FEATURE SPACE GAUSSIANIZATION

George Saon, Satya Dharanipragada and Dan Povey

IBM T. J. Watson Research Center, Yorktown Heights, NY, 10598
e-mail: {saon,dsatya,dpovey}@watson.ibm.com

ABSTRACT

We propose a non-linear feature space transformation for speaker/environment adaptation which forces the individual dimensions of the acoustic data for every speaker to be Gaussian distributed. The transformation is given by the preimage under the Gaussian cumulative distribution function (CDF) of the empirical CDF on a per dimension basis. We show that, for a given dimension, this transformation achieves minimum divergence between the density function of the transformed adaptation data and the normal density with zero mean and unit variance. Experimental results on both small and large vocabulary tasks show consistent improvements over the application of linear adaptation transforms only.

1. INTRODUCTION

Speaker adaptation is a key technique that is used in most state-of-the-art speech recognition systems. Traditionally, it consists in finding one or more linear transforms such that, when it is applied to either the Gaussian means [6] or, as in constrained MLLR, to the feature vectors themselves [5], the likelihood of the acoustic data associated with an utterance is maximized with respect to an initial word hypothesis. The utterance is then re-decoded after applying the transforms to either the models or to the features or both (as for unconstrained variance transforms).

In recent years, the family of feature space transformations for speaker adaptation has been extended by Dharanipragada and Padmanabhan [4] through the addition of a new class of non-linear transforms obtained by matching the overall cumulative distribution function (CDF) of the adaptation data to the CDF of the training data on a per dimension basis. In addition to having more potential over linear transforms for severely mismatched decoding conditions, this non-linear mapping also has the advantage that it does not require a first pass decoding step, i.e. it is completely unsupervised.

Independently, Chen and Gopinath [2] have proposed a Gaussianization transformation for high-dimensional data modeling which alternates passes of linear transforms for achieving dimension independence and passes of marginal Gaussianization of the individual dimensions through univariate techniques. At first sight, the two previous techniques have little in common. The link becomes apparent if we use the distribution matching technique (also called histogram equalization) to match the CDF of the speaker data to the CDF of a Gaussian on a per dimension basis which is exactly what we propose to do in this paper. Indeed, marginal Gaussianization can be performed either parametrically by assuming a Gaussian CDF mixture model for the data as in [2] or non-parametrically by using the empirical CDF or a binned version thereof as in [4]. The advantage of the latter is that it bypasses the problems associated with choosing the size (complexity) of the mixture models while having the drawback that it requires more adaptation data to get a reliable estimate of the CDF if no smoothing is to be performed.

There are two advantages of Gaussianization for ASR systems. The first one has to do with the fact that, in most systems, the HMM output distributions are modeled with mixtures of diagonal covariance Gaussians. It is therefore reasonable to expect that gaussianizing the features will enforce this particular modeling assumption. The second advantage is that both test and *training* speakers are warped to the same space which naturally leads to a form of speaker adaptive training (SAT) [1] through non-linear transforms. The benefit of retraining the models on CDF-warped training data in the context of the histogram equalization algorithm has been highlighted in [7].

The paper is organized as follows: in section 2, we outline the derivation of the transform. In section 3, we present some experimental evidence of its utility followed by some concluding remarks in section 4.

2. GAUSSIANIZATION TRANSFORM

Let $\mathbf{X} \in \mathbb{R}^n$ be the random variable (r.v.) describing the adaptation data for a given speaker. The differentiable and invertible function $T : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a Gaussianization transformation if the random variable $\mathbf{Y} = T(\mathbf{X})$ is normally distributed i.e.

$$T(\mathbf{X}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Finding the joint Gaussianization transform is in general a difficult problem (see [2]). We will make the simplifying assumption that the dimensions of $\mathbf{X}$ are statistically independent. The problem can be recast as finding n independent mappings $T^{(1)}, \ldots, T^{(n)}$ such that

$$T^{(i)}(X^{(i)}) \sim \mathcal{N}(0,1), \qquad 1 \leq i \leq n$$

where $X^{(i)}$ represents component i of the random variable $\mathbf{X}$. From now on, we will deal only with one-dimensional problems and for the sake of clarity we will drop the superscripts related to the dimension whenever they are not necessary. Consider X the r.v. corresponding to a particular dimension and let p_X be its probability density function. Moreover, let us denote the standard normal PDF by ϕ and its CDF by Φ that is:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

and

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)dt$$

Correspondingly, let F_X be the CDF of X i.e.

