2009 Seventh International Workshop on Content-Based Multimedia Indexing

Combining Cohort and UBM Models in Open Set Speaker Identification

Anthony Brew and Pádraig Cunningham
Machine Learning Group
School of Computer Science and Infomatics
University College Dublin
anthony.brew@ucd.ie, padraig.cunningham@ucd.ie

Abstract

In open set speaker identification it is important to build an alternative model against which to compare scores from the 'target' speaker model. Two alternative strategies for building an alternative model are to build a single global model by sampling from a pool of training data, the Universal Background (UBM), or to build a cohort of models from selected individuals in the training data for the target speaker. The main contribution in this paper is to show that these approaches can be unified by using a Support Vector Machine (SVM) to learn a decision rule in the score space *made up of the output scores of the client, cohort and UBM model.*

1 Introduction

In machine learning terms, speaker identification is a binary classification task where the challenge is to determine whether or not utterances come from a target speaker. *Open set* speaker identification is a particularly challenging variant of this problem where the system is required to distinguish the target speaker from speakers for whom there was no training data.

The iconic version of the speaker identification (speaker verification) problem comes from the domain of biometrics where the objective is to identify individuals for security reasons. However, the problem of speaker identification is at least as important in multimedia annotation [14, 3]. It is important for indexing news and film archives [22], for indexing telephone conversations [12] and in pervasive computing scenarios [18].

In this paper we present a machine learning framework for speaker identification where the system is required to handle non-class examples for which there is no training data. This arises for instance when a news archive is to be annotated with speaker labels. This situation is depicted in Figure 1 where the target examples are shown as + and the open set nature of the problem is depicted by non-target examples X,Y,Z. Annotated data for these non-target speakers is not available at training time, instead data on speakers A,B,C,D,E,F is available from which a non-target model can be constructed. All of this non target data can be sampled to produce a single non-target model or a subset of these non-target speakers can be selected (e.g. A,B and E) as good representatives on which to build a cohort of non-target models (see section 2). The main contribution in this paper is a machine learning framework that integrates both of these strategies. This integration strategy based on the idea of score spaces is described in section 4.

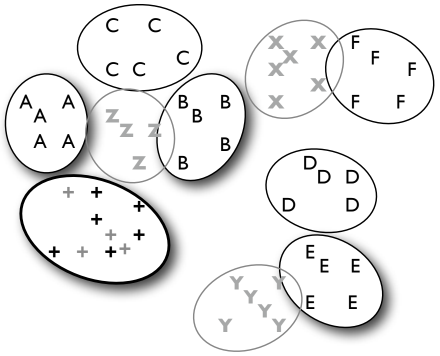

Figure 1. In open set speaker identification the system is required to distinguish the target speaker (+) from speakers for which there is no data available at training time – examples X,Y and Z in this figure.

In the next section the alternative strategies for building Gaussian mixture models for target and non-target data are explored. In section 3 we describe how support vector machines (SVM) can be used to bring the power of a discriminative classifier to this classification task and in section 4 we describe how we develop on the score space idea to improve the classification accuracy. The evaluation methodology is

978-0-7695-3662-0/09 $25.00 © 2009 IEEE
DOI 10.1109/CBMI.2009.30

explained in section 5 and the results are discussed in section 6.

2 Gaussian Mixture Models

In most speaker verification and identification systems, some form of spectral based parameterisation is used to encode the speech in machine readable form. Typically short-term analysis (about 20 ms) is used to compute a sequence of magnitude spectra or ‘frames’. Most commonly, the frames obtained are then converted into cepstral coefficients and the frequency scale warped into the Mel scale [3].

