

A New Strategy of Direct Access for Speaker Identification System Based on Classification

Hery Heryanto^{*1}, Saiful Akbar², Benhard Sitohang³

Data and Software Engineering Research Group
School of Electrical Engineering and Informatics, Bandung Institute of Technology
Jl. Ganesa No 10, Bandung 40132

*Corresponding author, email: h3ry.heryanto@gmail.com¹, saiful@informatika.org²,
benhard@stei.itb.ac.id³

Abstract

In this paper, we present a new direct access strategy for speaker identification system. DAMClass is a method for direct access strategy that speeds up the identification process without decreasing the identification rate drastically. This proposed method uses speaker classification strategy based on human voice's original characteristics, such as pitch, flatness, brightness, and roll off. DAMClass decomposes available dataset into smaller sub-datasets in the form of classes or buckets based on the similarity of speaker's original characteristics. DAMClass builds speaker dataset index based on range-based indexing of direct access facility and uses nearest neighbor search, range-based searching and multiclass-SVM mapping as its access method. Experiments show that the direct access strategy with multiclass-SVM algorithm outperforms the indexing accuracy of range-based indexing and nearest neighbor for one to nine percent. DAMClass is shown to speed up the identification process 16 times faster than sequential access method with 91.05% indexing accuracy.

Keywords: direct access, speaker identification, MFCC, multiclass classification, speaker indexing

Copyright © 2015 Universitas Ahmad Dahlan. All rights reserved.

1. Introduction

Direct Access Method (DAM) is a data access method for identifying an object based on the original characteristics of the object. For example, we can identify a speaker by listening to parts of their speech. A speech contains the speaker's original characteristics which is unique for each speaker. There are two important processes in DAM, 1) original characteristics extraction to facilitate the direct access, also known as feature extraction process and 2) direct access method that access the object based on the original characteristics of the object [1].

In voice biometrics, speaker identification is different than speaker verification. Speaker identification is a process of comparing an impostor's signal with a number of speaker models in a dataset. The comparison performed is 1:n, whereas in speaker verification the comparison 1:1, that is the impostor's signal compared to the speaker's claimed identity. In this paper, we focus on the data access method. Since there are n comparison performed, speaker model's access time is another challenge for a speaker identification system. A speaker identification baseline system requires a relatively long time for identifying a speaker in a number of speaker models in the dataset. In [2], Heryanto et.al shows that for 1,000 speaker models, we will need 58 seconds to identify an identity of the speaker. Every increase of 100 speaker models, it will take 5 to 6 seconds longer than the previous speaker models number.

There are several related work on speaker indexing to speed up speaker identification process by avoiding the 1:n identification process. In [3], Kwon proposed an indexing method based on unsupervised speaker indexing to identify speakers during a talk that consist of two and four people. The index is built by the speaker change detection and Sample Speaker Models (SSM). Kwon successfully achieved 92.5% indexing accuracy for two people talks and 89.6% for four people talks. SSM outperforms Universal Background Model method for more than 20%. In Kwon's experimental result, the number of speaker models used was 100 speaker models and indexing accuracy achieved was 87.2% [3].

Other work by Schmidt et al [4] uses Local Sensitive Hashing (LSH) and fast nearest neighbor search algorithm for speaker indexing. Schmidt proposed an indexing method using i-

Vector. LSH is commonly used in audio signal or music classification. LSH produces an efficient music retrieval method. Schmidt's proposed method performs vector factor analysis from the speaker modeling process. LSH algorithm generates i-Vector from the Gaussian Mixture Model (GMM) super vector. GMM is a popular technique for speaker modeling and claimed to be the most accurate modeling technique by previous researches. LSH method with i-Vectors produces sufficiently high indexing accuracy at 93.8% with the access time of 16 times faster than the linear or sequential search. LSH itself has a major challenge that is how to determine the vector that represents the speakers. In the Schmidt's paper, it is unclear what kind of vector representation was built. We have difficulties in determining the vector representation of the existing speakers. Based on our exploration results, Mel-Frequency Cepstral Coefficient (MFCC) feature does not have a structured pattern that represents the speaker's original characteristics. In addition, Indrawan et al. proposed in [5, 6] a direct access strategy for fingerprint identification system to improve the performance of the identification process. Indrawan's model uses a modified hash function for retrieving a candidate list based on fingerprint's local and global features. We found that it is difficult to apply this strategy to audio data because of characteristic differences between image and audio data. Audio data, especially speech signals, is very difficult to be visualized on a speaker identification system, so it is hard to find a pattern that represents a speaker.

