Anchor Models and WCCN Normalization For Speaker Trait Classification

Yazid Attabi 1,2, Pierre Dumouchel 1,2
1 École de technologie supérieure, Montréal, Canada
2 Centre de recherche informatique de Montréal, Montréal, Canada
yazid.attabi@crim.ca, pierre.dumouchel@crim.ca

Abstract
This paper presents an improved version of anchor model applied to solve the two-class classification tasks of the INTERSPEECH 2012 speaker trait Challenge. To build the anchor model space of each task, we include the class models of all tasks. The introduction of within-class covariance normalization (WCCN) applied to the log-likelihood scores of the anchor space not only improves the results compared to the unnormalized version but also exceeds the performance of GMM or GMM-UBM systems. Even if Euclidean distance gives worst performances compared to cosine metric, we find that after normalization both metrics give similar results so they can be used interchangeably.

Index Terms: anchor model, WCCN, speaker trait classification, GMM model, Interspeech 2012 challenge

1. Introduction
Anchor model is a method which was first introduced for speaker indexing in large audio databases in [1] and then was extended for speaker identification [2] and speaker verification [3] problems. In this method, a speaker is relatively characterized with respect to a set of reference speaker models called anchor models. The relative position obtained by the projection of a speaker in the anchor space, which is formed by the set of anchor speakers, is named Speaker Character Vectors (SCV) [3]. Different metrics are used to compute the similarity between SCV such as Euclidean [1], angular [1] [2] or correlation [4] metrics. Although these systems are newer, their performance based on the equal error rates criterion can not compete with those of GMM-UBM method. In [3], a new qualitative measurement based on the rank metric is introduced. This new metric improves the performance compared to the quantitative distances but remains below the performance of GMM-UBM method.

In this paper we present an improved anchor models method by introducing within-class covariance normalization (WCCN) on the log-likelihood scores in the anchor space. This method was successfully applied to solve a multi-class problem of emotion recognition from children's speech [5]. We experiment this model to solve two-class classification problems of the INTERSPEECH 2012 speaker trait Sub-Challenges namely, Personality, Likability and Pathology [6]. We show that WCCN normalization applied for Euclidean or angular metrics improves significantly the performance particularly for the first one.

In the next section we describe the feature vectors used in this study and the gaussianization method, which serves as canal compensation technique. Section 3 describes the GMM method used as both, a standalone baseline system and also as front-end system for the anchor models described in section 4. After a brief introduction of WCCN technique in section 5, experimental setup and results are presented in section 6 which is followed by the conclusion.

2. Feature extraction
In this section we describe the cepstral feature extraction as well as the gaussianization processing used as feature normalization technique.

2.1. Cepstral Features
The Mel-Frequency Cepstral Coefficients (MFCC) are used as features to model the varying nature of speech with respect to the type of speaker trait. The MFCC have been successfully applied for speech, emotion and speaker recognition. The MFCC vector is composed of the first 12 coefficients including C0 (as the energy component) calculated at a rate of 10 ms using a 25 ms Hamming window. First and second derivatives are computed using a 5-frame window for each MFCC vector in order to compute the temporal characteristics. Two sets of MFCC vectors are experimented depending on whether silences are kept or removed from the audio files before the MFCC extraction.

2.2. Gaussianization
A gaussianization is a two-step parameter processing technique aimed to reduce disparities introduced by the variability in the signal acquisition conditions. As first step, a global linear transformation is applied in order to obtain the independence of the different parameters of the multi-dimension acoustic vector. As second step, a short-time Gaussian warping transformation is applied separately on each dimension in order to reduce the mismatch between the actual statistical distribution of the acoustic features and the Gaussian distribution assumed in the acoustic models [7].

