Anchor Models for Emotion Recognition from Speech

Yazid Attabi and Pierre Dumouchel, *Member*, *IEEE*

Abstract—In this paper, we study the effectiveness of anchor models applied to the multiclass problem of emotion recognition from speech. In the anchor models system, an emotion class is characterized by its measure of similarity relative to other emotion classes. Generative models such as Gaussian Mixture Models (GMMs) are often used as front-end systems to generate feature vectors used to train complex back-end systems such as support vector machines (SVMs) or a multilayer perceptron (MLP) to improve the classification performance. We show that in the context of highly unbalanced data classes, these back-end systems can improve the performance achieved by GMMs provided that an appropriate sampling or importance weighting technique is applied. Furthermore, we show that anchor models based on the euclidean or cosine distances present a better alternative to enhance performances because none of these techniques are needed to overcome the problem of skewed data. The experiments conducted on FAU AIBO Emotion Corpus, a database of spontaneous children's speech, show that anchor models improve significantly the performance of GMMs by 6.2 percent relative. We also show that the introduction of within-class covariance normalization (WCCN) improves the performance of the anchor models for both distances, but to a higher extent for euclidean distance for which the results become competitive with cosine distance.

Index Terms—Anchor models, children's speech, emotion recognition, GMM model, skewed distribution, WCCN

1 Introduction

Automatic emotion recognition (AER) from speech has garnered increasing interest in recent years given the broad field of applications that can benefit from this technology. For example, a speaker emotional state recognition system can be used to develop a more natural and effective human-machine interaction system that incorporates an interface exhibiting greater sensitivity toward user behavior. Used in a distance learning context, a tutoring system could detect bored users and allow for a change of style and level of the supplied material, or provide an emotional encouragement [1]. AER may also be used to

1. support the driving experience and incite better driving practices, given that driver emotion and driving performance are often intrinsically linked [2];
2. detect the presence of extreme emotions, especially fear, in the context of public place surveillance [3];
3. automatically prioritize messages accumulated in the mailbox with different criteria such as emotional urgency, mood valence (happy versus sad), and arousal (calm versus excited) [4];
4. use the special features carried by emotions to develop more robust and accurate speaker verification systems [5];
5. assess the urgency of a call to assist in decision taking in the context of a call center offering medical advice to patients [6];
6. or to improve customer service in the context of commercial call centers [7].

Several approaches were investigated to enhance emotion recognition performance particularly the discriminative and generative ones [8]. Potentially promising methods that are yet to be deeply explored are those based on the similarity approach. The similarity-based methodology is, however, a natural way to approach the problem of emotion recognition from speech, where the concept of closeness or distance between classes is clearly present and illustrated in the mapping of categorical emotions onto the dimensional space. Thus, in the dimensional emotion theory, similarity of each emotion class can be easily measured to other classes with respect to some criterion (axis) such as *valence* or *arousal*.

The simplest and most common similarity-based method is the nearest neighbor algorithm which is widely tested. A more sophisticated *emotion profile (EP)-based representation* method was developed by Mower et al. [9]. In this method, emotions are expressed in terms of the presence or absence of a set of component emotions such as anger, happiness, neutrality, and sadness. The EPs are constructed using SVM with Radial Basis Function (RBF). Emotion-specific SVMs are trained for each class as self versus other classifiers. Each EP contains n-components, one for the output of each emotion-specific SVM. The profiles are created by weighting each of the n-outputs (± 1) by the distance between the individual point and the hyperplane boundary. The final

- *Y. Attabi is with the École de Technologie Supérieure (ÉTS) and the Centre de Recherche Informatique de Montréal (CRIM), Montréal, Canada. E-mail: yazid.attabi@crim.ca.*
- *P. Dumouchel is with the Software Engineering and Information Technology Department, École de Technologie Supérieure (ÉTS), Montréal, Canada. E-mail: pierre.dumouchel@etsmtl.ca.*

Manuscript received 8 Mar. 2013; revised 24 June 2013; accepted 10 July 2013; published online 31 July 2013.
Recommended for acceptance by S. Narayanan.
For information on obtaining reprints of this article, please send e-mail to: taffc@computer.org, and reference IEEECS Log Number TAFFC-2013-03-0026.
Digital Object Identifier no. 10.1109/T-AFFC.2013.17.

emotion is selected by classifying the generated profile in a speaker-dependent fashion using Naïve Bayes. In [10], [11], a precursor method based on the similarity concept, named WOC-NN, has been proposed. In this new framework, each emotion is represented by a neighborhood pattern composed of a set of emotion classes ranked according to their closeness or distance to each other. Classification is carried out by computing distances between the test data neighborhood pattern and the specific patterns of each emotion class issued from training.

In this paper, we will investigate in more depth a similar but different method, commonly referred to as anchor models in the speech community, which we have presented in [12].

The anchor model was first introduced for speaker indexing in large audio databases [13] and then extended for speaker identification [14], speaker verification [15], and recently for speaker trait classification [16] problems. In this method, speaker identity is characterized by its relative position in an anchor space. This space is formed by a set of reference speaker models. Different metrics are used to compute the relative position of a given speaker with respect to the set of reference speakers such as euclidean [13], angular [14], [15] or correlation [17] metrics. Several studies show that euclidean distance achieves worse results compared to cosine distance [14], [15], [18]. In [15], a new qualitative measurement based on the rank metric was introduced. This new metric improves the performance compared to the quantitative distances but remains below the performance of the Gaussian Mixture Model-Universal Background Model (GMM-UBM) method. Mami and Charlet [14] have shown that anchor models perform better than GMMs when there is little amount of training data. In [19], the application of Linear Discriminant Analysis (LDA) postprocessing on coordinate vectors of the anchor space allows anchor models to outperform GMMs. Furthermore, a system based on the combination of probabilistic and deterministic anchor model approaches has been proposed in [18] and achieves better results than the GMM-UBM-based system. The probabilistic approach aims to model the intraspeaker variability. Instead of representing the location of a speaker's utterances by only one point in the anchor model space, they are modeled using a normal distribution.

For emotion recognition from speech, the anchor model approach was tested in [20] as a combination method of different classifiers to improve system performance. The experimental framework used in [20] was adopted from language recognition and was composed of two parts: front- and back-end systems. The anchor model was used as back end to fuse two subsystems, namely, prosodic GMM-SVM (support vector machine) and prosodic statistics-SVM systems. Finally, an SVM classifier was used to train the back-end emotions in the anchor model space. The reported results show that the anchor models fusion method significantly improves recognition performance compared to the *sum* rule fusion when tested on two of three corpora.

In this paper, we study the anchor model system acting as a feature extractor rather than as a combination method to recognize emotion from speech. We apply this system to the specific task of recognizing emotional speech of children interacting with a pet robot called Aibo [21]. The corresponding spontaneous emotional speech FAU AIBO Emotion Corpus, described in Section 2, was introduced and made publicly available in Interspeech 2009 Emotion Challenge to provide the community with a medium sized database containing more spontaneous and less prototypical data to reflect more realistic scenarios. Cepstral features are extracted to train GMM models that are used as front end of an anchor model system described in Section 3. In this study, we show that anchor models is an efficient method to classify emotions in the context of highly unbalanced classes as is the case for the FAU AIBO Emotion Corpus. Contrary to speaker diarization and verification problems, anchor models using simple distance metrics such as the *cosine* metric without any preprocessing step achieve better results than GMM models. We also show that the application of within-class covariance normalization (WCCN) on the log-likelihood scores in the anchor space improves even more the performance as detailed in Section 4. In addition, as shown in Section 6, anchor models perform better than more complex and sophisticated classifiers such as SVM-based back-end systems. In Section 5, we investigate the effectiveness of representing at prediction step each emotion class by a set of representative vectors in contrast to a unique vector.

