Original Article
International Journal of Fuzzy Logic and Intelligent Systems
Vol. 14, No. 4, December 2014, pp. 240-248
http://dx.doi.org/10.5391/IJFIS.2014.14.4.240

ISSN(Print) 1598-2645
ISSN(Online) 2093-744X

Text-independent Speaker Identification Using Soft Bag-of-Words Feature Representation

Shuangshuang Jiang[1], Hichem Frigui[1], and Aaron W. Calhoun[2]
[1]Multimedia Research Lab, CECS Dept., University of Louisville, Louisville, KY 40292, USA
[2]Pediatrics Dept., University of Louisville, Louisville, KY 40202, USA

Abstract

We present a robust speaker identification algorithm that uses novel features based on soft bag-of-word representation and a simple Naive Bayes classifier. The **b**ag-**o**f-**w**ords (BoW) based histogram feature descriptor is typically constructed by summarizing and identifying representative prototypes from low-level spectral features extracted from training data. In this paper, we define a generalization of the standard BoW. In particular, we define three types of BoW that are based on crisp voting, fuzzy memberships, and possibilistic memberships. We analyze our mapping with three common classifiers: Naive Bayes classifier (NB); K-nearest neighbor classifier (KNN); and support vector machines (SVM). The proposed algorithms are evaluated using large datasets that simulate medical crises. We show that the proposed soft bag-of-words feature representation approach achieves a significant improvement when compared to the state-of-art methods.

Keywords: Speaker identification, Clustering, Bag-of-Words (BoW) feature representation, Fuzzy membership, Possibilistic membership, Naive Bayes classifier

1. Introduction

The Simulation for Pediatric Assessment, Resuscitation, and Communication (SPARC) group within the Department of Pediatric Critical Care Medicine at the University of Louisville makes extensive use of simulation in training teams of nurses, medical students, residents, and attending physicians. These simulation sessions involve trained actors simulating family members in various crisis scenarios. Sessions involve 4 to as many as 9 people and last approximately 20 minutes to one hour. They are scheduled approximately twice per week and are recorded as video data. After each session, the physician/instructor must manually review and annotate the recording and then debrief the trainees on the session. The goal is to enhance the care of children and strengthen interdisciplinary and clinician-patient interactions [1].

The physician responsible for the simulation has recorded 100's of sessions, and has realized that the manual process of review and annotation is labor intensive and that retrieval of specific video segments (based on speaker or what was said) is not trivial. Using machine learning methods, we have developed a speaker segmentation and identification system that can provide the physician with automated and efficient methods to semantically index and retrieve specific segments from the large collections of simulation sessions.

Received: Nov. 2 2014
Revised : Dec. 4, 2014
Accepted: Dec. 11, 2014

Correspondence to: Corresponding author name (h.frigui@louisville.edu)

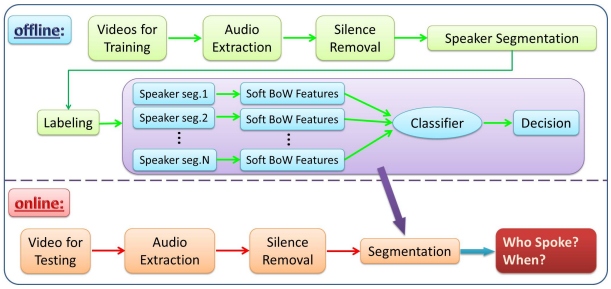


Figure 1. Overview of the proposed speaker segmentation and identification system

The architecture of this system is illustrated in Figure 1. It has two main components. The first one is for offline training, and the second one is for online testing. In the offline training, first audio streams are extracted from the training videos. Then, the divide-and-conquer ($DAC3$) based speaker segmentation method [2] is used to partition the speech sequence into homogeneous segments. Finally, a classifier is trained to discriminate between segments that correspond to different speakers. In the online testing, the input consists of an unlabeled video recording. First, the audio component is extracted and segmented. Then, each segment is labeled by the classifier. As a result, our system will identify "who spoke and when". In this paper, we focus on developing an efficient and accurate algorithm for speaker identification.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the proposed soft bag-of-words feature representation method. Section 4 discusses our experiments. Finally, concluding remarks are given in Section 5.