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} p_X(t)dt$$

We aim at finding a differentiable and invertible transform T which minimizes the Kullback-Leibler divergence between p_Y and ϕ where p_Y is the PDF of $Y = T(X)$. Stated otherwise, we look for

$$\begin{aligned} \hat{T} &= \operatorname*{argmin}_{T} D(\phi \parallel p_Y) \\ &= \operatorname*{argmin}_{T} \int_{\mathbb{R}} \phi(y) \log \frac{\phi(y)}{p_Y(y)} dy \end{aligned} \tag{1}$$

Now p_X and p_Y are related through the following equation

$$p_Y(y) = \frac{p_X(T^{-1}(y))}{|T'(T^{-1}(y))|} = p_X(T^{-1}(y))|T^{-1'}(y)| \tag{2}$$

where $|T^{-1'}(y)|$ represents the absolute value of the determinant of the Jacobian of the transformation which for one-dimensional transforms is simply the derivative. Assuming that T is monotonically increasing (recall that T is invertible) we can drop the absolute value in (2).

It is known that the divergence is minimized when the two distributions are pointwise the same, that is

$$\phi(y) = p_Y(y) = p_X(\hat{T}^{-1}(y))\hat{T}^{-1'}(y), \qquad \forall y \in \mathbb{R} \tag{3}$$

Next, we will attempt to solve the differential equation (3) in order to find $\hat{T}$. First, since (3) holds for all y, we can integrate both sides from $-\infty$ to $\hat{T}(x)$ and we get

$$\begin{aligned} \int_{-\infty}^{\hat{T}(x)} \phi(y)dy &= \int_{-\infty}^{\hat{T}(x)} p_X(\hat{T}^{-1}(y))\hat{T}^{-1'}(y)dy \\ &= \int_{\hat{T}^{-1}(-\infty)}^{x} p_X(t)dt \end{aligned} \tag{4}$$

where the latter equality follows from applying the substitution rule $t = \hat{T}^{-1}(y)$ in the second integration. Now, assuming $\hat{T}^{-1}(-\infty) = -\infty$, we further get

$$\Phi(\hat{T}(x)) = F_X(x) \tag{5}$$

or equivalently:

$$\hat{T}(x) = (\Phi^{-1} \circ F_X)(x), \qquad \forall x \in \mathbb{R} \tag{6}$$

which means that the desired transformation is given by the preimage of F_X under the Gaussian CDF Φ. It can be easily verified that $\hat{T}$ is monotonically increasing with $\hat{T}^{-1}(-\infty) = -\infty$ which is consistent with our previous assumptions. Also note that if $\hat{T}$ is a solution to (1) then $-\hat{T}$ is a solution as well.

Now, since F_X is not available we can approximate it with the empirical CDF

$$F_0(x) = \frac{1}{N} \sum_{i=1}^{N} \theta(x - x_i) \tag{7}$$

with θ denoting the step function and where $x_1, \ldots, x_N$ are N samples drawn from p_X (the adaptation data for a particular dimension). This is in contrast with the work of [2] where the author uses a mixture of Gaussian CDF's as an approximator for F_X. From a practical standpoint, we note that

$$F_0(x_i) = \frac{rank(x_i)}{N} \tag{8}$$

where $rank(x_i)$ is the rank of x_i in the sorted list of samples. Combining (6) with (8) yields the final form of the Gaussianization transform

$$y_i = \Phi^{-1}\left(\frac{rank(x_i)}{N}\right), \qquad 1 \leq i \leq N \tag{9}$$

3. EXPERIMENTS AND RESULTS

We experimented with two different databases: an in-car small vocabulary task and the Switchboard corpus which is a large vocabulary conversational telephone speech database. The Gaussianization transform is implemented as a simple table lookup where the entries are given by the inverse Gaussian CDF (Φ^{-1}) sampled uniformly in $[0, 1]$. In our experiments, we used one million samples. For each dimension of a speaker's data, we first sort all the samples then apply equation (8) to locate the table entry. Figure 1 shows a typical transform and the corresponding original and transformed distributions.