2 Data and Features Description

2.1 Corpus

The proposed framework is tested using the FAU AIBO Emotion Corpus [21]. The data set consists of spontaneous recordings of German children (21 male and 30 female) interacting with a pet robot. The corpus is composed of 9,959 chunks for training and 8,257 chunks for testing, which were collected at two different schools. A chunk is an intermediate unit of analysis between the word and the turn manually defined based on syntactic-prosodic criteria. The average length of the chunk is about 1.7 s. The chunks are labeled into five emotion categories: *Anger* (**A**), *Emphatic* (**E**), *Neutral* (**N**), *Positive* (**P**, composed of *motherese* and *joyful*), and *Rest* (**R**, consisting of emotions not belonging to the other categories such as *bored*, *helpless*, ...). The distribution of the five classes is highly unbalanced. For example, the percentage of training data of each class is as follows: **A** (8.8 percent), **E** (21 percent), **N** (56.1 percent), **P** (6.8 percent), and **R** (7.2 percent).

2.2 Feature Extraction

Mel-Frequency Cepstral Coefficients (MFCC) are used as features to model the varying nature of speech with respect to the type of emotion. The MFCC vector is formed of the first 12 coefficients including *C0* (energy component) calculated at a rate of 10 ms using a 25-ms Hamming window. First and second derivatives are computed using a five-frame window for each MFCC vector to compute the temporal characteristics. Cepstral features are extracted using the HTK toolkit [22]. Silences are removed from the audio files before MFCC extraction.

3 Anchor Models

In an anchor models system, an emotion class is characterized by its measure of similarity relative to other emotion classes. The set of these reference models is called anchor models and forms the anchor space. Three steps characterize the design of an anchor model system: building the anchor space, mapping the acoustic features onto the anchor space, and classifying test emotional speech.

3.1 Building the Anchor Space

In the case of a pattern recognition problem with unlimited number of classes such as in the speaker verification task, we need to find a set of speakers or virtual speakers (by clustering speakers) which is the most representative of all speakers. When the problem at hand involves a limited number of classes, as for the emotion recognition task, we have the opportunity to model the entire set of emotion classes. Thus, in this type of multiclass problem, all classes have the advantage of being well represented in the anchor space. Therefore, we can point out two main differences between the anchor models in speaker recognition and in emotion recognition. First, for speaker recognition the anchor space has a high dimension, composed of hundreds of speaker models. For emotion recognition, the anchor space dimension is very small because of the limited number of emotion classes available. Second, in speaker recognition, the speaker to characterize in the anchor space, during training or test stage, does not usually belong to the set of anchor models. On the other side in emotion recognition, the emotion appertains to the set of anchor models, owing that all emotion class models are used as anchor models.

If each of the C emotion classes is modeled by a GMM λ_i using their MFCC speech features, the reference space could be defined by the set $\Gamma = \{\lambda_A, \lambda_E, \lambda_N, \lambda_P, \lambda_R\}$.

GMM is a generative model widely used in the field of speech processing. It is a probabilistic method that offers the advantage of adequately representing speech signal variability using a mixture of sufficient number of Gaussians. Given a GMM modeling a D-dimensional vector, the probability of observing a feature vector given the model is computed as follows:

$$P(\mathbf{x} \mid \lambda) = \sum_{k=1}^{m} w_k \, N(\mathbf{x}; \mu_k, \mathbf{\Sigma}_k), \tag{1}$$

where m, w_k, $\boldsymbol{\mu}_k$, and $\mathbf{\Sigma}_k$ correspond to the number of Gaussians, weight, mean vector, and diagonal covariance matrix of the kth Gaussian, respectively. GMM parameters are estimated using the maximum-likelihood (ML) approach based on the expectation maximization (EM) algorithm [23].

Note that we opted for the use of GMM rather than hidden Markov model (HMM) in light of previous results achieved on the FAU AIBO Emotion Corpus in the Interspeech 2009 Emotion Challenge 2009. For the set of HMM baseline systems studied in [8], the single-state HMM (namely a GMM) performs slightly better than its tristate and slightly worse than its five-state representation. In [24], with a higher number of Gaussian mixtures (rather than the two used in [8]), the GMM model outperforms the HMM of the baseline system.

3.2 Mapping onto the Anchor Space

Let $\mathbf{X}$ be an utterance of emotional speech represented by a sequence of frames, $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$. $\mathbf{X}$ is mapped onto the anchor space by computing the vector $\mathbf{L}(\mathbf{X})$, where each element of the vector represents the mean log-likelihood score of $\mathbf{X}$ against each model λ_i. We refer to this new representation as *Emotion Characterization Vector* (ECV) by analogy to the terminology used in speaker recognition. It is

$$\mathbf{L}(\mathbf{X}) = \begin{bmatrix} \frac{1}{T}\log P(\mathbf{X} \mid \lambda_1) \\ \vdots \\ \frac{1}{T}\log P(\mathbf{X} \mid \lambda_C) \end{bmatrix}, \tag{2}$$

where $\log P(\mathbf{X}|\lambda_i)$ is the log likelihood of the feature vectors $\mathbf{X}$ given a GMM λ_i that belongs to the set of class models $\{A, E, N, P, R\}$ and $\mathbf{L}(\mathbf{X})$ represents the ECV of $\mathbf{X}$. Assuming the independence of the frames, $\log P(\mathbf{X}|\lambda_i)$ is computed according to

$$\log P(\mathbf{X} \mid \lambda_i) = \sum_{n=1}^{T} \log P(\mathbf{x}_n \mid \lambda_i). \tag{3}$$

Two types of ECV vectors are computed using (4) depending on the values of $\mathbf{X}$: 1) a *class representative* ECV vector for each emotion class computed during training stage and 2) a *test* utterance vector at prediction phase. A class representative vector $\mathbf{L}^i$ for emotion class i is estimated using all training utterances of class i according to

$$\mathbf{L}^i = \frac{1}{n_i} \sum_{q}^{n_i} \mathrm{L}(\mathbf{X}_q^i), \tag{4}$$

where $\mathbf{X}_q^i$ represents the qth utterance of class i and n_i the number of training utterances of class i.

3.3 Emotional Speech Classification

To classify a test speech, the distance between the ECV of the test data and those of each class representative is computed using either euclidean or cosine distance metrics defined as

- *Euclidean metric*:

$$d(\mathbf{L}_1, \mathbf{L}_2) = \sqrt{|\mathbf{L}_1 - \mathbf{L}_2|^2}. \tag{5}$$

- *Cosine metric*:

$$d(\mathbf{L}_1, \mathbf{L}_2) = 1 - \frac{\langle \mathbf{L}_1, \mathbf{L}_2 \rangle}{\|\mathbf{L}_1\| \|\mathbf{L}_2\|}, \tag{6}$$

where $\langle \mathbf{L}_1, \mathbf{L}_2 \rangle$ is the dot product of the vectors $\mathbf{L_1}$ and $\mathbf{L_2}$.