3. GMM model
The cepstral feature vectors are modeled using a Gaussian Mixture model (GMM). GMM is a generative model widely used in the field of speech processing. It is a semi-parametric probabilistic method that offers the advantage of adequately representing speech signal variability. Given a GMM modeling a $d$-dimensional vector, the probability of observing a feature vector given the model is computed by the following equation:
\[
P(x|\lambda) = \sum_{i=1}^{n} w_i N(x; \mu_i, \Sigma_i)
\]

where \( m, w_i, \mu_i \) and \( \Sigma_i \) correspond to the number of Gaussians, weight, mean vector and covariance matrix of the \( i \)-th Gaussian, respectively. In all our GMMs, we used diagonal covariance matrices instead of full covariance matrices. GMM parameters are estimated using two methods. The first one uses Maximum Likelihood (ML) approach based on the Expectation Maximization (EM) algorithm [8]. The latter uses the maximum a posteriori (MAP) in order to overcome the problem of sparse data. For each task, the speaker trait models are generated from a well-trained initial GMM model, called Universal Background Model (UBM), and from a limited amount of training data via the MAP adaptation [9]. For each of the three corpora, an UBM model is trained using all the data of the training partition via the EM algorithm.

The classification of a test frame sequence \( X = \{x_1, x_2, \ldots, x_n\} \) is based on the Bayes decision. Using an equal prior for all classes, this classification is obtained by computing directly the log-likelihood of the test recording given each GMM associated with each task classes. This test recording is classified as speaker trait of class \( C \) if the log-likelihood value of equation (2) is greater than a certain threshold \( \varepsilon \): 

\[
\log \frac{P(X|\lambda)}{P(X|\lambda')} \geq \varepsilon ,
\]

where \( \log P(X|\lambda) \) is the log-likelihood of the test sequence \( X \) given a GMM \( \lambda \).

4. Anchor models
An anchor model system allows representing each speaker trait relatively to a set of speaker trait models. The reference set of speaker trait models is called anchor models. In a multi-class problem, as in [5], we can define the anchor models space by selecting all or a sub-set class models of the problem. On the other side, for a two-class problem as is the case for the speaker trait challenge, extra class models are needed to define the anchor space. Given that not only class models of a given task can serve as a set of reference models, we can include class models of other tasks and of other corpora to build the anchor space of the two-class task. Thus, for a given task, we define a \( C \)-dimensional anchor space generated by all class models of the seven tasks, namely \( C = 14 \) models \( \# \text{tasks} \times \# \text{classes per task} = 7 \times 2 \). The speaker trait models of the anchor space are defined using GMMs. We will name the new representation of a speaker trait obtained by the projection of an utterance from the acoustic parameters space into the anchor model space as Speaker Trait Vector (STV) by analogy to the terminology used in the speaker diarization and verification. The STV vector of an utterance is calculated by computing its log-likelihood score for each speaker trait model as in (3).

5. WCCN Normalization
The Within-Class Covariance Normalization (WCCN) is a technique introduced in [10] to train a generalized linear kernel of an SVM-based (Support Vector Machine) system in order to minimize the expectation of false positives and false negatives errors. The generalized linear kernel \( k(L_1, L_2) \) is expressed as:

\[
k(L_1, L_2) = L_1^T R L_2 ,
\]

where \( L_1 \) and \( L_2 \) are two given instances and \( R \) is a positive semi definite matrix. The closed-form solution is reached by setting \( R = W^{-1} \), where \( W \) is the expected within-class covariance matrix of the data defined as:

\[
W = \sum_{i=1}^{n} p(i) \cdot S_i ,
\]

where \( p(i) \) and \( S_i \) represent the prior probability and the within covariance matrix of class \( i \) respectively.

WCCN has also been successfully applied on the i-vector feature space [11]. An i-vector is a low dimensional representation of a high dimensional supervector, which is in turn obtained by the concatenation of all GMM mean vectors.

In this paper, we apply WCCN on log-likelihood scores of the anchor vector. In addition to the WCCN normalization, we also experiment the normalization using the full and diagonal overall covariance matrix over all of the data.

6. Experiments
For each task, we use the training partition task to train the GMM and UBM models as well as the WCCN matrices. The development set was used for parameters tuning such as the
number of Gaussian components of the GMM and for the choice of the parameters estimation training method.