Based on the three previous works, we propose a new direct access strategy for improving the speed of the identification process. The first challenge is how to determine the original characteristics of the speaker other than MFCC to be a direct access facility. The second challenge is how to determine the speaker indexing strategy that can speed up the identification process without drastically decreasing the identification rate.

2. Research Method

Direct Access Method based on Classification (DAMClass) is a strategy for speaker identification's data access method. This proposed strategy uses classification technique as the basis for speaker indexing. The main objective of this strategy is to decompose the dataset into as small as possible sub-datasets while maintaining the accuracy of the identification process. The smaller sub-dataset narrows search space and thus speeds up the data identification process. The data that is used as a reference is the speaker's data, in the form of feature vector. The vector represents the speaker's speech signal. Our proposed strategy maps the speech signal based on the original characteristics using the direct access method.

Speech signal is a non-stationary signal, it needs to be decomposed into smaller frames with a duration of 20-30ms in order to generate "quasi" stationary signal. The most popular feature in a speaker identification system is MFCC. MFCC is a feature that is obtained from the spectrum of the speech signal in the form of double values that has 12 + 1 dimensions. MFCC is a replication of the human auditory system and gives the highest identification rate than other features [7-9]. The first dimension of MFCC is the energy coefficient of the speech signals, while the 12 other coefficients are the values of MFCC. The identification process cannot directly be based on the value of MFCC because the MFCC value is unstructured, thus requiring a statistical approach for obtaining the speaker model. The baseline system usually uses GMM algorithm for speaker modeling because this technique produces a higher identification rate compared to other algorithms such as Vector Quantization (VQ) [10-11].

For the same reason as in [10, 11], DAMClass also cannot use MFCC for speaker indexing. DAMClass uses some original characteristics from the speech signal that are pitch, flatness, brightness, and roll-off. These features are analyzed and selected from 78 audio features with several audio feature extraction toolboxes, including: MIR ToolBox [12], Audio Feature Extraction [13], and several other applications. The original characteristics or direct access facility selection is based on biometric system parameter criteria, such as: universal, distinctive, and permanent [14].

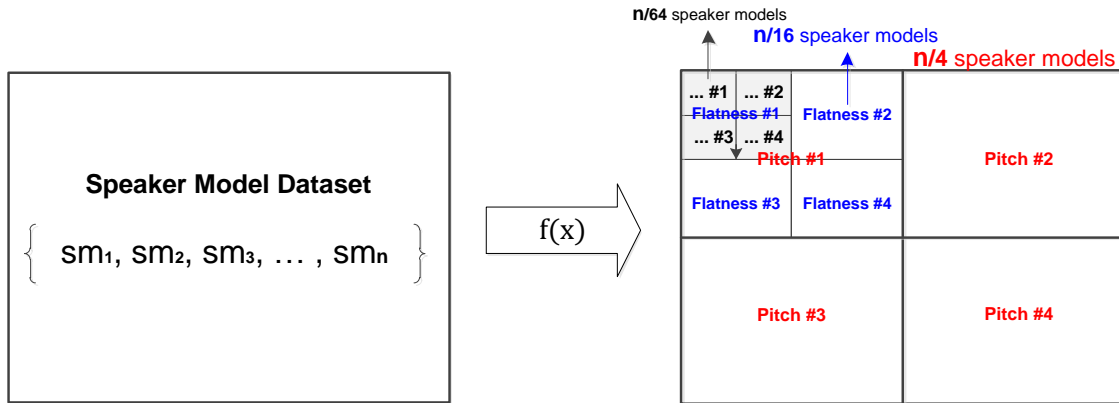


Figure 1. Dataset Decomposition using DAMClass

Figure 1 shows an illustration of decomposition process of the n data in a dataset into smaller sub datasets based on the original characteristics of the speech signal. Function $f(x)$ maps the speech signal from the dataset into a specific sub dataset. This function splits the search space into 4 smaller partitions based on the speech signal's pitch. The function uses range-based mapping or indexing. Later pitch subdataset is decomposed into 4 smaller search space based on flatness and so on. We set 4 classes for each layer and layer using a direct access facility or original characteristic. Equation 1 is a mathematical model of the dataset decomposition with DAMClass.