In Speaker Recognition a popular technique is to build a Gaussian Mixture Model (GMM) for each speaker in the training set. Each individual model is trained on many frames x_i of speech from a set of training utterances from each individual speaker separately [15]. A score that gives an indication of whether a given utterance $X = x_1, x_2, ...x_n$ originates with speaker ‘a’ is found using the mean log-likelyhood $\frac{1}{n}\log(P(\mathbf{X}|\theta_{\mathbf{a}}))$ that the utterance came from a’s model $\theta_{\mathbf{a}}$:

$$P(\mathbf{X}|\theta_{\mathbf{a}}) = \prod_{x_j \in \mathbf{X}} \sum_{i}^{N} \frac{\alpha_{a,i}}{(2\pi)^{d/2}|\Sigma_{a,i}|} e^{\left(\frac{1}{2}(\mathbf{x_j}-\mu_{\mathbf{a,i}})^T \Sigma_{a,i}^{-1}(\mathbf{x_j}-\mu_{\mathbf{a,i}})\right)} \tag{1}$$

where N is the number of models in the mixture, $\theta_{\mathbf{a}} = \{\mu_{\mathbf{a}}, \mathbf{\Sigma_a}, \alpha_{\mathbf{a}}\}$. However it has been shown that performance can be improved by normalizing the raw speaker model log-likelihood scores by using the log-likelihood score from a background speaker model to form a log-likelihood ratio [15, 7, 16, 1].

Two methods for building a background speaker model are the Universal Background Model (UBM) and cohort sets. The UBM [15] is a single model built on a representative sample of speech from a large pool of speakers, whereas the cohort approach uses a composition of C speakers models θ_{c_i} around the client speaker as an alternative model [16]. The decision function for a UBM is given by (2) and for the cohort approach is given by (3).

$$\frac{1}{n}\log(P(\mathbf{X}|\theta)) - \frac{1}{n}\log(P(\mathbf{X}|\theta_{\mathbf{UBM}})) \quad > \quad t_\theta \tag{2}$$

$$\frac{1}{n}\log(P(\mathbf{X}|\theta)) - f(\{\frac{1}{n}\log(P(\mathbf{X}|\theta_{\mathbf{c_i}})\}) \quad > \quad t_\theta \tag{3}$$

$$0 \leq i \leq C$$

The function $f(\cdot)$ refers to how a cohort score may be selected from the cohort set. This function for example can be the maximum or average score of the component models [16]. As each utterance X is of variable length n, the fraction $\frac{1}{n}$ is used to normalize the log-likelihood for varying utterance durations [15].

3 Support Vector Machines in Speaker Recognition

As utterances are of variable length and there is an abundance of frame based information, generative models such as the GMM have become very popular in Speaker Recognition. However the problem is not to generate a confidence in whether an utterance is from a given speaker or not but to simply make a binary decision. Hence a better solution in theory would be to use a discriminant classification framework [20]. Over the past number of years SVMs have become a popular discriminative classifier in the machine learning community. Initial work by by Schmidt and Gish [17] highlighted the drawback of using SVMs on the frame based features used by generative models, namely that SVMs become inefficient when the number of training frames is large.

For UBM-based speaker identification, the Bayes decision rule (2) will be optimal as long as the client and impostors are well modeled. Bengio and Mariethoz [2] suggested that the probability estimates are not perfect and that it would be better to look at the projection of the training data into the score space using the mapping (4) and then to use an SVM in this new projection space to make classification decisions.

$$\phi_{UBM}(X) = \begin{pmatrix} \frac{1}{n}log(P(\mathbf{X}|\theta)) \\ \frac{1}{n}log(P(\mathbf{X}|\theta_{\mathbf{UBM}}) \end{pmatrix} \tag{4}$$

In its basic form, an SVM is a binary linear classifier. Given a binary set of linearly separable training data, there are many possible linear discriminative classifiers. SVMs are described in detail by Vapnik [20] and in Burges’ tutorial [4].

Let the separating hyperplane be defined by $\mathbf{x}\cdot\mathbf{w}+b = 0$, where $\mathbf{w}$ is its normal vector. For a binary labeled linearly separable dataset $\{\mathbf{x}_i, y_i\}, y_i \in \{-1, +1\}, i = 1, \ldots, N$ the optimal boundary chosen by an SVM is the one that maximises the margin between the two classes. This is done by minimising $\|\mathbf{w}\|^2$ subject to $(\mathbf{x}_i \cdot \mathbf{w} + b)y_i \geq 1$ for all $i = 1, \ldots, N$.