### 6.1. Corpus

Each of the three Sub-Challenges is associated with its own corpus which we will describe briefly.

The Speaker Personality Corpus (SPC) is composed of clips randomly extracted from the French news bulletins aimed to serve for analysis and comparison. There are 322 clips of 10 seconds each. Each clip, associated to only one person, is annotated with a score for each of the OCEAN Big-Five dimensions mapped after that into two classes depending on whether the associated score is above or below the average: Openness (O, NO), Conscientiousness (C, NC), Extraversion (E, NE), Agreeableness (A, NA), Neuroticism (N, NN).

The speaker likability study is conducted on “Speaker Likability Database” (SLD) corpus a subset of the German Agender database recorded over fixed and mobile telephone lines at sample rate of 8 kHz. The utterances of 800 speakers selected are originally rated on a seven point Likert scale converted subsequently into binary classes ‘Likable’ (L) and ‘non-likable’ (NL).

For the Pathology Sub-Challenge, the "NKI CCRT Speech Corpus" (NCSC) is used. This corpus is collected from 55 patients reading Dutch text of neutral content at the department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute. The 2386 segmented sentences are evaluated on an intelligibility scale from 1 to 7 which are then converted into binary classes Intelligible (I) and non-intelligible (NI).

More details about these three corpora can be found in [6].

![Figure 1 UA recall performance comparison of five systems experimented on development data: 4 anchor based-systems (ANC) using Euclidean (-Euc) or Cosine (-Cos) without or with WCCN normalization (-Norm), and a GMM system as fifth system.](image)

### 6.2. Results and discussion

In this section, we study the effect of WCCN normalization on the different tasks. Figure 2 shows the relative gain in the performance when WCCN is applied for both metrics. As we can see, a significant improvement in performance is achieved which can reach 23.6% and 18.5% for Euclidean and Cosine metrics respectively. We also observe that the overall improvement over all tasks is more important for the Euclidean metric (equal to 14.2%) than for cosine distance (7.2%). When we compare the performance based on the absolute value of unweighted average (UA) recall measure as depicted in Figure 2, we observe that the anchor model based on Euclidean metric without WCCN gives the worst results compared to cosine. After normalization, both metrics gives similar performances. We also note that anchor models after WCCN outperforms GMM system. These observations corroborates with results reported in [5].

To visualize the effect of WCCN on the data, we plot in Figure 3 the STV values of the class E before (top figure) and after (bottom) normalization. As we can see, WCCN maps the data in a novel space where the distribution of the data with respect to the anchor models are more scattered. We can assume that more the models are scattered in the likelihood space more the anchor models has the capability to better represent any speaker trait relatively from its most closer to the most distant one and then generates different behavior of an utterance over different models.

![Figure 2. WCCN normalization effect on classification performance obtained on development set, evaluated for each task and for both Euclidean and Cosine metrics.](image)

Table 1 summarizes the parameter values of the anchor model systems optimized for each task. First, we note that except for (N)L task, the feature gaussianization decreases the classification performance. This can be explained by the fact that the feature distribution shape of the SLD corpus on which (N)L is experimented is collected from telephone speech which is more corrupted - given the large variation in the signal acquisition environment used, than those of NCSC corpus for example collected from microphone speech. We also observe that GMM parameters are generally better estimated with ML method except for the SLD corpus. Unexpectedly, we note that removing the silence from the audio features does not always improve performances unlike speaker verification where silence is considered as harmful information. Further studies are needed to explain these results. Finally, we observe that in general, the WCCN normalization with full within-class covariance matrix gives better results except for the (N)I task where the normalization with a diagonal overall covariance matrix is better.
Compared to the baseline results provided by the organizers in [4], we remark that the normalized anchor models gives better results on the development set. On the other hand, Random Forest in particular generalizes better than anchor models on test data. We suspect that anchor model can generalize better if the selection method of the set of anchor models is refined. In fact in this study, we have constructed the anchor model space using the brute force method by including all class models of the all tasks. As shown in Figure 3, WCCN helps to increase the scatter of the models but doesn’t totally eliminate all the redundancy on the behavior of the data with respect to models. An effective method of subset selection of models used to build an optimal anchor space will help to increase the robustness of the anchor models.

