Universal Background Models*

Douglas Reynolds

MIT Lincoln Laboratory, 244 Wood St., Lexington, MA 02140, USA
dar@ll.mit.edu

Synonyms

UBM; World model; General model; Person-independent model

Definition

A Universal Background Model (UBM) is a model used in a biometric verification system to represent general, person-independent feature characteristics to be compared against a model of person-specific feature characteristics when making an accept or reject decision. For example, in a speaker verification system, the UBM is a speaker-independent Gaussian Mixture Model (GMM) trained with speech samples from a large set of speakers to represent general speech characteristics. Using a speaker-specific GMM trained with speech samples from a particular enrolled speaker, a likelihood-ratio test for an unknown speech sample can be formed between the match score of the speaker-specific model and the UBM. The UBM may also be used when training the speaker-specific model by acting as a the prior model in MAP parameter estimation.

Main Body Text

Likelihood Ratio Test

To understand the development and use of a Universal Background Model (UBM), we first must describe the likelihood-ratio test for which it is intended. Given an observation, O, and a hypothesized person, P, the task of verification is to determine if O was from P. This verification task can be restated as a basic hypothesis test between

$$H_0 : O \text{ is from person } P$$
$$H_1 : O \text{ is } \underline{\text{not}} \text{ from person } P$$

Using statistical pattern recognition techniques, the optimum test[1] to decide between these two hypotheses is a likelihood ratio test given by

$$\frac{p(O \mid H_0)}{p(O \mid H_1)} \begin{cases} \geq \theta \text{ Accept } H_0 \\ < \theta \text{ Reject } H_0 \end{cases}, \tag{1}$$

*This work was sponsored by the Department of Defense under Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the United States Government.

[1] Strictly speaking, the likelihood ratio test is only optimal when the likelihood functions are known exactly. In practice this is rarely the case.

where $p(O \mid H_i)$, $i = 0, 1$ is the probability density function for the hypothesis H_i evaluated for the measurement Y, also referred to as the "likelihood" of the hypothesis H_i given the measurement[2]. The decision threshold for accepting or rejecting H_0 is θ. The basic aim in developing a verification system is to determine techniques to compute this likelihood ratio function, usually by finding method to represent and model the two likelihoods, $p(O \mid H_0)$ and $p(O \mid H_1)$.

The first step in a verification system is to extract from the observation features that convey person-dependent information, such as vocal-tract related spectral measurement when the observations are speech samples in a speaker verification system. The output of this stage is typically a sequence of feature vectors representing the observation, $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$. These feature vectors are then used to compute the likelihoods of H_0 and H_1.

In statistical pattern recognition based verification systems, H_0 is represented by a model denoted λ_P, that characterizes the distribution of features derived from observations from the person P in the feature space of $\mathbf{x}$. For example, one could assume that a Gaussian mixture model (GMM) distribution best represents the distribution of feature vectors for H_0 so that λ_P would be denoting the weights, means can covariance matrix parameters of a GMM. The alternative hypothesis (see entry on GMM for more details of this model). H_1, is likewise represented by a model $\lambda_{\overline{P}}$. The likelihood ratio statistic is then formed as

$$\mathrm{LR}(X) = \frac{p(X \mid \lambda_P)}{p(X \mid \lambda_{\overline{P}})}. \tag{2}$$

Alternative Hypothesis Modeling

While the model for H_0 is well defined and can be estimated using training samples from P, the model for $\lambda_{\overline{P}}$ is less well defined since it potentially must represent the entire space of possible alternatives to the person P.

From the area of speaker recognition, two main approaches have been taken for this alternative hypothesis modeling. Here we will refer to "speakers" and speech samples, but this equally applies to other biometric measurements and features. The first approach is to use a set of other speaker models to cover the space of the alternative hypothesis. In various contexts, this set of other speakers has been called likelihood ratio sets [1], cohorts [2] and background speakers [3]. Given a set of N background speaker models $\{\lambda_1, \ldots, \lambda_N\}$, the alternative hypothesis model is represented by

$$p(X \mid \lambda_{\overline{P}}) = \mathcal{F}(p(X \mid \lambda_1), \ldots, p(X \mid \lambda_N)), \tag{3}$$

where $\mathcal{F}()$ is some function, such as average or maximum, of the likelihood values from the background speaker set. The selection, size, and combination of the background speakers has been the subject of much research (for example [2, 4, 3]). In general, it has been found that to obtain the best performance with this approach requires the use of speaker-specific background speaker sets. This can be a drawback in an applications using a large number of hypothesized speakers, each requiring their own background speaker set.