2. Related Works

2.1 Speaker Recognition

Speaker identification, classifying speech utterances into different speaker classes, and speaker verification, verifying a person's claimed identity from his/her voice, are generally referred to as speaker recognition [3]. Several features and classification methods have been proposed for this task. For instance, Mel frequency cepstral coefficients (MFCC) [4], which take into account how humans perceive the difference between sounds of different frequencies, is one of the most commonly used features [5, 6]. Perceptual linear predictions (PLP) [7] and linear prediction cepstral coefficients (LPCC) [8] are two other common features that rely on psychophysically based spectral transformations and linear prediction. These features may not always achieve good performance, especially in noisy environment, and many other features have been proposed. For instance, in [9], Wang proposed combing MFCC features and phase information for speaker identification. In [10], Li proposed an auditory based feature extraction algorithm by using a set of modules cochlear filter banks. In [11], Gabor filtering was applied to speech spectrum features, and nonnegative tensor factorization method was used to extract more robust features. Other feature representations can be found in [3, 5, 12].

Most of the above methods extract features from small overlapping windows. Thus, each speech segment can be represented by a large number of features. Moreover, since speech segments can have different durations, they will be represented by different number of features. To overcome this limitation, the above features are usually summarized by a small fixed number of representatives. For instance, Gaussian mixture model (GMM), with universal background model (UBM) for speaker adaptation [13], has been widely used for speaker recognition [5, 14]. In [14], GMM adaptation was applied to UBM to learn each speaker. Then, log-likelihood scores with nonlinear normalization were used for speaker discrimination. In [5], the GMM mean supervector, that represents the variable size segments with a fixed dimension by concatenating all adapted Gaussian mean vectors, was combined with a support vector machines (SVM) classifier. This approach was proven to be one of the most effective methods for speaker recognition.

Another alternative approach, called possibilistic histogram features (PHF) was proposed in [1, 15]. The PHF is inspired by the "bag of words" concept used in information retrieval. It identifies a fixed set of representative prototypes, and each audio segment is mapped to the closest prototype. The relative frequency of occurrence of each prototype is used as the feature vector of each audio segment. The PHF has been used with a KNN classifier and its performance was constrained. On one hand, a reduced vocabulary cannot represent all variations within the features. On the other hand, a larger vocabulary can improve the feature representation, but can also degrade the KNN classifier.

2.2 Feature Representation with Bag-of-Words (BoW)

The bag-of-words model has been widely used in various applications, such as document classification, computer vision, speech and speaker recognition, etc. In document classification, the feature is constructed based on the frequency of occurrence of each word [16]. Generally, there are two different models

to represent the document. One model uses a vector of binary attributes to indicate whether a word occurs or does not occur in the document. This representation can be modeled as a multivariate Bernoulli distribution. Another model takes the number of word occurrences into account, and represents the document by a sparse histogram of words frequencies. This representation can be modeled as a multinomial model. For both models, the Naive Bayes classifier is commonly used for classification.

In computer vision, a bag of *visual* words is a vector of frequency counts of a vocabulary of local image features. It has been used mainly in image/video scenes classification and retrieval [17, 18]. In [17], a "bag of key points" method was proposed based on vector quantization of affine invariant descriptors of image patches. Two different classifiers, Naive Bayes and SVM, were applied for semantic visual categories classification. Similarly, in [18], a set of viewpoint invariant region descriptors were extracted to search and localize all the occurrences of a given query object in a video. In this approach, a visual vocabulary was built through vector quantizing the descriptors into clusters. Using the standard indexing method used in text retrieval, the term frequency-inverse document frequency (TF-IDF) was computed and the cosine similarity was used for retrieval.