When the data is not linearly separable, the above inequalities cannot be satisfied. By introducing ‘slack’ variables ξ_i which represent the amount each point in the dataset is misclassified, the objective function to be minimised can be reformulated as:

$$\frac{1}{2}\|\mathbf{w}\|^2 \quad + \quad C\sum_i \xi_i \tag{5}$$

$$\text{subject to } (\mathbf{x}_i \cdot \mathbf{w} + b)y_i \quad \geq \quad 1 - \xi_i, \forall i = 1, \ldots, N$$

The second term on the right-hand side of (5) is the empirical risk associated with points that are misclassified or lie within the margin. C is a hyper-parameter that trades off the effects of minimizing the empirical risk against maximizing the margin.

The Lagrangian dual formulation of (5) (which can be solved using standard convex optimisation techniques) is

$$\alpha = \max_{\alpha} \left(\sum_{i} \alpha_i + \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \right) \tag{6}$$

$$\text{subject to } 0 \leq \alpha_i \leq C \text{ and } \sum_{i} \alpha_i y_i = 0$$

Other approaches using SVMs in speaker identification have also used the generative models to project data into a new space for and SVM to learn a discriminative boundary. For the most part this work has focused on extending the more popular UBM based approach[9, 21, 8, 5, 11]. An extensive literature search suggests that the cohort based strategy has been largely ignored in this work that incorporates SVMs as a discriminative classifier – thus this is our focus in this paper. It would appear logical that scores found from speakers that sound 'like' the target speaker should be helpful in building a discriminative classifier.

In this paper we extend the idea of using the score space generated from the UBM and target model[2] to include the scores from individuals in a cohort set. Each speaker c_i in the set provides a model that induces a new dimension in the cohort score space when an utterance X is passed through the corresponding model in the transform yielding a vector in $\mathbb{R}_{C+1}$. We investigate the use of these spaces alone and by appending the scores to the score space obtained using a UBM.

$$\phi_c(X) = \begin{pmatrix} \frac{1}{n} log(P(\mathbf{X}|\theta_{\mathbf{c_0}})) \\ \frac{1}{n} log(P(\mathbf{X}|\theta_{\mathbf{c_1}})) \\ \vdots \\ \frac{1}{n} log(P(\mathbf{X}|\theta_{\mathbf{c_C}})) \end{pmatrix} = \begin{pmatrix} s_0 \\ s_1 \\ \vdots \\ s_C \end{pmatrix} \tag{7}$$

This paper proceeds by examining some score spaces derived from UBM and cohort models built on the YOHO dataset [6] and demonstrates how the use of the score space can greatly improve the performance of speaker verification using cohort sets.

4 Speaker Identification Using Cohort Model Score Spaces

For a given 'target' speaker and a cohort set of C background speakers, speaker identification using a cohort model (3) may be written in terms of the cohort score transform (7) as:

$$s_0 - f(s_1, \cdots, s_C) > t \tag{8}$$

This classification rule is the Bayes decision boundary between the target model θ and an alternative model given by the statistic $f(\cdot)$ and the cohort speakers models $\theta_1, \ldots, \theta_C$. Here t is a threshold that controls the tradeoff between the false positive and false negative rate.

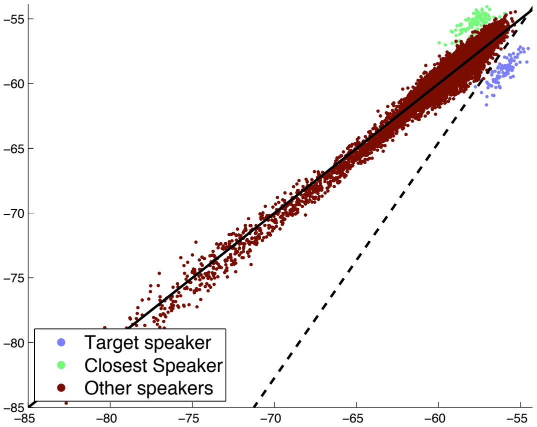

Figure 2. Data projected into 2D score space using two models. The x-axis is the log-likelihood score obtained from the target speaker and the y-axis is the log-likelihood score of the cohort model.

This boundary can be visualised by mapping the data into a 2D space using

$$\phi_f(X) = \begin{pmatrix} s_0 \\ f(s_1, \ldots, s_C) \end{pmatrix} \tag{9}$$

A simple example of this is shown in Figure 2 where a cohort set of size one was used. When the data that is not used in the cohort model or the target model is projected into the score space it resides around the Bayes decision boundary with some elements lying 'closer' to the target speaker and some elements lying closer to the cohort.