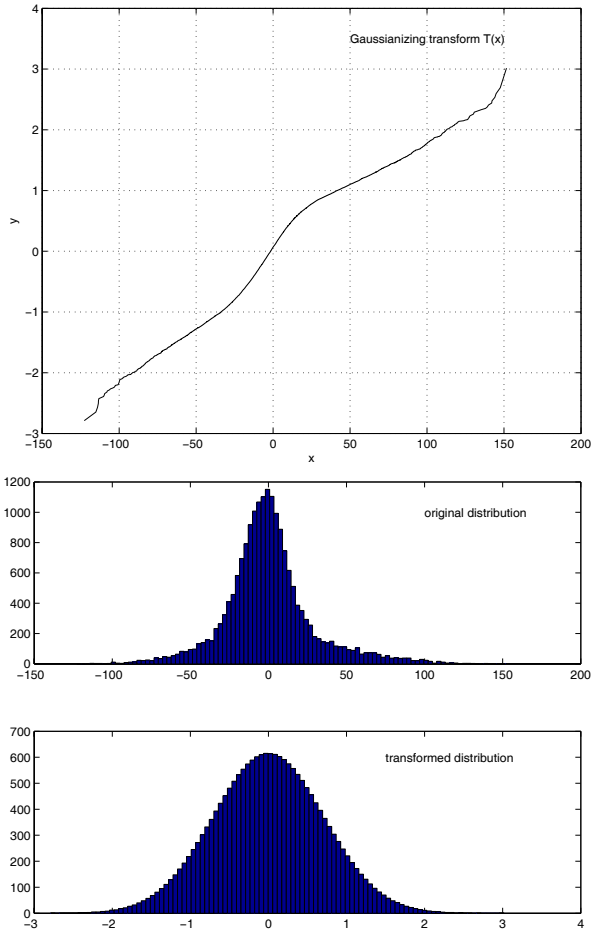

Figure 1: Example of transform and distributions.

3.1. In-car Database

We evaluated the Gaussianization transform on an in-car database. The training data consisted of speech collected in several stationary and moving (30 mph and 60 mph) cars with microphones placed at a few different locations – rear-view mirror, visor and seat-belt. We created additional data by synthetically adding noise, collected in a car, to the stationary car data. Overall, with the synthesized noisy data, we have about 480 hours of training data.

The acoustic model comprised of context-dependent subphone classes (allophones). The context for a given phone is composed of only one phone to its left and one phone to its right and does not extend over word boundaries. The allophones are identified by growing a decision tree using the context-tagged training feature vectors and specifying the terminal nodes of the tree as the relevant instances of these classes. Only the clean (stationary car) data was used to grow the decision tree. Each allophone is modeled by a single-state Hidden Markov Model with a self loop and a forward transition. The training feature vectors are poured down the decision tree and the vectors that are collected at each leaf are modeled by a Gaussian Mixture Model (GMM), with diagonal covariance matrices. The Gaussians were distributed across the states using BIC based on a diagonal covariance system. The acoustic models used separate digit phonemes with a total of 89 phonemes. Overall, we had 680 HMM states in our acoustic model.

Standard 13-dimensional MFCC vectors were extracted at 15 ms intervals. Each cepstral vector was concatenated with 4 preceding and 4 succeeding vectors to create a composite vector of dimension 117. This composite vector was then projected onto a $n = 39$ dimensional space using Linear Discriminant Analysis (LDA). The projected features were further transformed using a Maximum Likelihood Linear Transform (MLLT) [5]. More details about the system can be found in [3].

We report word error rates on a test set comprised of small vocabulary grammar based tasks (addresses, digits, command and control) and consists of 73743 words. Data for each task was collected at 3 speeds: idling, 30mph and 60mph.

Five different models, each with about 10K Gaussians, were evaluated on this test set and their results are reported in Table 1:

- A baseline model trained on 39-dimensional LDA+MLLT features.
- A model where each training and test speaker underwent a non-linear Gaussianization.
- A model where each training and test speaker data was transformed with a linear FMLLR transform.
- A model where each training speaker data was Gaussianized and where each test speaker data was gaussianized followed by a linear FMLLR transform.
- A model where each training and test speaker data was Gaussianized followed by a linear FMLLR transform.