$$D = \cup D_i : i \in I \quad (1)$$

where D is the speaker dataset and $D_i : i \in I$ is the sub datasets resulting from DAMClass decomposition process. DAMClass then one sub dataset as a candidate list by performing queries based on inputted query point. For example, in Figure 1, there are 3 original characteristics for retrieving the candidate list, i.e.: pitch, flatness, and flatness. Suppose the inputted query point (1, 2, 4), then the retrieving process of the candidate list can be described with relational algebra as follows:

$$Q1 = \pi_{filename} \sigma_{pitch=1} dataset \quad (\text{step 1})$$

$$Q2 = \pi_{filename} \sigma_{flatness=2} Q1 \quad (\text{step 2})$$

$$Q3 = \pi_{filename} \sigma_{flatness=4} Q2 \quad (\text{step 3})$$

$Q3$ is the resulting candidate list which contains several speaker models that will be matched with the impostor signal. Speaker matching is performed using Expectation Maximization algorithm. A speaker model with the highest similarity score is then chosen as the speaker identity of the impostor signal. Figure 2 describes the flow of dataset decomposition using DAMClass. Each layer uses one direct access facility and assumed there are four classes in each layer. A mathematical model that represents the data access method in DAMClass is given in Equation 2.

$$Q_n = \pi_{filename} \sigma_{index_n=daf_n} Q_{n-1} \quad (2)$$

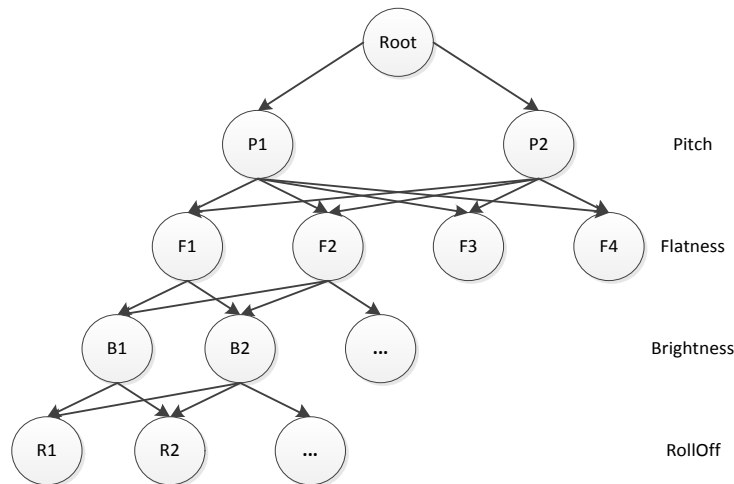


Figure 2. DAMClass: Speaker Dataset Decomposition Strategy

DAMClass speeds up the identification process of the classification strategy based on the original characteristics of the speaker models. The new access time when using DAMClass strategy is given in equation 3.

$$T' n = \frac{T n}{c^k} + \alpha k \tag{3}$$

where T is the old access time, n is the number of speaker models, c is the number of classes in each layer (the number of classes in each layer is assumed to be equal), k is the number of layers, and α is the time of speaker classification for each class (the value of α is less than 1 second, normally 0.2 second).

In this paper, DAMClass uses several algorithms for mapping a speech signal into available classes or buckets. The first algorithm that we use is Nearest Neighbor (NN) Classification with 3 distance metrics namely Euclidean, Manhattan, and Mahalanobis, illustrated in Figure 3. The second algorithm that we use is Range-Based Indexing (RB) and the last one a multiclass SVM Mapping (MSVM).