The BoW has also been used for the analysis of speech data. In [19], the high-frequency keywords (e.g. *you know*, *um*, *right*, etc.) were selected by computing the frequent, reflexive words and word pairs, and modeling them via word-based HMM models. Integrating this advantage of text-dependent modeling into the traditional GMM-based text-independent speaker recognition was shown to improve the performance. In [20], a bag-of-words (BoW)-style feature representation, which quantizes the observed direction of arrival (DOA) powers into discrete "word" samples, was developed to solve the speaker-clustering problem. In this approach, a time-varying probabilistic model was combined with the DOA information calculated from a microphone array to estimate the number and locations of the speakers.

2.3 BoW Feature Representation with Naive Bayes Classifier

Assume that we have a set of labeled speech segments $X = \{X^i\}$, C classes $[S_1, ..., S_j, ..., S_C]$, and representative vocabularies (i.e. codebook or cluster centers) $V = \{v_t\}$. Let $f_t(X^i)$ denotes the relative frequency of the occurrence of word v_t in segment X^i. To classify a new test sample, X^s, Bayes' rule is applied and the maximum a posteriori score is used for prediction:

$$P(S_j|X^s) \propto P(S_j)P(X^s|S_j) = P(S_j)\prod_{t=1}^{|V|} P(v_t|S_j)^{f_t(X^s)} \tag{1}$$

In (1), $P(S_j)$ is the a priori probability of class S_j, and the class-conditional probability $P(v_t|S_j)$ denotes the probability of word v_t occurring in class S_j and can be estimated using:

$$P(v_t|S_j) = \frac{\sum_{X^i \in S_j} f_t(X^i)}{\sum_{n=1}^{|V|}\sum_{X^i \in S_j} f_n(X^i)} \tag{2}$$

In order to avoid the zero probability estimation in (2), the Laplace smoothing is frequently used, and (2) can be replaced with:

$$P_{Lap}(v_t|S_j) = \frac{1 + \sum_{X^i \in S_j} f_t(X^i)}{|V| + \sum_{n=1}^{|V|}\sum_{X^i \in S_j} f_n(X^i)} \tag{3}$$

3. Soft BoW Audio Feature Representation

In this paper, we propose a generalization of the BoW feature representation. In addition to the standard binary voting, where each sample contributes to each keyword with a binary value (1 if the keyword is the closest one to the sample and 0 otherwise), we propose a generalization that uses soft voting. We discuss the advantages and disadvantages of each voting scheme. We also show that the soft BoW representations with a Naive Bayes classifier outperform existing methods for speaker identification.

3.1 Visual Vocabulary Construction

Assume that each speaker i has a training set of N^i low-level features, that is, $\mathbf{X}^i = \{\boldsymbol{x}_j^i | j = 1, ..., N^i\}$ where $\boldsymbol{x}_j^i \in \Re^D$ is a D dimensional feature vector extracted from the j^{th} segment of the i^{th} speaker.

The first step consists of summarizing each $\mathbf{X}^i$ by a set of representative prototypes $\{\boldsymbol{p}_1^i, \boldsymbol{p}_2^i, ..., \boldsymbol{p}_{K^i}^i\}$. This quantization step is achieved by partitioning $\mathbf{X}^i$ into K^i clusters and letting $\boldsymbol{p}_k^i$ be the centroid of the k^{th} partition. Any clustering algorithm can be used for this task. In this paper, we report the results using the Fuzzy C-means (FCM) [21] algorithm. The FCM partitions the N^i samples into K^i clusters by minimizing the

sum of within-cluster distances, i.e.,

$$\mathbf{J}(\mathbf{U};\mathbf{X}^i)=\sum_{j=1}^{N^i}\sum_{t=1}^{K^i}\mu_{tj}^m d^2(\boldsymbol{x}_j^i,\boldsymbol{p}_t^i). \tag{4}$$

