Some like it Gaussian ...*

Pavel Matějka[1], Petr Schwarz[2], Martin Karafiát[2], and Jan Černocký[2]

[1] VUT Brno, Faculty of Elec. Eng. and Communication, matejkap@feec.vutbr.cz
[2] VUT Brno, Fac. of Inf. Technology, schwarzp|karafiat|cernocky@fit.vutbr.cz

Abstract. In Hidden Markov models, speech features 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 have tested two methods of the transform estimation (global and speaker based). The results are reported on recognition of isolated Czech words (SpeechDat-E) with CI and CD models and on medium vocabulary continuous speech recognition task (SPINE). Gaussianized data provided in all three cases results superior to standard MFC coefficients proving, that the gaussianization 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 HMMs [1, 5]. Here the distributions are modeled by Gaussian mixtures. In ideal situation, the data should have Gaussian distribution per class (for example phonemes). Unfortunately, we do not know a-priori to which class a given feature vector belongs. Therefore our approach will be the global gaussianization of the data. Gaussianization was already tested on the speaker verification task [2]. This work presents its results while applied to standard MFC coefficients for speech recognition.

2 Gaussianization of data

The goal of gaussianization is to transform data, which has non Gaussian distribution denoted $p(x)$ to data, which has Gaussian distribution $\mathcal{N}(y, \mu, \sigma)$. It is feasible if we find the transform function $y = f(x)$. First it is necessary to estimate the distribution of original data $\hat{p}(x)$ using histogram and define the target Gaussian distribution – for simplicity, we choose zero mean and unity variance: $\mathcal{N}(y, 0, 1)$. Cumulative distribution functions $\hat{P}(x)$ and $P_{\mathcal{N}}(y)$ are computed from distributions. It is then easy to find, for each value of $\hat{P}(x)$ the corresponding value of $P_{\mathcal{N}}(y)$ which leads directly to the transform function $y = f(x)$. Care must be taken at the edges of this function, as the edge values of histograms are not reliably estimated.

* Supported by Grant Agency of Czech Republic under project No. 102/02/0124.

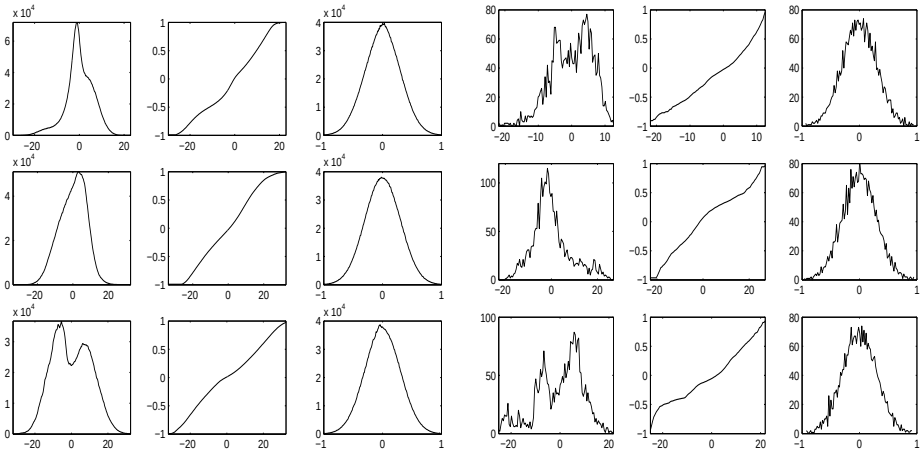

Fig. 1. Examples of distributions and transformation functions with global and per-speaker gaussianization. c_1, c_2, and c_0 are shown.

There are several approaches to the estimation of transform function. It is possible to compute it over the *whole database, per speaker, per utterance* 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 – the unit is too small, the recognition becomes worse, because there are not enough data for reliable estimation of the transform function. In this paper, results for the first two cases are presented.

3 Experiments and Results

SpeechDat – context independent models Czech SpeechDat-E database [4] was used in the first experiments to assess the gaussianization. The database consists of 1052 sessions containing various items, 700 sessions were used for training and 352 for tests.

Phoneme-based recognizer of isolated words was *trained* on phonetically balanced words from the training data (`W*` items), after discarding of some corrupted ones, 2777 words were used. For *tests*, 1394 words were used. The size of context-independent (CI) phoneme set was 42. Phoneme models had standard architecture, 3 emitting states with no state skip. The number of Gaussian components was a parameter in experiments (see the tables).

