Robust Automatic Speech Recognition for the 4th CHiME Challenge Using Copula-based Feature Enhancement

Alireza Bayestehtashk [1], *Izhak Shafran*[2]

[1]Oregon Health & Science University
[2]Google Inc

bayesteh@ohsu.edu, izhak@google.com

Abstract

In this paper, we investigate the application of the copula model for enhancing features in automatic speech recognition task. We compute a set of utterance-specific nonlinear transformations based on the copula model and use these transformations to obtain the enhanced features for every utterance in the dataset. These features improve the performance of the baseline system by about 4.3%, 1.4%, and 0.5% (absolute) respectively for 1-channel, 2-channel and, 6-channel. Further gains were obtained when our system was combined with the baseline system using minimum Bayes risk decoding to achieve 4.3%, 2.4%, and 1.2% absolute WER improvements for the respective channels.

1. Background

Generally, the mismatch between the training and testing conditions degrades the performance of machine learning tasks including automatic speech recognition (ASR). In real-world ASR applications, it is impractical to obtain training data that is representative of wide range of background noise and reverberations under which utterances are spoken, even when training data is modified using additive noise and simulated reverberations such as in multi-style training (MTR). These variations are currently modeled implicitly by the ASR acoustic models, such as deep neural networks (DNNs), recurrent neural networks (RNNs) and Gaussian mixture models (GMMs). The typical input features presented to the acoustic models are the logarithm of the mel-warped frequencies after passing it through a filter bank or mel-warped cepstral coefficient (MFCC).

The strategies to compensate the mismatch between the training and testing can be categorized into model based and feature based methods. The model-based methods attempt to model the variations associated with speech and neglect other variations such as background noise or channel distortion.

Feature mismatch reduction: In this approach features are extracted in a manner that minimizes the effect of additive and convolutional noise. 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. One such previously developed approach shares the same motivation as our work [1]. They learn a coarse transformation so that the histogram of their test features matches those of their training features.

These approaches are *ad hoc* in that they treat each feature component independently and do not take into account the joint distribution of the feature vector. Moreover, they do not consider the influence of the transformation in computing the likelihood of the input signal. 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. In our previous study [2], we showed that the CMVN and histogram equalization are two special cases of copula-based models.

In state-of-the-art ASR systems, CMVN is the only feature processing used to address mismatch between the training and testing condition. This assumes that components of input feature vectors are statistically independent, which is typically a poor assumption. In the section below, we propose a method to avoid this assumption and address the mismatch using a very flexible multivariate distribution – the multivariate copula model.

2. The Multivariate Copula Model

The standard multivariate distribution estimation methods such as GMM entirely focus 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. 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. The theory formally states 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 $[0, 1]$:

$$f(X) = c(F_1(x_1), F_2(x_2), \ldots, F_n(n))\Pi_{i=1}^{n} f_i(x_i) \qquad (1)$$

where $\{f_i(x_i)\}_{i=1}^{n}$ are the marginal density functions of f, $\{F_i(x_i)\}_{i=1}^{n}$ their corresponding marginal cumulative distribution functions, and $c(\cdot)$ is the Copula density function.

Equation (1) shows that any continuous density function can be constructed by combining a Copula density function and a set of marginal density functions.

Gaussian Copula model: Gaussian Copula density function is the most common multivariate Copula function:

$$c_{gaus}(U; R) = \frac{1}{|R|^{\frac{1}{2}}} \exp\{-\frac{1}{2}U^T(R^{-1} - I)U\} \qquad (2)$$

where R is the correlation matrix.

The Gaussian Copula model can be constructed by substituting the Gaussian Copula density function into Equation (1):

$$f(X; R, \Lambda) = c_{gaus}(U; R) \prod_{i=1}^{n} f_i(x_i; \lambda_i) \qquad (3)$$

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

The main difference between the Gaussian Copula model in Equation (3), 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 function and this capability makes the Gaussian Copula model more flexible than the Gaussian distribution.

In our previous work, we have shown how to compute the optimal feature transformation to minimize the KL distance between two multivariate Gaussian copula distributions [2].