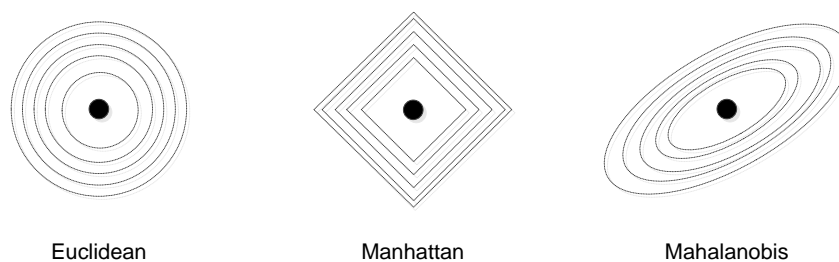


Figure 3. Distance Metrics for Nearest Neighbor Method in DAMClass

Algorithm 1 and 2 are examples of the proposed algorithm which is used for speaker model mapping based on the original characteristics or direct access facility. Algorithm 1 is a speaker model mapping based on the strategy of Normalized Range Based Indexing (NRB). This strategy maintains the balance of the speaker models number in each class so the access time of speaker identification process is evenly distributed among each class. NRB rebuild the indexes when the number of speaker models in each class is unbalance. It covers the weakness of RB Indexing which uses fixed lower and upper bound.

Algorithm 1: Normalized Range-Based Indexing

```

1: [min max num] ← getRange(dataset,direct_access_facility[1..l])
2: for i ← 1,...,l do
    min[i,1] ← min[l]
    for j ← 1,...,c do
        max[i,j] ← getValue(int(num*i/c))
        if i < c then
            min[i,j+1] ← max[i,j] + 0.0001
        end if
    end for
end for
3: setClass(dataset, {(min1,max1), ..., (minc,maxc)})

```

Algorithm 2 maps the speaker model based on multiclass SVM algorithm, where the mapping process is based on the value of each direct access facility and a feature vector of the speech signal itself. This algorithm recapitalizes all speaker models for each speaker and looks for speaker model classes with the highest frequency based on the speaker's identity. The purpose of this algorithm is to map the speaker models that are deviated from the characteristic of the speaker itself. Multiclass SVM uses some existing kernel, including: Linear Kernel, RBF (Gaussian) kernel and polynomial kernel.

Algorithm 2: Multiclass SVM Mapping

```

1: [min max num] ← getRange(dataset,direct_access_facility[1..l])
2: for i ← 1,...,l do
    min[i,1] ← min[l]
    for j ← 1,...,c do
        max[i,j] ← getValue(int(num*i/c))
        if i < c then
            min[i,j+1] ← max[i,j] + 0.0001
        end if
    end for
end for
4: setClass(impostor_value, {(min1,max1), ..., (minc,maxc)})
5: trainset[] ← getMode(speakerid,direct_access_facility[1..l])
6: msvmclassify(trainset[],kernel_type)
7: svmMap ← mapSpeakerModel(dataset,msvm_model)

```

After the index was built, the candidate list search process can be performed with various access methods. The first option is to use kNN Search, which is the most commonly used algorithm for searching a query point. The input of this algorithm is in the form of a direct access facility vector, ex: impostor ← (pitch_value, flatness_value, brightness_value, rolloff_value) ← (118.0897, 0.3651, 0.5430, 1269.0000). This algorithm is run in phases. In each phase, it searches for k-closest point by calculating the distance between impostor's vector and speaker model's vector. The second option is to do a range query from the query point that represents the impostor signal with some lower and upper bound, which obtained from index building process. This can be done with both NRB Indexing and multiclass SVM Mapping. In NRB Indexing, range query is performed on each layer with direct access facility parameter values prevailing in that layer. In Multi-Class SVM mapping, range query is performed on the candidate list using direct access facility vector in each layer. The query process have one value when the speaker identity of the impostor signal is contained in candidate list, otherwise query process returned zero value. Indexing accuracy is calculated by dividing the number of one valued queries by the number of executed queries in a batch of processes. The mathematical model is given in equation 4 and 5.

$$f(x) = \begin{cases} 1 & \text{if } \pi_{COUNT(filename)\sigma_{speakerid=speakerid_x}} list > 0 \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$Acc = \frac{1}{n} \sum_{i=1}^n f(x_i) \quad (5)$$

where $f(x)$ is the query process for an impostor signal. When the identity of the impostor signal is found in the list then the query's value is one otherwise it is zero. Acc is indexing accuracy that is calculated from the number of queries that have one value divided by n queries performed in the experiment.