In (4), d refers to the Euclidean distance between $\boldsymbol{x}_j^i$ and $\boldsymbol{p}_t^i$, and $U=[\mu_{tj}]$ represents the membership of feature vector $\boldsymbol{x}_j^i$ in cluster t [22] and satisfies the constraints:

$$\begin{cases} \mu_{tj}\in[0,1] \\ \sum_{t=1}^{K^i}\mu_{tj}=1 \end{cases} \tag{5}$$

Each prototype, p_k, is a representative of cluster c_k that summarizes a group of similar speech segments. Let σ_k be the variance of all features x_j assigned to cluster c_k. After clustering, the K^i prototypes obtained by partitioning the data of speaker i, X^i, are all combined to form a dictionary or a codebook with $K=\sum_{i=1}^{N_{sp}}K^i$ words, where N_{sp} is the number of speakers.

Instead of using the original feature space X, the **b**ag-**o**f-**w**ords based **h**istogram **f**eature **d**escriptor (BoW-HFD) approach maps it to a new space H characterized by the K clusters that capture the characteristics of the training data. Formally, this mapping is defined as

$$\begin{aligned} M: \quad & \boldsymbol{x}\longrightarrow\boldsymbol{H} \\ M(\boldsymbol{x}_j) \quad = \quad & \boldsymbol{h}_j=[f_1(\boldsymbol{x}_j),...,f_K(\boldsymbol{x}_j)] \end{aligned} \tag{6}$$

In (6), $f_i(\boldsymbol{x}_j)\in[0,1]$ is a measure of belongingness of feature $\boldsymbol{x}_j$ to cluster i represented by prototype $\boldsymbol{p}_i$. This measure could be *crisp*, *fuzzy*, or *possibilistic* [22]. These different mappings are described in the following subsections.

3.2 Crisp Mapping

In crisp mapping, each feature vector x_j is assigned a binary membership value to each "word" i based on the distance between them. This mapping considers only the closest word (i.e. prototype) to word i and is defined as:

$$f_i^c(\boldsymbol{x}_j)=\begin{cases} 1 & \text{if } i=\underset{k}{argmin}\parallel x_j-p_k\parallel^2 \\ 0 & otherwise \end{cases} \tag{7}$$

This mapping is used in the standard BoW approach [17] and considers only the closest word. Thus, it is reasonable if $\boldsymbol{x_j}$ is close to one word and far from the other words. However, if $\boldsymbol{x_j}$ is close to multiple words (i.e., $\boldsymbol{x}_j$ is located close to the clusters' boundaries), then, crisp mapping will not preserve this information.

3.3 Fuzzy Mapping

Instead of using binary voting (as in eq. (7)), fuzzy mapping uses soft labels to allow for partial or gradual membership values. This type of labeling offers a richer representation of belongingness and can handle uncertain cases. In particular, a sample x_j votes to each word i in the codebook with a membership degree $f_i^f(x_j)$ such that:

$$\begin{cases} f_i^f(x_j)\in[0,1] \\ \sum_{i=1}^{|K|}f_i^f(x_j)=1 \end{cases} \tag{8}$$

Many clustering algorithms use this type of labels to obtain a fuzzy partition. In the proposed fuzzy BoW (F-BoW) approach, we use the memberships derived within the Fuzzy C-Means (FCM) [21] algorithm, i.e.,

$$f_i^f(\boldsymbol{x}_j)=\frac{1}{\sum_{t=1}^{|K|}(\frac{D_{ji}}{D_{jt}})^{\frac{2}{m-1}}} \tag{9}$$

where $m\in(1,\infty)$ is a constant that controls the degree of fuzziness. In (9), D_{jt} is the distance between feature vector x_j and the "word" summarizing cluster t. To take into account the shape of the clusters, we use

$$D_{jt}=\sum_{k=1}^{M}\frac{||x_{jk}-p_{tk}||^2}{\sigma_{tk}^2} \tag{10}$$

where σ_{tk}^2 is the variance of feature k of cluster t and M is the dimensionality of the feature space.