3. Experimental Setup & Results

Akin to speaker adapted training, we estimate the acoustic models in 3 stages: (a) estimate a canonical multivariate copula distribution of the 13-dim MFCC features using all the utterances in the single channel noisy training data; (b) transform each utterance in the training data to reduce the KL distance between the multivariate distribution of the given utterance and the canonical distribution; and (c) train a standard acoustic model in the transformed feature space. At test time, we transform the features of each utterance to the canonical multivariate copula distribution space before decoding.

Compared to the performance of the baseline system [3], tabulated in Table 1 for different conditions, our copula-based system, in Table 2 shows significant improvement in several conditions, but not all. Note, 5gkn stands for 5-gram Knesser-Ney smoothed LM provided with the baseline system. The gains are particularly remarkable in single channel input for which it is well-suited. Note, we haven't applied any special processing for multi-channel case and hence didn't expect gains there. The gains are highest in bus background noise and our hypothesis is that there is more structure and correlation in the noise in this case for which the multivariate copula is an apt representation. We expect applying copula-based feature enhancement to give further improvements when it is applied to frequency spectrum before the filterbank and MFCC where the noise components can be modeled in a fine grained manner. Finally, our copula-based system is sufficiently different from the baseline system that we are able to obtain additional gain through system combination using MBR, as reported in Table 3.

Table 1: Average WERs of the baseline systems trained on single channel data.

Track	System	Dev		Test	
		simu	real	simu	real
1ch	DNN	17.4	16.5	26.0	30.0
	smbr	15.8	14.6	24.0	27.1
	smbr+5gkn	13.9	12.3	22.1	24.3
	smbr+rnn	12.8	11.5	20.8	22.9
2ch	GMM	18.7	16.3	27.3	28.7
	DNN	13.5	12.2	20.4	22.4
	smbr	12.1	10.8	18.8	20.0
	smbr+5gkn	10.7	9.6	16.4	17.6
	smbr+rnn	9.3	8.4	15.2	16.2
6ch	GMM	14.2	12.7	21.1	21.7
	DNN	10.1	9.5	15.9	16.6
	smbr	9.0	8.2	14.2	14.7
	smbr+5gkn	7.8	7.0	12.1	12.8
	smbr+rnn	6.7	6.0	10.9	11.3

Table 2: Average WERs of the baseline systems trained on single channel features after copula-based transformation.

Track	System	Dev		Test	
		simu	real	simu	real
1ch	GMM	23.0	19.8	30.0	29.4
	DNN	17.6	15.4	24.9	24.4
	smbr	16.5	13.9	23.5	23.1
	smbr+5gkn	14.7	12.1	21.7	20.1
	smbr+rnn	13.2	10.7	20.4	18.6
	copula+baseline	12.1	9.8	19.2	18.6
2ch	GMM	18.1	15.2	24.9	24.4
	DNN	13.9	12.1	20.4	19.8
	smbr	12.7	10.7	19.1	18.2
	smbr+5gkn	10.9	9.1	17.2	16.4
	smbr+rnn	9.6	8.0	15.6	14.8
	copula+baseline	8.8	7.3	13.9	13.8
6ch	GMM	14.4	12.5	19.7	19.3
	DNN	10.8	9.6	16.0	15.4
	smbr	9.8	8.2	15.2	14.5
	smbr+5gkn	8.2	7.1	13.0	12.2
	smbr+rnn	7.1	6.1	11.7	10.8
	copula+baseline	6.3	5.4	10.1	10.1

Table 3: Average WERs after combining the baseline and copula-based system using MBR decoding.

Track	Envir.	Dev		Test	
		simu	real	simu	real
1ch	BUS	10.3	12.6	13.8	26.0
	CAF	15.7	10.5	23.5	20.8
	PED	9.3	6.6	18.8	15.7
	STR	12.9	9.6	20.6	11.9
2ch	bus	7.2	9.2	10.0	19.4
	CAF	11.8	7.5	16.2	14.1
	PED	6.9	4.9	14.2	12.0
	STR	9.1	7.7	15.2	9.7
6ch	bus	5.3	6.8	6.7	13.3
	CAF	7.7	5.1	11.2	9.5
	PED	5.1	3.9	10.0	8.5
	STR	7.2	5.7	12.5	9.1

4. References

[1] 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.

[2] A. Bayestehtashk, I. Shafran, and A. Babaeian, "Robust speech recognition using multivariate copula models," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2016, pp. 5890–5894.

[3] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech and Language*, to appear.