3. Experiment Results and Discussion

In this experiment, we build our own dataset and Hyke dataset as a comparison in validating our proposed models. We have collected 142 speakers, consist of 97 males and 45 females. We use Bahasa as a spoken language in our dataset. Each data's utterance duration starts from one second to 30 seconds. The speaker speech was recorded with a headset and each speaker asked to pronounce 16 pieces of text which contains a combination of numbers, phrases, sentences and paragraphs. The speech was recorded in a hall room sized 20 x 30 meters. Some background noises that recorded are the sound of vehicles, people's conversations, and the sound of chair movements or footsteps. Meanwhile, the Hyke dataset that we use is collected by Microsoft Research India with English as a spoken language [15].

In our experiment, we use our speaker identification system framework that we proposed in [2] as the baseline system. We add our experiment modules in the existing framework. For the identification process, we use speaker identification system that was built by Dijk in Eindhoven University of Technology [16].

3.1. DAMClass using Nearest Neighbor Search

The first access method that we test is Nearest Neighbor because this method is most commonly used for a query point. We use three distance metrics, namely: Euclidean, Manhattan, and Mahalanobis. First, we implement this method on the Bahasa dataset with 2,259 speaker models. The experiment result is given in Figure 4. In this dataset, Mahalanobis distance produces the highest indexing accuracy compared to the two other distances. The highest indexing accuracy is about 96.80% with 200 speaker models in the candidate list.

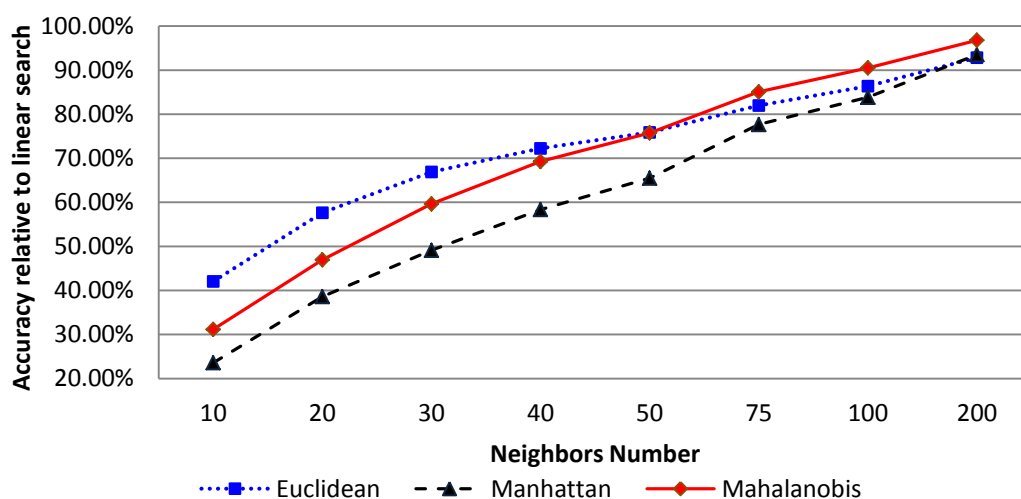


Figure 4. Nearest Neighbor Indexing in Bahasa Dataset

In this experiment session, we use Digit dataset which is a subset of Bahasa dataset, and Hyke dataset. The number of speaker models in the Digit dataset is 710 with 142 speakers. Hyke dataset has 415 speaker models with 83 speakers. The experiment results showed that the Mahalanobis distance produces the highest indexing accuracy about 100% when the number of speaker models in the candidate list is 200 and the dataset is Hyke dataset (access time is two times faster than sequential search). The indexing accuracy of Hyke dataset showed an inconsistent pattern because there is a lot of noisy speech signal in Hyke dataset. The experiment result is given in Table 1. The ratio is smaller than the ratio in the previous experiment session's result that has far more number of speaker models.

