Feature Gaussianization for Speech Recognition

Pavel Matějka, Jan Černocký,
Faculty of Electrical Engineering and Communication,
Brno University of Technology
Purkyňova 118, 612 00 Brno, Czech Republic
Phone: +420 5 41149156, E-mail: matejkap@feec.vutbr.cz

In Hidden Markov models, speech data are modeled by Gaussian distributions. In this paper, we propose to Gaussianize the features to better fit to this modeling. A distribution of the data is estimated and a transform function is derived. We test three methods of the transform estimation (global, speaker based, frame based) and report results on the SPINE 2000 task with Sphinx recognizer. We conclude that the proposed method is a cheap way to increase the recognition accuracy.

1 Introduction

Gaussianization is a process, where the data are transformed to data with Gaussian distribution. This idea was inspired by data distribution modeling in HMM. Here the distributions are modeled by Gaussian mixtures. In ideal situation, the data should have Gaussian distribution per class (for example phonemes). But we don't know the classes apriori. Therefore our approach will be the global gaussianization of the data. This gaussianization process is universal, because it can be added to every speech recognizer (fig. 1) only by inserting this to data preprocessing.
In section 2 we present general method of data gaussianization. In section 3 several principles of gaussianization are presented. And section 4 mentions the results of our experiments

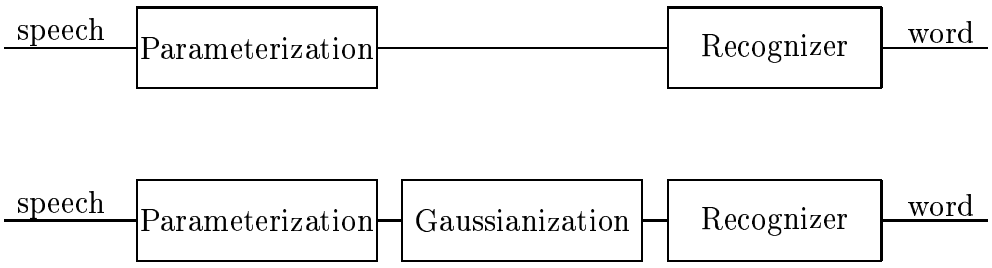

Fig. 1: Simple addition of gaussianization to recognizer

2 Gaussianization of data

The goal of gaussianization is to transform data, which has non Gaussian distribution to data, which has Gaussian distribution. It is feasible if we find the transform function. First it is necessary to estimate distribution of original data and define a Gaussian distribution to which we want to approach the distribution of our destination data. Cumulative distribution function is computed from distributions. Final step is to find for each point of this function of original data projection to appropriate point in Gaussian distribution function. Then we obtain the final transform function. Simple example of non Gaussian signal and its transformation is shown in fig. 2. One speech feature, on the left panel of fig. 2, has non Gaussian distribution. It could be approximated by a mixture of 2 Gaussians with different mean values and dispersions. Then transform function is computed (fig. 2 b). Distribution of transformed data is shown in (fig. 2c).

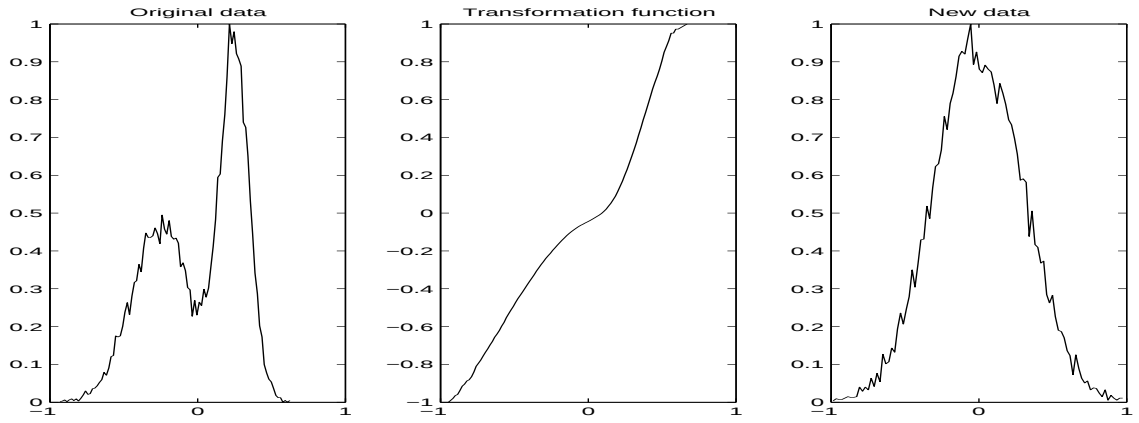

Fig. 2: Example of a)original distribution b)transform function c)new distribution

3 Methods of gaussianization

There are several approaches to the estimation of transform function. It is possible to compute it over the whole database, speakers, utterances or only for some time section around the destination frame. We may expect better results if we use smaller unit, because the method can adapt better to data. But it is true only to some extent - if we use smaller one the recognition becomes worse, because there are not enough data for reliable estimation of the transform function. In this paper results from all these categories are presented.