In this section, each emotion class of the C classes is represented at test phase by a unique ECV vector. The decision rule is formulated as follows:

$$\text{emotion} = \underset{i=1,\ldots,C}{\arg\min}(d(\mathbf{L}_T, \mathbf{L}_i)), \tag{7}$$

where d represents the metric used to compute the distance between $\mathbf{L}_T$, the ECV of the test data, and $\mathbf{L}_i$, the representative ECV of the emotion class i.

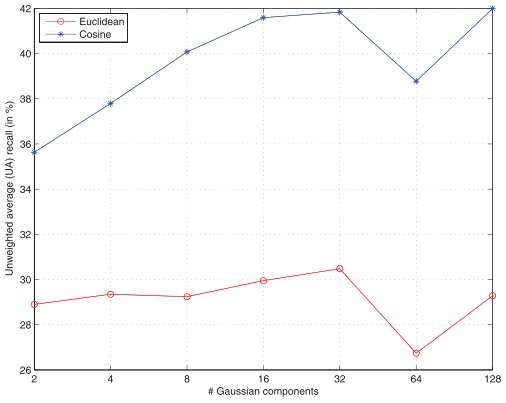

Fig. 1. UA recall results achieved using ninefold cross validation on FAU AIBO Emotion training data with respect to the number of Gaussians of the GMM. Euclidean and cosine distance-based anchor models systems are compared.

3.4 Experimental Setup

In this section, the performances of anchor models are evaluated for both the euclidean and cosine metrics. The model parameters such as the number of GMM Gaussian components are tuned based on the training data using the ninefold cross-validation protocol. Each of the nine partitions contains a disjoint set of speakers. The results are optimized via maximization of the unweighted average (UA) recall measure and second the weighted average (WA) recall (i.e., accuracy) given that FAU AIBO Emotion classes are highly unbalanced. Note that a baseline classifier that predicts all the test data as being in the same class as the dominant class, namely *Neutral*, will achieve 65 percent of accuracy but only 20 percent of UA recall. Note that this is the same proposed measure for the Interspeech 2009 Emotion Challenge. Therefore, our results can be compared with the state of the art.

Fig. 1 shows the results obtained for each system evaluated using ninefold cross validation on the training data. We observe that the euclidean distance-based anchor model achieves very poor performances compared to the cosine distance-based system.

The results suggest that speech emotion utterances mapped onto the anchor models space are more discernible through their directions rather than their Cartesian coordinates. This implies that noise adversely affects features in their magnitude and therefore degrades the likelihood score of data against each emotion model by the same multiplicative constant.

To illustrate this mismatch on the length of the vector $\mathbf{L}$, the mean and variance of the Cartesian values of each variable of the ECV vectors are plotted in Fig. 2. The training data are used to compute the statistics of each emotion class separately in each plot. In the optimal case, the mean log-likelihood score of an emotion class data will get the maximum value for the component corresponding to its own model. For the other components of the ECV vector, more a model of another emotion class is close to its own model more the log-likelihood score is higher and vice versa. The amount of score reflects the degree of similitude of a given class relatively to other classes.

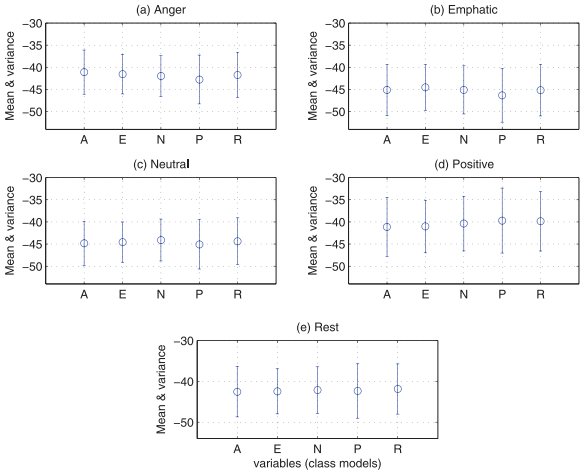

Fig. 2. Each graphic plots the mean and variance parameters representing the data distribution of one emotion class with respect to each emotion model (variables of ECV vectors). In this figure, the plotted values represent the statistics of the Cartesian coordinates of the ECV vectors (see (2)).

In Fig. 3, we have plotted the mean and variance parameters of the angular value of each Cartesian coordinate variable of the ECV vectors mapped in Fig. 2. The ith variable of the angular vector represents the angle between the ECV vector and the ith axis of the euclidean space. Formally, the angular vector of an ECV vector $\mathbf{L}$ is computed as follows: Let $\mathbf{L} = (l_1, l_2, \ldots, l_C)^T$ be an ECV vector and $(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_C)$ be the standard basis, where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T, \ldots,$ and $\mathbf{e}_C = (0, 0, \ldots, 1)^T$. The angular value between $\mathbf{L}$ and the ith standard axis is equal to

$$\text{angle}(\mathbf{L}, \mathbf{e}_i) = \arccos\left(\frac{<\mathbf{L}, \mathbf{e}_i>}{\|\mathbf{L}\| \times \|\mathbf{e}_i\|}\right), \tag{8}$$

after simplification we get

$$\text{angle}(\mathbf{L}, \mathbf{e}_i) = \arccos\left(\frac{l_i}{\|\mathbf{L}\|}\right). \tag{9}$$

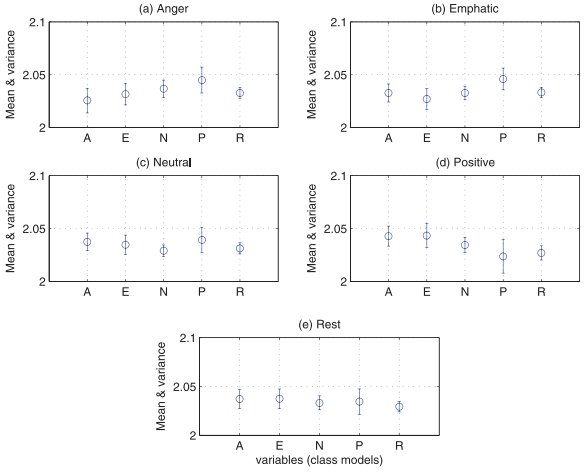

Fig. 3. Each graphic plots the mean and variance parameters representing the data distribution of one emotion class with respect to each emotion model (variables of ECV vectors). In this figure, the plotted values represent the statistics of the angular values of the ECV vectors with respect to the standard basis (see (9)).

We have also

$$\cos(\text{angle}(\mathbf{L}, \mathbf{e}_i)) = \frac{l_i}{\|\mathbf{L}\|}. \quad (10)$$

Fig. 2 reveals that the variables (which represent the output of the density function of the GMM models of the emotion classes) are not discriminative by their Cartesian values due particularly to the large scale of their variance. Furthermore, the variance-scaling problem is less pronounced with the angular values of the variables as depicted in Fig. 3.

It is interesting to note that (10) resembles the length normalization formula that has recently been used in [25] and [26] as a preprocessing step to enhance recognition performance. In fact normalizing the length means using the cosine of the angles of the variables that are more discriminative as depicted in Figs. 2 and 3.