3.4 Possibilistic Mapping

The fuzzy membership in (9) is a relative number that depends on the distance of $\boldsymbol{x_j}$ to all prototypes. Thus, it does not distinguish between samples that are equally close to multiple prototypes and samples that are equally far from all prototypes.

An alternative approach to generate soft labels is based on possibility theory [22]. Possibilistic labeling relaxes the constraint in (8) that the memberships across all words must sum to one. It assigns "typicality" values, $f_i^p(\boldsymbol{x}_j)$, that do not consider the *relative* position of the point to all clusters. As a result, if x_j is a noise point, then $\sum_{t=1}^{|K|}f_t^p(\boldsymbol{x}_j)\ll 1$, and if x_j is typical of more than one cluster, we can have $\sum_{t=1}^{|K|}f_t^p(\boldsymbol{x}_j)>1$. Many robust partitional clustering algorithms [23, 24] use this type of

labeling in each iteration. In this paper, we use the membership function derived within the Possibilistic C-Means [22], i.e.,

$$f_i^p(\boldsymbol{x}_j) = \frac{1}{1 + (\frac{D_{ji}}{\eta_j})^{\frac{2}{m-1}}} \tag{11}$$

In (11), η_j is a cluster-dependent resolution/scale parameter [22] and $m \in (1, \infty)$.

Robust statistical estimators, such as M-estimators and W-estimators [25], use this type of memberships to reduce the effect of noise and outliers.

4. Experimental Results and Discussion

4.1 Data Collection

Multiple data sets are used to validate and compare our proposed soft BoW-based audio feature representation with Naive Bayes classifier for speaker identification. In particular, we use 15 medical simulations videos. We only use the audio information for speaker identification as it contains most conversation information. This is because the video resolution is low and has no additional information (people are sitting with little movement and just talking). As shown in Table 1, each simulation has four speakers (patient, patient's friend, doctor, and nurse). Videos are recorded in different rooms, and have different quality with different levels of background noise and frequent interruptions. The content of the conversations involve similar topics. For all experiments reported in this paper, we use a k-fold cross validation with k = 5. That is, for each video, we keep 80% of data for training and use the remaining 20% for testing. We repeat this process 5 times by testing different subsets and report the average of the 5 numbers.

Table 1. Data collections used to validate the proposed speaker identification approach

Videos	Lengths	# of Spkers	# of Seg.	Avg. Len. (in sec)
Med1	6m35s	4	202	1.96
Med2	7m13s	4	324	1.34
Med3	10m20s	4	304	2.04
Med4	18m02s	4	526	2.06
Med5	9m40s	4	295	1.97
Med6	7m22s	4	218	2.03
Med7	10m16s	4	303	2.03
Med8	4m33s	4	134	2.04
Med9	6m54s	4	215	1.93
Med10	5m32s	4	169	1.96
Med11	6m43s	4	206	1.96
Med12	7m45s	4	236	1.97
Med13	12m1s	4	361	2.00
Med14	5m34s	4	165	2.02
Med15	7m55s	4	230	2.07

4.2 Preprocessing

First, the audio component is extracted from the video. All speech files are single-channel data sampled at 22.05kHz frequency. Then, since silence segments provide no information about the speakers and actually may reduce the correct speaker identification rate, each audio stream is processed to identify and remove silence segments. We use a trainable support vector machines (SVM) classifier [26] based on 3 low-level audio features (short-time energy, zero crossing rate, and spectral centroid) to discriminate between speech and nonspeech audio.

The remaining speech segments are decomposed into small frames using a 25ms analysis window with 10ms overlap. From each window, we extract MFCC [4], PLP [7], LPCC [8], and GFCC [11] features. For GFCC, instead of using tensor decomposition as proposed in [11], we simply average all Gabor filtered spectrum features along the scales and phases to reduce the computational complexity.

For each extracted feature, we use the BIC algorithm [27] to identify changing points within the audio stream and partition it into homogeneous segments. Each segment will then be processed by our proposed algorithm to identify the speaker. Table 1 displays the number of speech segments identified by BIC for each video. The average length of all segments is short due to the frequent interruption during the conversation. We should note here that each video segment is processed independently since it involves different speakers. The reported results are the average over the 15 datasets.