In addition to the nature of $f(\cdot)$ the composition of the cohort set has also been a topic of research [15, 7]. By using a distance measure between models proposed by Reynolds a selection of models can be identified to make up the the cohort set. It has been shown that while it is important to select models that are close it is also useful to have diversity in the cohort [13]. In this work we use cohort score spaces created by using C close but maximally spread (CMS) speakers. This is where, for the target speaker, a set of N closest models are found. From this set C $(C < N)$ models that are maximally spread are then used as the cohort, full details can be found in [15].

When the world model is a UBM it has been shown that the Bayes decision surface is not the optimal boundary and a better boundary can be found using a support vector machine (SVM) [2]. This work was extended to show that this boundary can be further improved by building an individual boundary for each speaker against the UBM [10].

It is clear in Figure 2 that the Bayes boundary is not the optimal boundary for verifying the identity of the target speaker when the cohort size is one. As the size of the cohort increases, the data will start to shift over the Bayes boundary towards the cohort model. We show that better rules in the score space defined by the mapping (7) can be found using an SVM to build a maximal margin linear classifier.

An important quality of (6) is that SVMs are trained and classify new points using only the inner product between points in the training set. Thus if we can define a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ between elements in the data set that satisfies Mercer's condition, we can build a maximal margin classifier using an SVM. Mercer's condition for a valid Kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$ requires that there exists a mapping $\phi(\mathbf{x}) : X \Rightarrow \mathbb{H}$ where $\mathbb{H}$ is a Hilbert Space: a vector space closed under dot products, ie

$$K(\mathbf{x}_i, \mathbf{x}_j) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle$$

The cohort model score transforms $\phi_f(\cdot), \phi_c(\cdot)$ perform such a mapping from the space of variable length utterances to $\mathbb{R}_n$. Thus we can define cohort model kernels between variable length utterances $\mathbf{X}, \mathbf{Y}$ as

$$\begin{aligned} K_f(\mathbf{X}, \mathbf{Y}) &= \langle \phi_f(\mathbf{X}), \phi_f(\mathbf{Y}) \rangle \\ K_c(\mathbf{X}, \mathbf{Y}) &= \langle \phi_c(\mathbf{X}), \phi_c(\mathbf{Y}) \rangle \end{aligned}$$

In the case of $\phi_c(\cdot)$, (7), learning the hyper-plane that separates the target speaker from the rest of the populace we find a hyper-plane

$$\mathbf{w} \cdot \phi_c(X) - b = 0$$

where $\mathbf{w}$ is the normal vector to the plane. Speaker verification is calculated using

$$\mathbf{w} \cdot \phi_c(X) > t$$

where t is a threshold set to give an appropriate trade-off between false positives and false negatives. The individual entries w_i that make up the vector $\mathbf{w}$ can be considered weights. This is equivalent to learning a weighted average statistic for classification.

By extending the mapping (4) used by Bengio and Mariethoz [2] by appending the scores of the cohort we effectively combine the UBM with the cohort based approach.

5 Experimental Setup

In this work, conventional Mel Frequency Cepstral Coefficients (MFCCs) [3] are used for speech parametrization. 20 MFCCs are used, extracted using a Hamming window of 20ms. The zeroth cepstral coefficients (the DC level of the log-spectral energies) are not used in the feature vector. Cepstral mean subtraction was also performed on the MFCCs at an utterance level.

Naive silence removal was performed from each utterance by training a bimodal Gaussian mixture model on the energy of each frame and then discarding frames that belonged to the lower of the two Gaussian components [3].

The individual models for the score transform and the UBM were computed using 64 and 1024 diagonal covariance mixture models respectively. The models were trained using expectation maximisation using 10 steps. This training procedure in turn was run 10 times and the model which achieved the highest expectation value on the training set was retained.

For each speaker in the cohort experiments described in Section 4 a set of 30 randomly chosen outliers was held out and cohort selection was performed on the remaining 107 speakers. The cohort speakers were selected using the distance metric described in [15].