In the next section, we propose to deal with the variance scale problem by applying within-class covariance normalization to further enhance the discrimination between different class models.

4 WCCN Normalization

WCCN is a technique introduced in [27] to train a generalized linear kernel of an SVM-based system to minimize the expectation of false-positive and false-negative errors. The generalized linear kernel $\mathrm{k}(\mathbf{L}_1, \mathbf{L}_2)$ is expressed as

$$\mathrm{k}(\mathbf{L}_1, \mathbf{L}_2) = \mathbf{L}_1^t \mathbf{R} \mathbf{L}_2, \quad (11)$$

where $\mathbf{L}_1$ and $\mathbf{L}_2$ are two given instances and $\mathbf{R}$ is a positive semidefinite matrix. The closed-form solution is reached by setting $\mathbf{R} = \mathbf{W}^{-1}$, where $\mathbf{W}$ is the expected within-class covariance matrix of the data defined as

$$\mathbf{W} = \sum_{i=1}^{C} p(i) \cdot \mathbf{S}_i, \quad (12)$$

where $p(i)$ and $\mathbf{S}_i$ represent the prior probability and within covariance matrix of class i, respectively. If we define $\mathbf{A}$ as the *Cholesky* factorization of $\mathbf{W}^{-1}$, namely $\mathbf{A}\mathbf{A}^T = \mathbf{W}^{-1}$, the new metric of cosine distance for instance, when WCCN is applied on ECV vectors, is expressed as

$$\mathrm{d}(\mathbf{L}_1, \mathbf{L}_2) = 1 - \frac{(\mathbf{A}^T \mathbf{L}_1)^T (\mathbf{A}^T \mathbf{L}_2)}{\|\mathbf{A}^T \mathbf{L}_1\| \|\mathbf{A}^T \mathbf{L}_2\|}. \quad (13)$$

WCCN has also been successfully applied on *i-vector* feature space [25]. 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. WCCN was also applied in [28] to improve the performance of SVM using likelihood scores as features.

Finally, an intraspeaker normalization method for anchor model-based speaker verification, called vectorial Z-normalization, was introduced in [29]. VZ-norm is an extension of Z-norm to the multivariate case that aims to normalize the score against the intraspeaker variability. VZ-norm is similar to WCCN in that the normalization in both methods is based on the use of the within-class covariance matrix. On the other side, the two methods differ in that no mean normalization is required for WCCN.

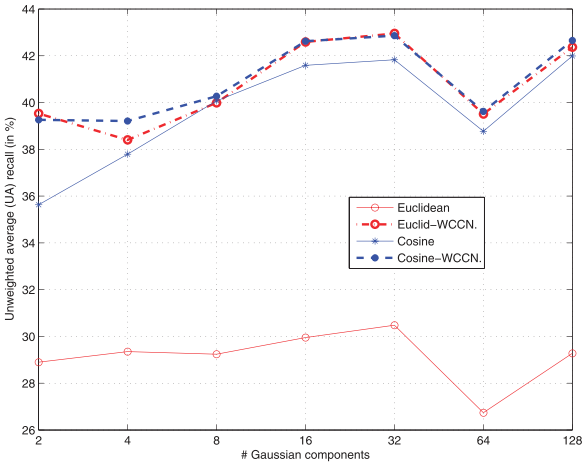

Fig. 4. Effect of WCCN normalization on UA recall performance of anchor models with respect to the number of Gaussians of the GMM. The results are obtained using ninefold cross validation on FAU AIBO Emotion training data. The performances of systems before and after WCCN normalization are also compared.

4.1 Results and Discussion

Fig. 4 shows the classification results for anchor models, based on euclidean and cosine distance metrics, and evaluated before and after WCCN normalization on the training data using ninefold cross validation. First, we observe that WCCN enhances performances for both metrics with a marked improvement for euclidean distance, the mean gains being 3.3 and 40 percent for cosine and euclidean, respectively. It is also interesting to note that after applying normalization, euclidean and cosine metrics show similar performances. We also note that the best performance is obtained for both metrics with a GMM model of 32 Gaussian components used as a front-end system. Accordingly, a 32 Gaussian GMM mixture is selected for the test experiment.

To visualize the effect of WCCN normalization on the data (log-likelihood scores), we have plotted the distribution of *anger* class utterances over the emotion models before and after normalization in Fig. 5. In the top of Fig. 5, we observe that the data of emotion class **A** exhibit almost the same behavior with respect to its own model (**A**) as toward any other model. This behavior makes it difficult to take advantage of the anchor model and get any discriminative information from learning the relative behavior of an utterance over different models.

As depicted in the bottom plot of Fig. 5, WCCN normalization has the effect of maximizing the discriminative capability between models in the anchor space, as evidenced by a greater distribution of the models over the entire range of possible anchor space scores.

Table 1 gives the performance results of anchor models on the test data. The models used in the test stage are trained with all the training data. The matrix A of WCCN is also estimated using log-likelihood scores of training data using ninefold cross validation. We note that the observations pertaining to the training data extend equally

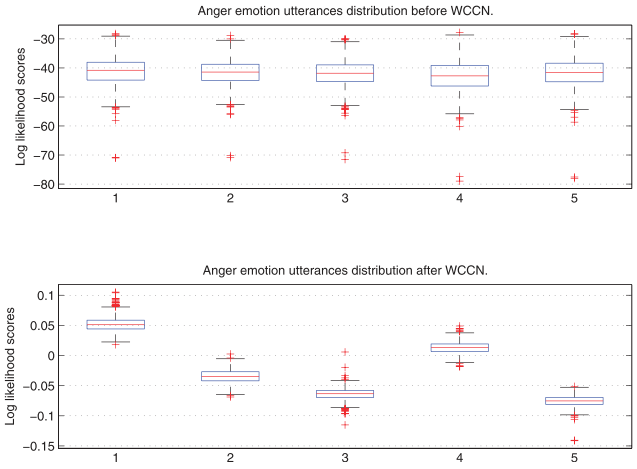

Fig. 5. Box plot of the distribution of the Anger (A) emotional speech over the five emotion models before (top) and after (bottom) WCCN normalization. On each box, the central mark is the median, the edges are the 25th and 75th percentiles, the whiskers denote the most extreme data, and outliers are plotted individually.

well to the test data. WCCN improves performance significantly (using the *McNemar* statistical test) for both metrics. We also note that euclidean and cosine distances achieve comparable performances after normalization. Interestingly, we observe that the results obtained for the test data are actually better than those of the training data. This is explained by the fact that the GMM models used in the test are more robust than those used in the cross-validation protocol. For test models we have used nine partitions instead of the eight used in the ninefold cross validation, namely a difference of three additional speakers' data. To verify this assertion, we evaluated the test data using the same models used for training data evaluation. The performance achieved for the anchor model system using euclidean distance, for example, drops from 44.19 to 42.18 percent, giving results worse than training data. This result confirms our assertion and emphasizes the importance of having more data and speakers for developing more robust systems.

5 Class Representative Vectors

One key difference between the anchor models system presented in this work and the k-nearest neighbor method is the identity of training data points that are used to compare the test data at classification step. In the k-NN method, the classification is approximated locally and is based on the k closest examples. A major inconvenience of this method stems from its sensibility toward the outliers contained in the training data. The anchor model offers the advantage of comparing the test data to a more reliable fixed vector used as class representative that is determined during the training stage.