The second major approach to alternative hypothesis modeling is to pool speech from several speakers and train a single model. Various terms for this single model are a general model [5], a world model and a universal background model [6]. Given a collection of speech samples from a large number of speakers representative of the population of speakers expected during recognition, a single model, λ_{bkg}, is trained to represent the alternative hypothesis. Research on this approach has focused on selection and composition of the speakers and speech used to train the single model [7, 8]. The main advantage of this approach is that a single speaker-independent model can be trained once for a particular task and then used for all hypothesized speakers in that task. It is also possible to use multiple background models tailored to specific sets of speakers [8, 9].

Universal Background Models

Most modern speaker verification system use a UBM for modeling the alternative hypothesis in the likelihood ratio test. Typically, GMMs are used for distribution models and a speaker-specific model are derived by using MAP estimation with the UBM acting as the prior model (see entry on GMMs for details on MAP estimation). In the GMM-UBM system we use a single, speaker-independent background model to represent $p(X \mid \lambda_{\overline{P}})$. The UBM is a large GMM (2048 mixtures) trained to represent the speaker-independent distribution of features. Specifically, we want to select speech that is reflective of the expected alternative speech to be encountered during recognition. This applies to both the type and quality of speech, as well as the composition of speakers. For example, for a verification system using telephone speech and only male speakers, the

[2] $p(A \mid B)$ is referred to as a likelihood when B is considered the independent variable in the function.

UBM would be trained using telephone speech from pool of male speakers. In the case where such specific prior knowledge of the gender composition of the alternative speakers not known, we would train using speech from both male and female speakers. Other than these general guidelines and experimentation, there is no objective measure to determine the right number of speakers or amount of speech to use in training a UBM.

Given the data to train a UBM, there are many approaches that can be used to obtain the final model. The simplest is to merely pool all the data and use it to train the UBM via the EM algorithm. One should be careful that the pooled data is balanced over the subpopulations within the data. For example, in using gender-independent data, one should be sure there is a balance of male and female speech. Otherwise, the final model will be biased towards the dominant subpopulation. The same argument can be made for other subpopulations such as speech from different microphones. Another approach is to train individual UBMs over the subpopulations in the data, such as one for male and one for female speech, and then pool the subpopulation models together. This approach has the advantages that one can effectively use unbalanced data and can carefully control the composition of the final UBM. Still other approaches can be found in the literature (see for example [8, 10]).

The concept of a UBM is also used for discriminative systems, such as Support Vector Machines (SVM), where explicit likelihood functions for the two hypothesis are not used. In this case, the UBM refers to the collection data from the general population used as negative examples when training a person specific discriminate function [11].

Related Entries

Speaker Recognition, Speaker Modeling, Speaker Matching, Gaussian Mixture Models

References

1. Higgins, A., Bahler, L., Porter, J.: Speaker verification using randomized phrase prompting. Digital Signal Processing **1** (1991) 89–106
2. Rosenberg, A.E., DeLong, J., Lee, C.H., Juang, B.H., Soong, F.K.: The use of cohort normalized scores for speaker verification. In: International Conference on Speech and Language Processing. (1992) 599–602
3. Reynolds, D.A.: Speaker identification and verification using Gaussian mixture speaker models. Speech Communication **17** (1995) 91–108
4. Matsui, T., Furui, S.: Similarity normalization methods for speaker verification based on a posteriori probability. In: Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification. (1994) 59–62
5. Carey, M., Parris, E., Bridle, J.: A speaker verification system using alphanets. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. (1991) 397–400
6. Reynolds, D.A.: Comparison of background normalization methods for text-independent speaker verification. In: Proceedings of the European Conference on Speech Communication and Technology. (1997) 963–967
7. Matsui, T., Furui, S.: Likelihood normalization for speaker verification using a phoneme- and speaker-independent model. Speech Communication **17** (1995) 109–116
8. Rosenberg, A.E., Parthasarathy, S.: Speaker background models for connected digit password speaker verification. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. (1996) 81–84
9. Heck, L.P., Weintraub, M.: Handset-dependent background models for robust text-independent speaker recognition. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. (1997) 1071–1073
10. Isobe, T., Takahashi, J.: Text-independent speaker verification using virtual speaker based cohort normalization. In: Proceedings of the European Conference on Speech Communication and Technology. (1999) 987–990
11. Campbell, W.M.: Generalized linear discriminant sequence kernels for speaker recognition. In: Proceedings of the International Conference on Acoustics, Speech, and Signal Processing. (2002) 161–164