4.3 Evaluation and Discussion

First, the same low-level features used to segment the audio stream (MFCC, PLP, LPCC, and GFCC) are also used for speaker identification. Next, bag of words features (C-BoW, F-BoW, and P-BoW) are constructed for each feature as described in Section 3. The initial number of prototypes is set to 100 per speaker, i.e. $K_i = 100$, resulting in a codebook with $K = 400$ words.

For each low-level feature, we evaluate the performance of the proposed mapping using 3 different classifiers: K-NN, Naive Bayes, and SVM [28]. K-NN has the advantage of incor-

Table 2. Classification rate of the K-NN classifier using the proposed soft bag of words representation of MFCC features and various distance measures

Dist. Type	C-BoW	F-BoW	P-BoW
Eu	0.756	0.775	0.77
CS	0.742	0.766	0.758
HI	0.752	0.765	0.752
JS	0.545	0.571	0.552
KS	0.792	0.809	0.799
KL	0.555	0.573	0.564
MD	0.793	0.808	0.803
DD	0.715	0.734	0.739
CD	**0.794**	**0.816**	**0.806**

porating various distance measures. Naive Bayes is a simple and efficient classifier that proved to be effective is classification problems that use the bag-of-word feature representation [16, 17]. SVM is one of the most commonly used classifiers. For each classifier, we compare the performance of the 3 proposed feature mapping methods.

For the K-NN classifier, first we experiment with several measures, as discussed in [29], to compute the dissimilarity between two histogram features (i.e. vectors mapped to histograms using bag of words representation). In particular, we use chi-square statistics (CS), histogram intersection (HI), Jensen-Shannon divergence (JS), Kolmogorov-Smirnov distance (KS), Kullback-Leibler divergence (KL), match distance (MD), diffusion distance (DD), and cosine distance (CD). The speaker recognition accuracies, averaged over the 15 datasets, using the MFCC features with a K-NN classifier (K=7), are displayed in Table 2. As it can be seen, the cosine distance has the best performance for the crisp, fuzzy, and possibilistic bag of words representations. Similar results are obtained for the PLP, LPCC, and GFCC features. Thus, for the remaining experiments, the cosine distance will be used within the K-NN classifier to compare it to other classifiers.

In Figure 2, we compare the speaker identification accuracy of the proposed soft BoW feature mappings using MFCC features with the K-NN, NB, and SVM classifiers. First, we notice that the NB classifier outperforms the K-NN and SVM classifiers for the crisp, fuzzy, and possibilistic cases. Second, on average, the soft (fuzzy and possibilistic) feature mappings outperform the crisp mapping. Similar results were obtained for the PLP, LPCC, and GFCC features.

In a second experiment, we compare our methods to 3 existing speaker identification algorithms: GMM-UBM [14], GMM

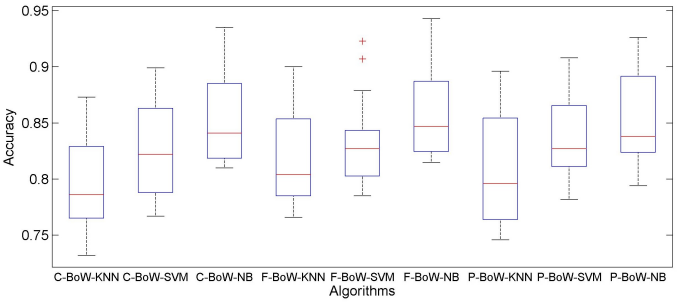


Figure 2. Performance of the crisp, fuzzy, and possibilistic BoW using MFCC features with the KNN, SVM, and NB classifiers

mean supervector [5] with K-NN classifier (SV-KNN) and SVM classifier (SV-SVM), and PHF [1] with KNN classifier (PHF-KNN). For the GMM-UBM-based speaker identification, the UBM is estimated using all training features, while GMM adaptation is applied to the UBM to get each training speaker model. Then, log-likelihood scores are used for the classification. For both the GMM-UBM and GMM mean supervector methods, we experiment with several values for the number of Gaussian components and set this parameter to 10. The results are reported in Figure 3. As it can be seen, for all 4 features, the proposed soft feature mapping coupled with the NB classifier outperform the state of the art methods.