3.1 Global gaussianization

Global gaussianization is the first method. This means that only one transform function is estimated for all data. The histogram (distribution) is made from all data

and then the transform function is computed.

3.2 Gaussianization per speaker

This method computes transform function from all utterances, which belong to one speaker. This means that there is one transform function for each speaker.

3.3 Gaussianization per frame

Transform function were computed for each frame. All utterances belonging to one speaker are concatenated and the transform function is computed from several frames around the actual one. The reason is that some utterances are too short. This technique is shown in fig. 3.

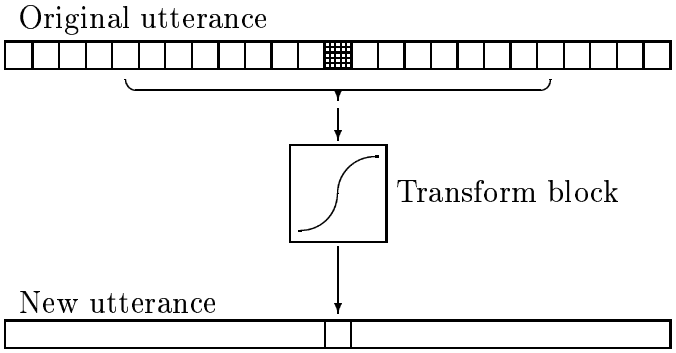

Fig. 3: Scheme of gaussianization per frame

4 Experiments and Results

4.1 The data

Database SPINE [3] was choosen for experiments. SPINE (Speech in Noisy Environments) is an evaluation run by the Naval Research Laboratory. The task a medium-sized vocabulary recognition on several military environments. The training and evaluation data from 2000 were used to assess performances of our features. These data come as stereo-recordings, but we disposed of data pre-segmented at CMU (Carnegie Mellon University) into speech and silence regions. The recognizer used – SPHINX – came also from CMU. The **training data** consists of 140 conversation (each has 2 channels) completed with 18 directories with DRT (Diagnostic Rhyme Test) words with added noises. There are 15847 files in the training set. The **evaluation data** consists of 120 conversations (each with 2 channels). There are 13265 files in the evaluation set. 12 first conversations were selected as the short evaluation set, including 1353 files. Every of the results reported here were obtained on this short set.

4.2 The experiments

For parameterization 12 MFCC coefficients [1] plus energy and their delta and delta-delta coefficients are used. These coefficients are computed in 25 ms window with 10 ms overlap. Each frame is weighted by Hamming window and is used preemphaze with $a = 0.97$.
Histograms (estimates of distributions) were computed with 200 points. This number was used for global and speaker gaussianization, but not for frame, because there is not enough data for computation. The transform function in gaussianization per frame was computed on 66 frames around the destination one. This is approximately 0.7 second in time domain. Histograms were computed with 10 points.
Table 1 presents results of recognition based on context-independent (CI) and context-dependent (CD) phonemes. The table shows WER (Word Error Rate) in % for each category of gaussianisation.

recognition	CI	CD			
type/mixtures	1	1	2	4	8
base line	72.3	49.9	44.2	38.8	36.6
global	66.3	45.8	40.2	37.7	35.3
speaker	**65.7**	**43.7**	**38.7**	**35.7**	**34.1**
frame	72.7	48.4	40.8	37.0	35.0

Tab. 1: Word Error Rate [%] for recognition

5 Conclusion

The paper demonstrates the possibility to decrease WER only with some conversion of data in preprocessing. The decreasing of number of Gaussian component in models has less effect on gaussianizated data than on the original one. It means less computation demands for training models and faster testing process. The best results was achieved with gaussianization per speaker. The improvement is about 2.5 % with CD models and 8 mixture Gaussian components.

References

[1] Hidden Markov Model Toolkit (HTK), `http://htk.eng.cam.ac.uk`

[2] PSUTKA J, Komunikace s počitačem mluvenou řeči, Academica, Praha, 1995

[3] ČERNOCKÝ J, TRAPS in all senses, report of post-doctoral research internship, OGI School of Science and Engineering,Portland, Oregon, 2001