As with the cohort experiments, in the UBM experiments a set of 30 speakers for each speaker (the same 30 as held out for the cohort experiments) was held out for testing and the UBM was trained on the remaining 107 speakers. To enable reasonable computation times the hold out set was 'rotated' so that the same UBM could be used for 30 target speakers. This rotational policy was also used for the cohorts so that the comparison was fair.

Experiments were performed using cohort sets of size 10. The selection strategies used was close but maximally spread (CMS), as outlined in section 4. The CMS cohort was based on finding the 20 closest speakers and then selecting 10 speakers from this set that when included with the target speaker model where maximally spread from one another. The statistic used to combine scores found in the cohort was the mean as this is considered to be 'smooth' statistic [19].

An SVM was trained for each subproblem so that a separate decision boundary for each speaker, cohort selection method, cohort size and statistic $f(\cdot)$ was found. The C parameter of the SVM was calculated using cross validation to find a value that worked well for all speakers, on each subproblem.

Errors are reported in Table 1 using the average Equal Error Rate (EER) found over the full 138 speakers. The EER is the operating point that balances false positive and false negative errors, a popular measure for assessing the performance of classifiers in this domain [3].

6 Results

Results in Table 1 show that a better linear boundary can be found when the scores from the generative models are used as a mapping into a new space for training a classifier rather than directly for classification as in their traditional setting.

The first row of Table 1 is the base line result using a cohort based model with a cohort size of 10. The results shown on the second line of the table shows that when the mapping $\phi_\mu(\cdot)$ is used a better decision boundary can be found. This mapping maps the utterances to a two dimensional space by taking the target score for the first dimension and the average of the scores from the cohort as the second dimension. This allows a better decision surface to be learnt but does not allow each member of the cohort to 'speak for themselves' in the training of the decision surface. This result reinforces the result found by Bengio and Mariethoz [2] for a UBM based model – in this case the non-target model is a cohort statistic rather than a UBM. The improvement in accuracy is more pronounced here as the average of the cohort scores does not provide a good statistical model of alternative speech.

The next major improvement can be seen when the data is projected directly into the cohort score space using $\phi_c(\cdot)$ (7), this result is shown on the third line of table 1. Each member of the cohort in this method generates a dimension of its own. The hyper-plane found in this space is a weighted combination of the cohort models scores. In effect we are allowing the SVM to learn the best weights for each member of the cohort based on the projection of the training data into the score space. This allows the learner to estimate the usefulness of each member of the cohort in the final classification. It is interesting to note that because this makes better use of the information from the cohort models it can match its UBM counterpart in accuracy.

We found that when we attempted to use Benjio and Mariethoz's UBM projection (line 5 Table 1) [2] we could only match the accuracy found by the traditional UBM (line 4 Table 1). The inability to improve the UBM result here may be due to differences in the setup of our experiment but also due to the high quality of the recordings in the YOHO dataset leading to very high quality models being built for the target speaker and the global model.

Our key result is shown on line 6 of Table 1. This is where the data is projected using the models that make up the cohort, the UBM and the model built on the target speaker. Similar to the result found for the individual cohort projection, it allows each model in the transform to 'speak for itself' whilst the classification boundary is being learnt from the data. While projections only using the cohort or the UBM have both been shown to improve accuracy, what is important here is that in combination they can go further improving the EER from 0.7% to 0.4%. This would indicate that the information encoded in the score space made up by the members of the cohort and score space of the UBM hold different information about the identity of the speaker and that in combination the benefits of both can be realised.

Decision	**EER**
Cohort (Average)	1.7%
SVM (ϕ_μ) Cohort	1.0%
SVM (ϕ_c) Cohort	0.7%
Universal Base Model (UBM)	0.7%
SVM UBM	0.7%
SVM UBM + Cohorts	0.4%

Table 1. Equal Error Rates (EER) found for speaker verification experiments using cohorts and the UBM as normal and as projections for an SVM to learn

7 Conclusions and Further Work

This paper investigated the use of a background model provided by a cohort of speakers to project the training data into a new space to make better classification predictions.

We have shown that using a background model provided by an average of the cohort scores can be improved by allowing each member of the cohort to 'speak for themselves' when the decision boundary is being learnt. This allows the SVM to learn better weights for combining the scores from each of the models that make up the target speaker's cohort. This more comprehensive use of the information coming from the cohort models brings the accuracy of the cohort based approach in line with the UBM alternative.