5.1 Unifold versus Multifold Representatives

In the design of the anchor models presented in the previous sections, each emotion class was represented by a unique ECV vector that is computed using all the training data associated with that class. Another alternative is to represent each emotion class by a set of representative vectors. A model with multiple class representatives could be particularly useful when the data have a multimodal distribution. In this section, we investigate the impact of duplicating the number of representative vectors on the anchor models performance. In the rest of the paper, we dub *multifold* the system based on more than one representative vectors as opposed to the *unifold* system based on a unique representative ECV.

If we assume that $\{\mathbf{L}_1^i, \mathbf{L}_2^i, \ldots, \mathbf{L}_r^i\}$ represents the set of representative vectors of an emotion class $\mathbf{E}_i$, the decision rule (7) becomes

$$\text{emotion} = \underset{i=1,\ldots,C}{\arg\min} \left(\sum_{j=1}^{r} \mathrm{d}(\mathbf{L}_T, \mathbf{L}_i^j) \right), \tag{14}$$

where $\mathbf{L}_i^j$ represents the jth representative vector of the ith emotion class.

Different methods can be used to select the class representatives of a multifold system. Two methods will be studied and compared to the performance of a unifold system:

1. *Random selection*. From each emotion class, r utterances are randomly selected from its training data.
2. *Clustering method*. The training data of each emotion class are clustered into $r(r \geq 2)$ clusters based on the distance between their ECV values.

We also investigate a weighted cluster version of anchor models. The aim is to reduce the effect of clusters composed of outlier instances. The contribution of each cluster in the computation of the distance will be proportional to the number of instances in the cluster. The new decision rule is formulated as follows:

$$\text{emotion} = \underset{i=1,\ldots,C}{\arg\min} \left(\sum_{j=1}^{r} \frac{n_i^j}{n_i} \times \mathrm{d}\left(\mathbf{L}_T, \mathbf{L}_i^j\right) \right), \tag{15}$$

where n_i and n_i^j represent the size of the training data of class i and the jth cluster of class i, respectively.

TABLE 1
Results of Different Anchor Models Systems Tested on Test Data

Anchor models systems	Recall [%]	
	UA	WA
Cosine	42.25	33.57
Euclidean	26.59	23.00
Cosine + WCCN	43.91	46.01
Euclidean + WCCN	**44.19**	47.44

5.2 Experiment Results

To select the number of representative vectors that optimize the performance of the multifold systems, we first carry out experiments on the training data according to different number of representative vectors used per class. For the multifold system based on random selection, 50 runs are executed. At each run, a new subset of class representative vectors is randomly selected and the UA recall performance is evaluated. The means of these runs are computed and plotted in Fig. 6. We observe that the overall performance increases sharply as the number of class representatives increases up to a value of 150 vectors

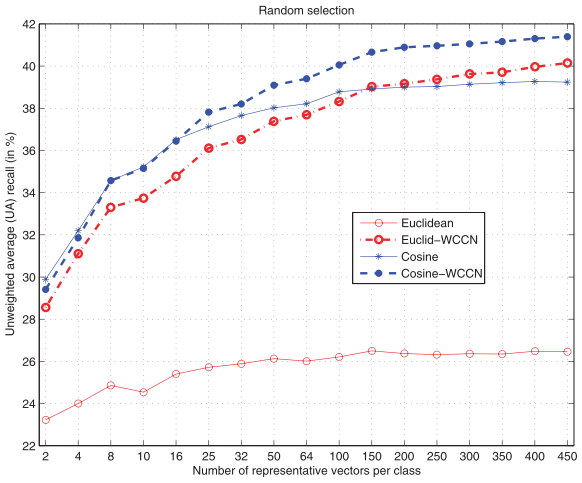

Fig. 6. UA recall mean results of 50 runs of the anchor models systems with respect to the number of ECV vectors per class selected as class representative vectors. At each iteration, a new subset of training data is randomly selected as class representative vectors. Performances are evaluated on the training set using ninefold cross validation.

for which the best results are obtained for the anchor models systems without WCCN normalization. For the normalized system versions, performances improve continually and slowly until 450 vectors.

The performance of the clustering-based multifold system is depicted in Fig. 7. The best performance is reached with only two clusters and the performance can drop drastically for a higher number of clusters. Furthermore, when the clusters are weighted proportionally to their size, the performances are more stable as the number of clusters changes, as depicted in Fig. 8. The performances then become less sensitive to clusters composed of outliers although the best results are not enhanced with the weighting operation. The results achieved on test data are reported in Table 2. As we can observe, increasing the number of representative vectors, however they are selected, does not improve performance. This result suggests that the data can be treated as a unimodal distribution.

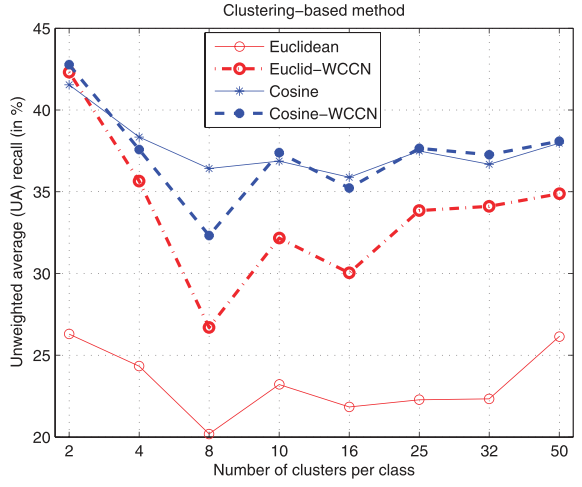

Fig. 7. UA recall results of anchor models systems with respect to the number of clusters per class. The centers of clusters are used as class representative vectors. Performances are evaluated on the training set using ninefold cross validation.

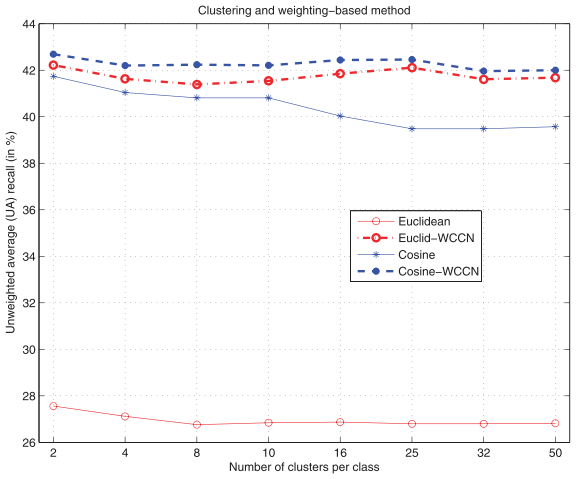

Fig. 8. UA recall results of anchor models systems with respect to the number of center clusters per class used as class representative vectors. Clusters are weighted with a value proportional to the class size. Performances are evaluated on the training set using ninefold cross validation.

