Using Full Covariance Matrix for CMU Sphinx-III Speech Recognition System

Marcin Płonkowski, Pavel Urbanovich,
The John Paul II Catholic University of Lublin, marcin.plonkowski@kul.pl

Sphinx III spech recognition system uses HMM with continuous observations modeled as multivariate Gaussian [1]. So in order to calculate the probability functions we have to invert the covariance matrix. Due to the fact that we use 39-dimensional features vectors, the covariance matrix will have a dimension of 39x39. The inverse operation of that matrix is computationally expensive. The second problem is the fact that, large amounts of training data are required for reliable full covariance estimation. If we do not have a very large dataset than the matrices are often poorly-conditioned, and do not generalise well [2]. In addition, rounding errors that will occur during the calculation will lead a determinant of a matrix to very small values and consequently to 0. This makes it impossible to invert the matrix.

Despite this shortcomings, full covariance systems have been successfully used for large vocabulary ASR. The most notable example being in the 2004 IBM system [3], where the computational cost was reduced by aggressively pruning Gaussians during the full covariance likelihood computation.

In speech recognition, we frequently assume that the feature vector dimensions are all independent of each other [4]. Then we might reduce the covariance matrix to a diagonal form. The determinant of the diagonal matrix and its inverse are easy to compute. However, due to this simplification, we lose information about the correlation of features.

In this article we analyze the speech recognition accuracy based on the publicly available AN4 database [5]. Authors analyzed use of full covariance matrix in speech recognition systems. The use of this type of matrix involves many problems, which in practice often worsen the results of the system. By using only a diagonal matrix, we lose a great deal of information about the correlation of learning vector coefficients. Hence, the authors proposed a hybrid system in which the full covariance matrix is used only at the initial stage of learning. At the further stage of learning, the amount of covariance matrix increases significantly, which, combined with rounding errors, causes problems with matrix inversion. Therefore, when the number of matrices with a determinant of 0 exceeds 1%, the system goes into the model of diagonal covariance matrices.

Thanks to this, the hybrid system has achieved a better result of about 11%. The disadvantage of this solution is almost twice the length of the algorithm's time.

References

[1] The Carnegie Mellon Sphinx Project: CMU Sphinx. http://cmusphinx.sourceforge.net/, Apr 2017
[2] Bell P., Full Covariance Modelling for Speech Recognition. PhD thesis, The University of Edinburgh 2010
[3] Chen S., Kingsbury B., Mangu L., Povey D., Saon G., Soltau H., Zweig, G., Advances in speech transcription at IBM under the darpa ears program. IEEE Transactions on Audio, Speech and Language Processing, vol. 14, no. 5, pp. 1596–1608, 2006
[4] Płonkowski M., Urbanowicz P., Tuning a CMU Sphinx-III Speech Recognition System for Polish Language, Przegląd Elektrotechniczny (4)/2014
[5] The CMU Audio Databases, AN4 database, http://www.speech.cs.cmu.edu/databases/an4/, Apr 2017