Table 1. Parameters of the best anchor model obtained with Cosine metric optimized separately for each task. The parameters are: N for WCCN normalization, S for either the silence is contained (+) in the MFCC files or not, G for feature Gaussianization, followed by the number of Gaussian components used in the front-end of the system.

<table>
<thead>
<tr>
<th>Task</th>
<th>Anchor model</th>
<th>Training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality Sub-Challenge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N)O</td>
<td>70.9 (69.4)</td>
<td>- + - 2  ML</td>
</tr>
<tr>
<td>(N)C</td>
<td>74.5 (75.4)</td>
<td>+ - - 32 ML</td>
</tr>
<tr>
<td>(N)E</td>
<td>85.3 (85.3)</td>
<td>+ - - 8 ML</td>
</tr>
<tr>
<td>(N)A</td>
<td>70.7 (70.5)</td>
<td>+ - - 16 ML</td>
</tr>
<tr>
<td>(N)N</td>
<td>71.4 (71.6)</td>
<td>+ + - 32 ML</td>
</tr>
<tr>
<td>Mean</td>
<td>74.6 (74.4)</td>
<td></td>
</tr>
<tr>
<td>Likability Sub-Challenge</td>
<td>63.5 (63.5)</td>
<td>+ + - 64 MAP</td>
</tr>
<tr>
<td>Pathology Sub-Challenge</td>
<td>67.2 (66.2)</td>
<td>+ - - 2  ML</td>
</tr>
</tbody>
</table>

For normalization, we use diagonal of the global covariance matrix.

7. Conclusions

In this paper we have proposed to solve the problem of the two-class speaker trait classification using anchor models. The anchor space is composed of a set of class models of all tasks. We showed that the introduction of the WCCN normalization on the log-likelihood scores improves significantly the classification performance particularly for the Euclidean distance. We found also that after normalization, Euclidean and cosine metrics gives similar performances. As future work, we will refine the construction of the anchor space by using subset selection techniques to improve robustness of the anchor models.

8. References


Table 2 Three Sub-Challenges results on test data

<table>
<thead>
<tr>
<th>Task</th>
<th>SVM</th>
<th>Random Forest</th>
<th>Anchor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Personality Sub-Challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N)O</td>
<td>57.8 (59.7)</td>
<td>59.0 (63.7)</td>
<td>59.6 (58.2)</td>
</tr>
<tr>
<td>(N)C</td>
<td>80.1 (80.1)</td>
<td>79.1 (79.1)</td>
<td>76.2 (76.2)</td>
</tr>
<tr>
<td>(N)E</td>
<td>76.2 (76.6)</td>
<td>75.3 (75.6)</td>
<td>68.9 (69.1)</td>
</tr>
<tr>
<td>(N)A</td>
<td>60.2 (60.2)</td>
<td>64.2 (64.2)</td>
<td>56.6 (56.7)</td>
</tr>
<tr>
<td>(N)N</td>
<td>65.9 (65.7)</td>
<td>64.0 (63.7)</td>
<td>62.8 (63.7)</td>
</tr>
<tr>
<td>Mean</td>
<td>68.0 (68.5)</td>
<td>68.3 (69.3)</td>
<td>64.8 (64.7)</td>
</tr>
<tr>
<td>Likability Sub-Challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N)L</td>
<td>55.9 (56.1)</td>
<td>59.0 (59.2)</td>
<td>56.5 (56.6)</td>
</tr>
<tr>
<td>Pathology Sub-Challenge</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(N)L</td>
<td>68.0 (66.2)</td>
<td>68.9 (67.5)</td>
<td>67.6 (65.6)</td>
</tr>
</tbody>
</table>