6 More Complex Back-End System

The likelihood probability values computed using GMM models could be used directly as final classification scores using *Bayes* decision rule (first system architecture type). As second system architecture type, the likelihood values could also be viewed as high-level feature entries for another but *simple* classifier without any learning stage. The anchor models system was an example of a basic classifier based on a similarity concept that makes use of likelihood scores as features. In the third case of architecture, the likelihood scores are used as input for a more complex back-end system with a more sophisticated training algorithm such as SVM or multilayer neural network (MLP). Such a two-stage architecture has already been successfully applied in the literature as is the case for unconstrained handwritten numeral classification [30] and offline signature verification [31] problems. In both studies, an HMM is used for the first stage to calculate similarity measures that populate feature vectors used to train an SVM (or ensemble of SVMs). An improvement of 1.23 percent has been achieved over HMM for handwritten numerals problem, while the reduction in individual error rates could reach 10 percent for signature verification.

In [20], a GMM/SVM architecture was used as a means of multisystem combination for emotion recognition but no comparison results were reported against the GMM model used as baseline. For the speaker recognition task, the

TABLE 2
Comparison between Multifold and Unifold Anchor Models Systems Evaluated on Test Data of FAU AIBO Emotion Corpus

System	UA	WA
Random	42.55 %	45.40 %
Clustering	43.94 %	48.84 %
Weighted-Clustering	43.41 %	49.73 %
Unifold	44.19 %	47.44 %

GMM-SVM approach is also investigated in [28] and [32]. In [28], the GMM-SVM system achieved comparable performance to the GMM-UBM system. In [32], where Gaussian distributions in the UBM were used as a reference space, GMM-SVM achieved better results than anchor models using euclidean distance as decision metric for both speaker verification and identification problems.

6.1 Processing Skewed Data

When a discriminative model such as SVM is used, we particularly need to deal with the problem of unbalanced data distribution. Without data sampling techniques, performance will be boosted in favor of the most represented class at the expense of the other classes. Several methods are proposed in the literature to mitigate the impact of a skewed class distribution:

1. Downsampling by reducing the size of the majority class to the size of the minority class [33];
2. oversampling by generating new samples of minority class using algorithms such as Synthetic Minority Oversampling Technique (SMOTE) [34]; and
3. ensemble sampling [35] that consists in undersampling the majority class into an ensemble of data subsets. Each subset is used in turn to train a separate classifier.

In [36], these three sampling methods were tested and compared using SVM as classifier. SVM was trained using the baseline feature set of Interspeech 2009 Emotion Challenge extracted from FAU AIBO Emotion Corpus. In terms of UA recall performance, these techniques are ranked as follows: SMOTE, downsampling followed by ensemble sampling.

In the same study, the author showed that importance weighting presents a better alternative than sampling methods for optimizing the unweighted average recall of skewed data. This technique consists in applying an importance value for each training data in the optimized objective function at the training stage. The value of the importance weight is inversely proportional to the class size. The objective function of SVM trained with hinge-loss is expressed as follows:

$$\min \mathbf{V}^T \mathbf{V} + c \sum_j \xi_j, \tag{16}$$

where c is the slope of the hinge function, $\mathbf{V}$ is a vector normal to the decision boundary, and ξ_j is a slack variable. After introducing the importance weighting, the objective function is rewritten as follows:

$$\min \mathbf{V}^T \mathbf{V} + c \sum_j \gamma_j \xi_j, \tag{17}$$

where γ_j represents the importance weight of data point j which is equal to $\frac{1}{C_j}$, the inverse of the size of the class that data point j belongs to.

6.2 Experiment Results

The goal of this section is to assess the relative efficiency of the three aforementioned system architectures toward recognition performance improvement of a multiclass emotion problem using anchor model features in the context of highly unbalanced classes. For this purpose, a GMM model is evaluated using *Bayes* decision rule, with equal prior probability classes, as first architecture. The test recording is classified according to the emotion class label that maximizes the log-likelihood value over all class models:

$$\text{emotion} = \arg \max_{i=1,\ldots,C} (\log P(\mathbf{X} \mid \lambda_i)). \tag{18}$$

Note that we have already tested the GMM-UBM system in [24] and found that it achieves slightly worse results than the GMM model. As second type of architectures, namely systems using GMM scores as features without any further training stage, we investigate the kNN method in addition to the anchor models described in Section 6. Finally, for the third type of system architectures that use a more sophisticated front-end system, we experiment four classifiers: SVM, MLP, logistic regression, and random forest.

As techniques used to overcome the problem of unbalanced data, the best three methods reported in [36] are investigated, namely, importance weighting, SMOTE, and downsampling techniques. Two variants of downsampling are tested. In the first, the *neutral* (majority) class is reduced to the size of second most frequent class (*Emphatic*). In the second, all majority classes are downsampled to the size of the minority class (*positive*). The experiments are conducted using WEKA software [37]. The importance weighting is also performed using WEKA software that gives the option to weight each data instance at the end of each data line in the ARFF file.

The results of different systems evaluated on the test data are reported in Table 3. Several observations can be reported. First, we note that when original data are used without any sampling or weighting techniques, complex back-end systems give the worst results. This is particularly true for SVM for which performances are comparable to the chance baseline system, classifying all data as being of majority class (**N**). Second, we observe that the introduction of importance weighting and sampling techniques remedy, in general, the problem of unbalanced data. Although the best technique depends on the type of classifier used, the technique of downsampling to the least frequent class generally leads to better results. For SVM with polynomial kernels of degree one or three and for MLP system, downsampling to the least frequent class outperforms not only other techniques but also represents the only technique that can improve the performance achieved by GMM. For SVM with *Radial Basis Function* kernel and random forest classifiers, the best results are achieved with SMOTE technique; however, the UA performances remain lower than GMM. Regarding importance weighting, this technique behaves differently depending on the classifier on which it is used. On SVM and particularly on RBF kernel, the effect observed on skewed data is reversed when the weights are applied. Namely, all data of one of the minority classes are correctly classified at the expense of other classes. On the other hand, importance weighting comes at the first position compared to the other techniques and succeeds in alleviating the problem of unbalanced data when tested with kNN and logistic regression methods. We also observe that downsampling to the least frequent class achieves good results as well, however slightly below those

TABLE 3
Comparison of the Three Different Types of Systems Evaluated on Test Data of FAU AIBO Emotion Corpus