The feature extraction was done using 13 MFC coefficients, including c_0 (similar to log energy, but not the same, as we are summing the *log* energies of frequency bands). Velocity and acceleration features were added. No normalization was applied. The recognizer with those features (MFCC_0_D_A in HTK notation) was the *baseline*.

	SpeechDat-E CI models			
#mixture components	1	2	4	8
baseline	80.92	85.58	88.59	90.24
glob gauss	82.42	86.44	89.45	90.39
spk gauss	88.88	91.68	93.47	95.05
	SpeechDat-E CD models			
#mixture components	1	2	4	8
baseline	94.98	96.27	96.63	96.77
glob gauss	94.40	95.12	95.98	96.27
spk gauss	96.48	96.56	97.06	97.13

Table 1. Recognition accuracy in SpeechDat-E experiments with CI and CD models

#mixture components	1	2	4	8
baseline	65.7	55.7	47.3	44.6
glob gauss	60.9	49.5	45.4	43.5
spk gauss	44.2	39.0	36.0	34.4

Table 2. Word error rates on SPINE.

The gaussianization was first done globally for all the speech data. The histogram in the $\hat{p}(x)$ estimation had 50 points. The results are denoted *glob gauss* and some distributions and transform functions are shown in left panel of Fig. 1. Next, the data were gaussianized per speaker. Again, histograms had 50 points. This time, they were less reliably estimated (right panel of Fig. 1). In tables, the result is denoted *spk gauss*.

The results summarized in Tab. 1 show clearly the power of gaussianization: even the global one helps the recognizer to gain several percent for low number of Gaussians. Per-speaker gaussianization increases the accuracy with 8 Gaussians by almost 5 percent and makes the CI recognizer comparable to the baseline CD one.

SpeechDat – context dependent models In this experiment, context-dependent triphones were created using a set of phonetic questions. The models were initialized on phonetically balanced words (2777), but then re-trained on the entire training part (36330 items). The number of logical triphones was 77662, the numbers of physical triphones and tied states varied from one experiment to another, but were around 7500 and 2400. The same feature extraction (MFCC_0_D_A) was used and data gaussianization was performed exactly in the same way as above.

The results in 1 show that in this case, the global gaussianization hits the recognition accuracy. We have however gained 0.36% in case of speaker-based processing, which is not a negligible improvement for accuracies around 96-97%.

SPINE Third set of experiments was run on SPINE data. 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. We disposed of data pre-segmented at CMU (Carnegie Mellon University) into speech and silence regions [3]. 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.

The same features as for SpeechDat (MFCC_0_D_A) were used as the baseline. Then, data were gaussianized globally and per speaker, again with histograms estimated on 50 points. Tab. 2 presents the word-error rates (WER) [3] of the system.

The improvement of 10% can be attributed to non-optimal baseline, but it clearly shows the power of gaussianization. On contrary to some feature extraction techniques, helping the CI models but hurting CD ones, our method integrates smoothly with CD models.

4 Conclusion

The paper demonstrates the possibility to decrease WER only with a "cheap" transform in feature extractor. Moreover, the decreasing of number of Gaussian components in models has less effect on gaussianized data than on the original ones. It means less CPU power for training models and faster recognition.

References

1. B. Gold and N. Morgan. *Speech and audio signal processing.* John Wiley & Sons, 2000.
2. J. Pelecanos and S. Sridharan. Feature warping for ro-bust speaker verification. In *Proc. Speaker Odyssey 2001 conference*, June 2001.
3. R. Singh, M. L. Seltzer, B. Raj, and R. M. Stern. Speech in noisy environments: robust automatic segmentation, feature extraction, and hypothesis combination. In *Proc. ICASSP 2001*, Salt Lake City, Utah, USA, May 2001.
4. H. van den Heuvel et al. Speechdat-east: Five multilingual speech databases for voice-operated teleservices completed. In *submitted to Eurospeech 2001*, Aalborg, Denmark, September 2001.
5. S. Young, J. Jansen, J. Odell, D. Ollason, and P. Woodland. *The HTK book.* Entropics Cambridge Research Lab., Cambridge, UK, 1996.

[3] WER is used in the SPINE community rather than accuracy.