The most important contribution of this paper is that when the UBM based approach is combined with the cohort model in a projective framework accuracies are further improved. This shows that both strategies hold separate information about the identity of a speaker, and when combined the benefits of both can be realised.

In its simplest form this work provides new features for classifiers to use in speaker identification. The use of a linear kernel in the generated space was driven by simplicity of parameter selection and should not suggest that we believe a linear relation between cohort scores is optimal. We believe there is merit is investigating other kernels and we are looking at combining this technique with other generative SVM methods to find further improvements.

References

[1] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score Normalization for Text-Independent Speaker Verification Systems. *Digital Signal Processing*, 10(1-3):42–54, 2000.

[2] S. Bengio, J. Mariethoz, and M. IDIAP. Learning the decision function for speaker verification. *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP'01). 2001 IEEE International Conference on*, 1, 2001.

[3] F. Bimbot, J. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, Petrovska-Delacretaz, and D. D., Reynolds. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing*, 4:430–451, 2004.

[4] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining and Knowledge Discovery*, 2(2):121–167, 1998.

[5] W. Campbell, D. Reynolds, and J. Campbell. Fusing Discriminative and Generative Methods for Speaker Recognition: Experiments on Switchboard and NFI/TNO Field Data. *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[6] J. Campbell Jr. Testing with the yoho cd-rom voice verification corpus. *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, 1, 1995.

[7] A. Higgins, L. Bahler, and J. Porter. Speaker verification using randomized phrase prompting. *Digital Signal Processing*, 1(2):89–106, 1991.

[8] P. Ho and N. Vasconcelos. A Kullback-Leibler Divergence Based Kernel for SVM Classification in Multimedia Applications. In *Proc. Advances in Neural Information Processing Systems*, volume 16, pages 1385–1392, 2004.

[9] J. Kharroubi, D. Petrovska-Delacretaz, and G. Chollet. Combining GMM's with Suport Vector Machines for Text-independent Speaker Verification. In *Seventh European Conference on Speech Communication and Technology*. ISCA, 2001.

[10] Q. Le and S. Bengio. Client Dependent GMM-SVM Models for Speaker Verification. In *Artificial Neural Networks and Neural Information Processing-ICANN/ICONIP 2003*. Springer, 2003.

[11] J. Louradour, K. Daoudi, and F. Bach. SVM Speaker Verification using an Incomplete Cholesky Decomposition Sequence Kernel. In *Proc. Odyssey: The Speaker and Language Recognition Workshop*, 2006.

[12] I. Magrin-Chagnolleau, F. Bimbot, and R. IRISA. Indexing telephone conversations by speakers using time-frequencyprincipal component analysis. In *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, volume 2, 2000.

[13] D. Reynolds. Comparison of background normalization methods for text-independent speaker verification. *Fifth European Conference on Speech Communication and Technology*, 1997.

[14] D. Reynolds. An overview of automatic speaker recognition technology. In *Proc. Int. Conference on Acoustics, Speech, and Signal Processing*, 2002.

[15] D. A. Reynolds. Speaker identification and verification using gaussian mixture speaker models. *Speech Communication*, 17(1):91–108, 1995.

[16] A. Rosenberg, J. DeLong, C. Lee, B. Juang, and F. Soong. The Use of Cohort Normalized Scores for Speaker Verification. *Second International Conference on Spoken Language Processing*, 1992.

[17] M. Schmidt and H. Gish. Speaker identification via support vector classifiers. *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1, 1996.

[18] V. Stanford, J. Garofolo, O. Galibert, M. Michel, and C. Laprun. The NIST Smart Space and Meeting Room projects: signals, acquisition annotation, and metrics. In *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on*, volume 4, 2003.

[19] D. Tax, M. van Breukelen, R. Duin, and J. Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485, 2000.

[20] V. Vapnik. *Statistical learning theory*. Wiley New York, 1998.

[21] V. Wan. *Speaker Verification using Support Vector Machines*. PhD thesis, University of Sheffield, 2003.

[22] X. Zhu, C. Barras, L. Lamel, and J. Gauvain. Speaker Diarization: From Broadcast News to Lectures. *LECTURE NOTES IN COMPUTER SCIENCE*, 4299:396, 2006.