Systems	Recall (%) A	E	N	P	R	Unweighted Average
Without back-end system						
GMM-Bayes	46,97	49,27	41,83	46,95	23,23	41.65
Similarity-based systems						
kNN (k=5)	28.3	37	74.3	11.3	2.2	30.62
kNN-S (k=41)	49.3	45	38.5	49.3	18	40.02
kNN-D1 (k=21)	30.8	63	56.7	23.9	2.2	35.32
kNN-D2 (k=75)	53.5	45.1	48.5	52.1	10.6	**41.96**
kNN-W (k=211)	51.1	48.1	47.2	53.1	15.1	**42.92**
Anchor models	55.97	47.02	49.79	55.35	12.82	**44.19**
Complex back-end systems						
SVM *(linear)*	0	4.8	98.9	0	0	20.74
SVM-S *(linear)*	40.8	52.9	53	32.9	16.7	39.26
SVM-D1 *(linear)*	3.3	67.8	68.4	3.7	0	28.64
SVM-D2 *(linear)*	56.6	42.7	47.1	59.2	14.3	**43.98**
SVM-W *(linear)*	83.8	48.1	0	0	0	26.38
SVM *(polynomial d=3)*	8.2	15.6	95.8	0	0	23.92
SVM-S *(d=3)*	39.1	53.1	55.5	32.9	16.2	39.36
SVM-D1 *(d=3)*	17.3	65.8	67.6	11.7	0	32.48
SVM-D2 *(d=3)*	55.3	47.8	50.2	52.1	13.4	**43.76**
SVM-W *(d=3)*	100	0	0	0	0	20
SVM *(RBF kernel)*	0	0	100	0	0	20
SVM-S *(RBF)*	60.1	31	28.4	23	55.6	39.62
SVM-D1 *(RBF)*	0	76.1	56.8	0	0	26.58
SVM-D2 *(RBF)*	70.9	3.4	1.4	0	75.1	30.16
SVM-W *(RBF)*	100	0	0	0	0	20
Logistic	21.1	31.8	87.4	9.4	0	29.94
Logistic-S	40.1	52.7	53.1	34.7	17.5	39.62
Logistic-D1	28.8	66.6	62.1	28.2	0	37.14
Logistic-D2	57.6	45.4	44.5	59.6	13.8	**44.18**
Logistic-W	58.6	47.3	41.8	59.2	15.1	**44.4**
MLP	18.2	22.5	91.1	10.3	0	28.42
MLP-S	38.6	40	65	35.7	11	38.06
MLP-D1	20.8	78.1	48.5	25.8	0	34.64
MLP-D2	55	60.7	32.5	62.9	0.7	**42.36**
MLP-W	0	80.2	0	0	5.2	26.44
Random forest	26.8	37.1	75.5	9.4	2.4	30.24
Random forest-S	38.3	44.3	51	28.2	13.6	35.08
Random forest-D1	33.7	57.2	47.7	23.5	5.8	33.58
Random forest-D2	48.6	39.2	33.9	32.4	14.1	33.64
Random forest-W	36.7	50.1	48.7	20.7	6.5	32.54

Systems Sys-X stands for system Sys using a technique X to overcome the problem of skewed data. X can take the value: W for importance weighting, S for SMOTE, D1 or D2 for downsampled to the second most frequent or least frequent class size, respectively. Results that outperform baseline system (GMM) are highlighted in boldface.

of importance weighting. Both techniques when used with NN and logistic regression outperform GMM models.

If we compare the two variants of downsampling together, we find that undersampling to the least frequent class size always gives better results for the tested classifiers compared to undersampling to the second most frequent class size. We also observe that the best technique used to overcome the problem of skewed data depends not only on the type of classifier but also on the type of features. Indeed, SVM trained with likelihood score data in this work achieves better results with downsampling technique, while importance weighting technique performs better when SVM is trained with suprasegmental acoustic features studied in [36].

TABLE 4
Comparison of Anchor Models Performances with State-of-the-Art Tested on FAU AIBO Emotion Test Corpus

Systems	Recall [%] UA	WA
Schuller et al. (IS2009 baseline system) [8]	38.20	39.20
Lee et al. (Bayesian logistic regression)[38]	41.30	43.90
Kockmann et al. (fusion of 2 joint factor analysis systems)[39]	41.70	-
Authors (WOC-NN)[11]	43.14	35.33
Schuller et al. (majority voting of best IS2009 contributions) [40]	**44.0**	-
Anchor models-Euclidean	**44.19**	47.44
Logistic-W	**44.40**	42.78

Finally, it is interesting to note that anchor models based on simple metrics such as euclidean is the only system that is capable of improving performance over GMM (6.2 percent relative) without use of any sampling or importance weighting techniques. Furthermore, euclidean metric outperforms all other complex systems even if these different techniques are applied except for logistic regression when tested with importance weighting, which gives slightly better results. This gives evidence that anchor models used with distance metrics are less sensitive to the skewed distribution owing to the use of a balanced number of global and reliable representative vectors for each class.

In Table 4, we compare anchor models to the state of the art using the same corpus and same experimental protocol. We observe that the anchor models system with euclidean distance outperforms the baseline [8], the best single [38], and combined [39] systems of Interspeech 2009 Emotion Challenge by 15.8, 7.1, and 6.1 percent relative, respectively. Furthermore, anchor model offers better results than WOC-NN (another similarity-based method) with a relative improvement of 2.6 percent in UA recall and interestingly also achieves a 33 percent relative gain in WA recall (accuracy). Finally, anchor model slightly outperforms the system obtained by the majority vote fusion of the best contributions of Interspeech 2009 Emotion Challenge [40].

7 Conclusion

In this paper, we have presented anchor models, a similarity-based method, to solve the multiclass emotion recognition problem. We have shown that after WCCN normalization, euclidean or cosine distances can be indifferently used as decision metric to significantly improve performance of the front-end system, namely the GMM model. A relative gain of 6.2 percent is achieved using euclidean distance. We also showed that some of the more

complex and sophisticated classifiers used as back-end systems can improve performance provided that an appropriate sampling or importance weighting technique is used. The best technique, selected to overcome the problem of skewed class distribution, is classifier and features dependent. Thus, by virtue of its algorithmic simplicity that does not require any parameter tuning, its low time execution complexity, and finally its insensitivity toward unbalanced data, the anchor models system based on distance metrics represent an attractive solution to improve on the performance of generative models such as GMM.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the many helpful comments and suggestions that they received from anonymous reviewers.

REFERENCES

[1] W. Li, Y. Zhang, and Y. Fu, "Speech Emotion Recognition in E-Learning System Based on Affective Computing," *Proc. Third Int'l Conf. Natural Computation (ICNC '07)*, pp. 809-813, 2007.

[2] C.M. Jones and I.-M. Jonsson, "Performance Analysis of Acoustic Emotion Recognition for In-Car Conversational Interfaces," *Proc. Fourth Int'l Conf. Universal Access in Human-Computer Interaction: Ambient Interaction*, pp. 411-420, 2007.

[3] C. Clavel et al., "De la Construction du Corpus Émotionnel au Système de Détection le Point de Vue Applicatif de la Surveillance dans les Lieux Publics," *Revue d'Intelligence Artificielle*, vol. 20, nos. 4/5, pp. 529-551, 2006.

[4] Z. Inanoglu and R. Caneel, "Emotive Alert: HMM-Based Emotion Detection in Voicemail Messages," *Proc. Int'l Conf. Intelligent User Interfaces (IUI '05)*, pp. 251-253, 2005.

[5] A.R. Panat and V.T. Ingole, "Affective State Analysis of Speech for Speaker Verification: Experimental Study, Design and Development," *Proc. Int'l Conf. Computational Intelligence and Multimedia Applications*, pp. 255-261, 2007.

[6] L. Devillers and L. Vidrascu, "Real-Life Emotion Recognition in Speech," *Speaker Classification II*, pp. 34-42, Springer-Verlag, 2007.

[7] C. Lee and S. Narayanan, "Towards Detecting Emotions in Spoken Dialogs," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 2, pp. 293-302, Mar. 2005.

[8] B. Schuller, S. Steidl, and A. Batliner, "The INTERSPEECH 2009 Emotion Challenge," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '09)*, 2009.

[9] E. Mower, M.J. Mataric, and S.S. Narayanan, "A Framework for Automatic Human Emotion Classification Using Emotional Profiles," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 5, pp. 1057-1070, July 2011.