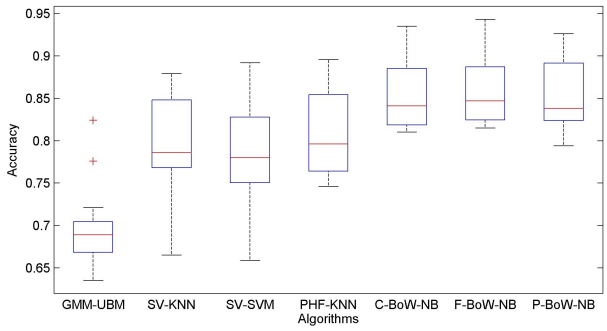
5. Conclusions

We proposed a soft feature mapping approach for speaker identification. Our approach uses bag-of-words model to extract robust histogram descriptors from low-level spectral features. We formulated three kinds of feature mapping methods using crisp, fuzzy, and possibilistic membership functions.

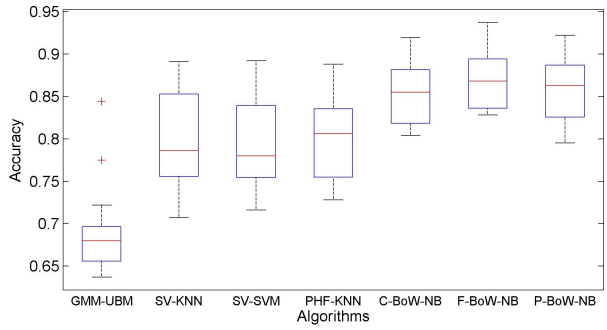
Using 15 datasets, we showed that the Naive Bayes is the best classifier to be used with our soft mapping. We also showed that the proposed approach outperforms commonly used methods.

The Proposed mappings provide more accurate speaker identification results. This allows the physicians to analyze the simulation sessions more easily and to identify and retrieve speech segments for a given speaker more accurately.

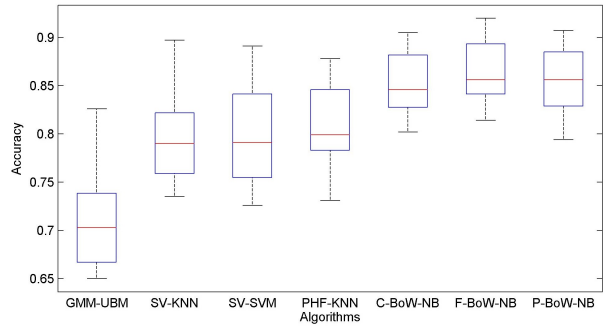
In our future work, we will focus on the fusion of multiple histograms that map different features (e.g. MFCC, PLP, LPCC, and GFCC) and applying ensemble learning approaches to further improve the accuracy of the speaker identification.



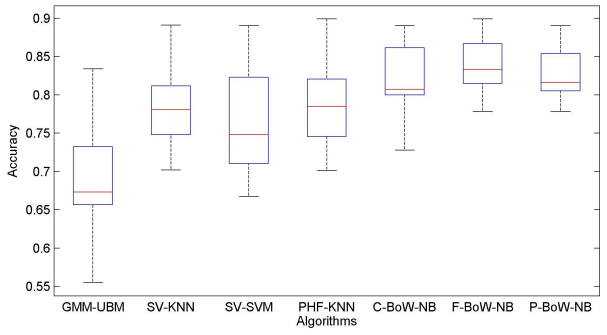
(a) MFCC feature



(b) PLP feature



(c) LPCC feature



(d) GFCC feature

Figure 3. Comparison of the classification accuracy of the proposed soft BoW feature mappings using the NB classifier with GMM-UBM, GMM mean supervector with K-NN (SV-KNN) and SVM (SV-SVM), and PHF with KNN.

