training set, akin to the speaker adapted training in ASR models. After applying the transform, we train a DNN using this new features. During the test, we estimate the distribution of the each utterance using the Gaussian copula model and then apply the necessary transform to map them to the distribution of training data. Figure 3 shows the results of the matching the noisy set by clean train set.

ASR experiments, we used the Aurora recipe provided in Kaldi [8]. We build our ASR by creating a GMM-HMM model using the MFCC features from scratch. Then, we train DNN1 using the alignment obtained form the GMM-HMM baseline using the filter banks. We use the alignment obtained from DNN1 to train DNN2. For comparison with results in the literature, we choose the number of hidden layers and other training parameters similar to [9] . In Table 1, we report the ASR performance of our models under different conditions. In addition to transforming each feature dimension, we also include second order statistics by computing $W = R_g^{1/2} R_f^{-1/2}$ where R_g is a global correlation matrix over training set and R_f is per utterance correlation matrix (Copula-DNN-C). The matrices were estimated using Toeplitz structure [10].

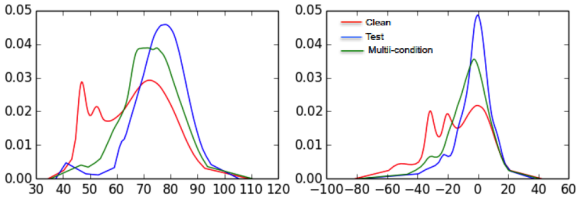

Fig. 2. Empirically computed marginals on the training set – clean and multi-conditional data – and the test set for first two filter bank features.

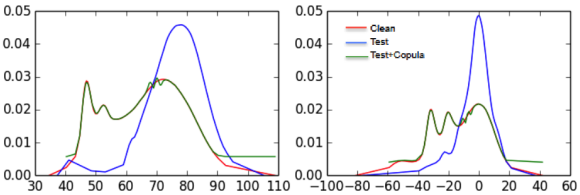

Fig. 3. Marginals for the test set before and after copula-based transformations for first two filter bank features.

Table 1. Comparison of the ASR performance with different acoustic models on Aurora 4.

	WER
GMM-HMM	19.41
DNN 1	13.95
DNN 2	13.38
DNN noise-aware training [11]	12.4
Copula-DNN 1 ($W = I$)	12.16
Copula-DNN 2 ($W = I$)	11.80
Copula-DNN-C ($W = R_g^{1/2} R_f^{-1/2}$)	11.56

6. CONCLUSION

We have presented a copula-based model for transforming the distribution of the test utterance to that of the training utterances using a Gaussian copula model. We characterize the theoretical properties of this mapping for three cases. On the Aurora 4 task, we demonstrate improvement in ASR performance over comparable baselines. The reported results can be further improved by combining this work with masking and refinements of DNNs, which are complementary to this work [12, 13].

7. REFERENCES

[1] A. Bayestehtashk and I. Shafran, "Efficient and accurate multivariate class conditional densities using copula," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, April 2015, pp. 3936–3940.

[2] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 14, no. 3, pp. 845–854, 2006.

[3] M. J. F. Gales, "Model-based approaches to handling uncertainty," *Robust Speech Recognition of Uncertain or Missing Data*, 2011.

[4] A. Acero, *Acoustical and environmental robustness in automatic speech recognition*. Springer Science & Business Media, 2012, vol. 201.

[5] A. Sklar, "Fonctions de repartition a n dimensions et leurs marges," *Publ. Inst. Stat. Univ. Paris 8*, pp. 229–231, 1959.

[6] P. Trivedi and D. Zimmer, "Copula modeling: An introduction for practitioners," *Foundations and Trends in Econometrics*, vol. 1, pp. 1–111, 2005.

[7] N. Parihar and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," *Inst. for Signal and Information Process, Mississippi State University, Tech. Rep*, vol. 40, p. 94, 2002.

[8] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[9] C. Weng, D. Yu, S. Watanabe, and B.-H. Juang, "Recurrent deep neural networks for robust speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing(ICASSP)*, May 2014, pp. 5532–5536.

[10] A. Bayestehtashk and I. Shafran, "Parsimonious multivariate copula model for density estimation," in *Proc. IEEE ICASSP*, May 2013, pp. 5750–5754.

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

[12] A. Narayanan and D. Wang, "Joint noise adaptive training for robust automatic speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 2504–2508.

[13] B. Li and K. C. Sim, "A spectral masking approach to noise-robust speech recognition using deep neural networks," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 8, pp. 1296–1305, Aug 2014.