[10] Y. Attabi and P. Dumouchel, "Weighted Ordered Classes—Nearest Neighbors: A New Framework for Automatic Emotion Recognition from Speech," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '11)*, 2011.

[11] Y. Attabi and P. Dumouchel, "Emotion Recognition from Speech: WOC-NN and Class-Interaction," *Proc. 11th Int'l Conf. Information Science, Signal Processing and Their Applications (ISSPA '12)*, 2012.

[12] Y. Attabi and P. Dumouchel, "Emotion Recognition from Children's Speech Using Anchor Models," *Proc. Workshop Child, Computer and Interaction (WOCCI '12)*, 2012.

[13] D. Sturim, D. Reynolds, E. Singer, and J. Campbell, "Speaker Indexing in Large Audio Databases Using Anchor Models," *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing (ICASSP '01)*, pp. 429-432, 2001.

[14] Y. Mami and D. Charlet, "Speaker Identification by Location in an Optimal Space of Anchor Models," *Proc. Int'l Conf. Spoken Language Processing (ICSLP '02)*, vol. 2, pp. 1333-1336, 2002.

[15] Y. Yang, M. Yang, and Z. Wu, "A Rank Based Metric of Anchor Models for Speaker Verification," *Proc. IEEE Int'l Conf. Multimedia and Expo (ICME '06)*, pp. 1097-1100, 2006.

[16] Y. Attabi and P. Dumouchel, "Anchor Models and WCCN Normalization for Speaker Trait Classification," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '12)*, 2012.

[17] M. Collet, D. Charlet, and F. Bimbot, "A Correlation Metric for Speaker Tracking Using Anchor Models," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. 713-716, 2005.

[18] M. Collet, Y. Mami, D. Charlet, and F. Bimbot, "Probabilistic Anchor Models Approach for Speaker Verification," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '05)*, 2005.

[19] Y. Mami and D. Charlet, "Speaker Identification by Anchor Models with PCA/LDA Post-Processing," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '03)*, 2003.

[20] C. Ortego-Resa, I. Lopez-Moreno, D. Ramos, and J. Gonzalez-Rodriguez, "Anchor Model Fusion for Emotion Recognition in Speech," *Proc. Joint COST 2101 and 2102 Int'l Conf. Biometric ID Management and Multimodal Comm. (BioID-Multicomm '09)*, pp. 49-56, Sept. 2009.

[21] S. Steidl, *Automatic Classification of Emotion Related User States in Spontaneous Children's Speech*. Logos Verlag, 2009.

[22] S. Young, P. Woodland et, and W. Byrne, "HTK: Hidden Markov Model Toolkit V1.5," technical report, Cambridge Univ. Eng. Dept. of Speech Group and Entropic Research Laboratories Inc., http://htk.eng.cam.ac.uk/, 1993.

[23] A. Dempster, N. Laird, and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statistical Soc.*, vol. 39, pp. 1-38, 1997.

[24] P. Dumouchel, N. Dehak, Y. Attabi, R. Dehak, and N. Boufaden, "Cepstral and Long-Term Features for Emotion Recognition," *Proc. 10th Ann. Conf. Int'l Speech Comm. Assoc. (InterSpeech '09)*, pp. 344-347, 2009.

[25] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788-798, May 2011.

[26] D. Garcia-Romero and C.Y. Espy-Wilso, "Analysis of I-Vector Length Normalization in Speaker Recognition Systems," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '11)*, Aug. 2011.

[27] Hatch and A. Stolcke, "Generalized Linear Kernels for One-versus-All Classification: Application to Speaker Recognition," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '06)*, 2006.

[28] X. Zhao, Y. Dong, H. Yang, J. Zhao, and H. Wang, "SVM-Based Speaker Verification by Location in the Space of Reference Speakers," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '07)*, 2007.

[29] D. Charlet, M. Collet, and F. Bimbot, "VZ-Norm: An Extension of Z-Norm to the Multivariate Case for Anchor Model Based Speaker Verification," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '07)*, 2007.

[30] K.T. Abou-Moustafa, M. Cheriet, and C.Y. Suen, "Classification of Time-Series Data Using a Generative/Discriminative Hybrid," *Proc. Int'l Workshop Frontiers on Handwriting Recognition (IWFHR '04)*, pp. 51-56, Oct. 2004.

[31] L. Batista, E. Granger, and R. Sabourin, "A Multi-Classifier System for Off-Line Signature Verification Based on Dissimilarity Representation," *Proc. Ninth Int'l Workshop Multiple Classifier Systems (MCS '10)*, pp. 264-273, Apr. 2010.

[32] Z. Lei, Y. Yang, and Z. Wu, "An UBM-Based Reference Space for Speaker Recognition," *Proc. Int'l Conf. Pattern Recognition*, pp. 318-321, 2006.

[33] G.M. Weiss and F. Provost, "The Effect of Class Distribution on Classifier Learning," technical report, Dept. of Computer Science, Rutgers Univ., 2001.

[34] N.V. Chawla, K.W. Bowyer, L.O. Hall, and W.P. Kegelmeyer, "Smote: Synthetic Minority Over-Sampling Technique," *J. Artificial Intelligence Research*, vol. 16, pp. 321-357, 2002.

[35] R. Yan, Y. Liu, R. Jin, and A. Hauptmann, "On Predicting Rare Cases with SVM Ensembles in Scene Classification," *Proc. IEEE Int'l Conf. Acoustics, Speech, and Signal Processing (ICASSP '03)*, 2003.

[36] A. Rosenberg, "Classifying Skewed Data: Importance Weighting to Optimize Average Recall," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '12)*, 2012.

[37] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann Ian, and H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explorations*, vol. 11, no. 1, pp. 10-18, 2009.

[38] C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion Recognition Using a Hierarchical Binary Decision Tree Approach," *Speech Comm. Sensing Emotion and Affect—Facing Realism in Speech Processing*, vol. 53, nos. 9/10, pp. 1162-1171, Nov./Dec. 2011.
[39] M. Kockmann, L. Burget, and J. Cernocký, "Brno University of Technology System for Interspeech 2009 Emotion Challenge," *Proc. Conf. Int'l Speech Comm. Assoc. (InterSpeech '09)*, 2009.
[40] B. Schuller et al., "Recognising Realistic Emotions and Affect in Speech: State of the Art and Lessons Learned from the First Challenge," *Speech Comm.*, vol. 53, no. 9, pp. 1062-1087, 2011.

Yazid Attabi received the engineer degree in computer engineering in 1994 from the Université des Sciences et de la Technologie Houari Boumedine, Algeria, and the master's of science degree in software engineering in 2009 from École de Technologie Supérieure (ETS), Montreal. He is currently working toward the PhD degree at ETS, Montréal, and is also with Centre de recherche informatique de Montreal, Canada. His research interests include machine learning approaches applied to emotion recognition from speech.

Pierre Dumouchel received the BEng degree from the Université McGill and the MSc and PhD degrees from INRS-Télécommunications. He has more than 25 years of experience in the field of speech recognition, speaker recognition, and emotion detection. He is the chairman and professor at the Software Engineering and IT Department at École de Technologie Supérieure, Université du Québec, Canada. He is a member of the IEEE.