Systems	0mph	30mph	60mph	all
Baseline	1.47	2.62	6.52	3.54
Gaussianized	1.32	2.36	4.69	2.79
FMLLR-SAT	1.16	1.77	3.80	2.25
Gaussianized+ FMLLR	0.93	1.72	3.33	2.00
Gaussianized+ FMLLR-SAT	1.05	1.71	3.39	2.06

Table 1: Word error rates on an in-car database of small vocabulary tasks

.

3.2. Switchboard database

The second set of experiments were conducted on the Switchboard database. The test set consists of 72 two-channel conversations (144 speakers) totaling 6 hours used by NIST during the RT'03 conversational telephone speech evaluation. The recognition system uses a phonetic representation of the words in the vocabulary. Each phone is modeled with a 3-state left-to-right HMM. Further, we identify the variants of each state that are acoustically dissimilar by asking questions about the phonetic context (within an 11-phone window) in which the state occurs. The questions are arranged hierarchically in the form of a decision tree, and its leaves correspond to the basic acoustic units that we model. The output distributions for the leaves are given by a mixture of at most 128 diagonal covariance Gaussian components totaling around 158K Gaussians. The Gaussians were trained on VTL-warped PLP cepstral features transformed to 60 dimensions through the application of LDA followed by MLLT. In addition, we performed speaker adaptive training in feature space by means of constrained MLLR transforms [5]. More details about the baseline system can be found in [9]. Feature space Gaussianization is applied on the final 60-dimensional SAT features (that is after VTLN, LDA+MLLT and the feature space MLLR transforms). In Table 2, we show a comparison between two sets of systems trained on original and gaussianized features: systems trained using maximum likelihood and systems trained using a minimum phone error (or MPE) criterion which is a variant of MMIE training [8].

4. CONCLUSION

We presented a non-linear dimensionwise Gaussianization transform for speaker/environment adaptation. This transformation achieves minimum divergence between the density function of the transformed adaptation data and the normal density with zero mean and unit variance. Clearly, the target distribution for the transformation can have an arbitrary form although the choice of a normal distribution facilitates the use of diagonal covariance Gaussians in the final acoustic model. We have presented experimental evidence on both a small and a large vocabulary task showing that non-linear Gaussianization provides additional gains on top of standard linear feature space transforms (11% relative improvement for the in-car database and 2% for Switchboard).

Features	ML	MPE
baseline (FMLLR-SAT)	30.9%	29.1%
FMLLR-SAT+Gaussianized	30.5%	28.5%

Table 2: Word error rates on original and gaussianized features using ML and MPE trained models.

5. REFERENCES

[1] T. Anastasakos, J. McDonough, R. Schwartz, J. Makhoul. A compact model for speaker-adaptive training. In Proc. ICSLP'96, Philadelphia, 1996.

[2] S. Chen and R. Gopinath. Gaussianization. In Proc. NIPS'00, Denver, 2000.

[3] S. Deligne, S. Dharanipragada, R. Gopinath, B. Maison, P. Olsen, H. Printz, A robust high-accuracy speech recognition system for mobile applications. In IEEE Transactions on Speech and Audio Processing, 10:08, 2002.

[4] S. Dharanipragada, M. Padmanabhan. A non-linear unsupervised adaptation technique for speech recognition. In Proc. ICSLP'00, Beijing, 2000.

[5] M. J. F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report CUED/F-INFENG, Cambridge University Engineering Department, 1997.

[6] C. J. Leggetter and P. C. Woodland. Speaker adaptation of HMMs using linear regression. Technical Report CUED/F-INFENG, Cambridge University Engineering Department, 1994.

[7] S. Molau, H. Ney and M. Pitz. Histogram based normalization in the acoustic feature space. In Proc. ASRU'01, Italy, 2001.

[8] D. Povey and P. Woodland. Minimum phone error and I-smoothing for improved discriminative training. In Proc. ICASSP'02, Orlando, 2002.

[9] G. Saon, B. Kingsbury, L. Mangu, G. Zweig and U. Chaudhari. An architecture for rapid decoding of large vocabulary conversational speech. In Proc. Eurospeech'03, Geneva, 2003.