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

References

[1] S. Jiang, H. Frigui, and A. Calhoun, "Semantic indexing of video simulations for enhancing medical care during crises," *11th International Conference on Machine Learning and Applications (ICMLA)*, pp. 520–525, 2012.

[2] S. Cheng, H. Wang, and H. Fu, "Bic-based speaker segmentation using divide-and-conquer strategies with application to speaker diarization," in *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 18 of *1*, pp. 141–157, 2010.

[3] J. Campbell, "Speaker recognition: a tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437–1462, 1997.

[4] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 155–210, 1937.

[5] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, pp. 12–40, Jan. 2010.

[6] X. Miro, S. Bozonnet, N. Evans, C. Fredouille, and G. F. abd O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[7] H. Hermansky, "Perceptual linear predictive (plp) analysis of speech," *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[8] X. Huang, A. Acero, and H. Hon, "Spoken language processing: a guide to theory, algorithm, and system development," *Prentice-Hall, New Jersey*, 2001.

[9] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining mfcc and phase information in noisy environment," in *ICASSP*, 2010.

[10] Q. Li and Y. Huang, "Robust speaker identification using an auditory-based feature," in *ICASSP*, 2010.

[11] Q. Wu, L. Zhang, and G. Shi, "Robust feature extraction for speaker recognition based on constrained nonnegative tensor," *J. Comput. Sci. Technol.*, vol. 25, no. 4, pp. 745–754, 2010.

[12] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, pp. 1557–1565, September 2006.

[13] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[14] R. Zheng and B. X. S. Zhang, "Text-independent speaker identification using gmm-ubm and frame level likelihood normalization," in *International Symposium on Chinese Spoken Language Processing*, pp. 289–292, Dec. 2004.

[15] H. Balti and H. Frigui, "Feature mapping and fusion for music genre classification," *ICMLA*, pp. 306–310, 2012.

[16] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," *AAAI-98 workshop on learning for text categorization*, pp. 41–48, 1998.

[17] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," *Proc. of ECCV International Workshop on Statistical Learning in Computer Vision*, 2004.

[18] J. Sivic, "Efficient visual search of videos cast as text retrieval," *IEEE Trans. on pattern analysis and machine intelligence*, vol. 31, no. 4, pp. 591–605, 2009.

[19] K. B. B. Peskin, "Text-constrained speaker recognition on a text-independent task," *In ODYS-2004*, pp. 129–134, 2004.

[20] K. Ishiguro, T. Yamada, S. Araki, T. Nakatani, and H. Sawada, "Probabilistic speaker diarization with bag-of-words representations of speaker angle information," *IEEE Trans. on audio, speech, and language processing*, vol. 20, no. 2, pp. 447–460, 2012.

[21] J. C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*. Norwell, MA, USA: Kluwer Academic Publishers, 1981.

[22] H. Frigui, "Membershipmap: Data transformation based on granulation and fuzzy membership aggregation," *IEEE Trans. Fuzzy Systems*, vol. 14, pp. 885–896, Dec 2006.

[23] R. Duda, P. Hart, and D. Stork, "Pattern classification, 2nd edition," *New York: John Wiley & Sons*, 2000.

[24] L. Kaufman and P. Rousseeuw, "Finding groups in data: An introduction to cluster analysis," *New York: Wiley*, 1990.

[25] F. R. Hampel, E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel, "Robust statistics the approach based on influence functions," *New York: Wiley*, 1986.

[26] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, first ed., 2000.

[27] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *in Proc. DARPA Broadcast News Transcription Understanding Workshop*, (Landsdowne, VA), 1998.

[28] C. M. Bishop, "Pattern recognition and machine learning," *Springer*, 2006.

[29] Y. Rubner, C. Tomasi, and L. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.