Table 1. Nearest Neighbor Indexing in Bahasa and Hyke Dataset

Dataset	Distance	Speaker Models Number in Candidate list							
		10	20	30	40	50	75	100	200
BAHASA	Euclidean	42.05%	57.64%	66.92%	72.25%	75.89%	81.97%	86.37%	92.85%
	Manhattan	23.58%	38.59%	49.11%	58.35%	65.50%	77.66%	83.84%	93.52%
	Mahalanobis	31.17%	46.94%	59.64%	69.27%	75.75%	85.08%	90.50%	96.80%
HYKE	Euclidean	19.26%	29.47%	37.44%	46.14%	53.38%	68.36%	80.68%	99.03%
	Manhattan	19.92%	30.43%	37.68%	46.38%	55.31%	69.57%	81.16%	99.03%
	Mahalanobis	20.05%	26.57%	32.37%	38.41%	41.55%	56.04%	67.39%	100%

3.2. Normalized Range Based Indexing

Based on the DAMClass strategy in Section 2, the second experiment uses the Normalized Range Based Indexing (NRB) on Bahasa and Hyke dataset. Experiment starts by determining the range of each class in the direct access facility layer which is pitch, flatness, brightness, and roll off. NRB will break the layer into 4 balanced classes. In initial experiment, the number of speaker models in the classes is unbalanced. This causes the access time to be very diverse so it is difficult to determine the actual access time. Table 2 shows the result of experiment that uses NRB in Bahasa dataset compared to Range-Based strategy.

Table 2. NRB Indexing in Bahasa dataset

Direct Access Strategy	Direct Access Facility							
	Pitch		P & Flatness		PF & Brightness		PFB & Roll Off	
	Accuracy	C. Num	Accuracy	C. Num	Accuracy	C. Num	Accuracy	C. Num
Range Based	97.75%	609	93.52%	266	85.20%	104	78.02%	66
Norm. Range Based	98.53%	558	91.05%	144	79.66%	67	71.71%	52

The experiment result confirms the existence of trade-off between accuracy and speed in NRB strategy. Table 2 shows that the existing classes without normalization improve indexing accuracy, but the average number of speaker models included in the list of candidates is much larger. NRB outperformed Range-Based when uses pitch as its direct access facility for 0.75% and the average number of speaker models included in the candidate list is also smaller.

3.3. Normalized Range Based Indexing vs. Multiclass SVM Mapping

Our third experiment is to compare the NRB access method to the multiclass SVM Mapping (MSVM). Both of methods use the lower and upper bound for each class in the pitch, flatness, brightness, and roll off layer. Lower and upper bound is obtained from the mapping process based on a range of values and balancing the number of members in each class. As shown in Table 3, the DAMClass strategy with MSVM Mapping is more effective than the NRB Indexing. MSVM Mapping uses Radial Basis Function (RBF) kernel because the early experiments showed that RBF kernel is the best kernel for the dataset compared to linear and polynomial kernel. RBF gives the highest classification accuracy and significantly faster during the training and testing process. The analysis result that has been done shows that MSVM Mapping can map the speaker models of each speaker properly, but there is a price to be paid. If the number of speaker models in its classes is not balanced, the access time would increase.

Table 3. NRB Indexing vs. MSVM Mapping

Direct Access Strategy	Direct Access Facility			
	Pitch	Flatness	Brightness	Roll Off
NRB – Digit Dataset	95.78%	89.85%	89.28%	88.15%
MSVM – Digit Dataset	95.89%	89.78%	92.06%	93.06%
NRB – Hyke Dataset	93.48%	88.65%	89.86%	89.86%
MSVM – Hyke Dataset	93.96%	90.34%	89.37%	89.37%

Figure 5 is a scatterplot that describes a comparison of several access methods in DAMClass. In the experiment, we modify the lower and upper bound by adding the value of tolerance for handle speaker models that are in the transition zone from one class to another. The result shows that adding the value of tolerance can improve indexing accuracy. This strategy outperformed the earlier direct access strategies in the same number of speaker models in the candidate list.

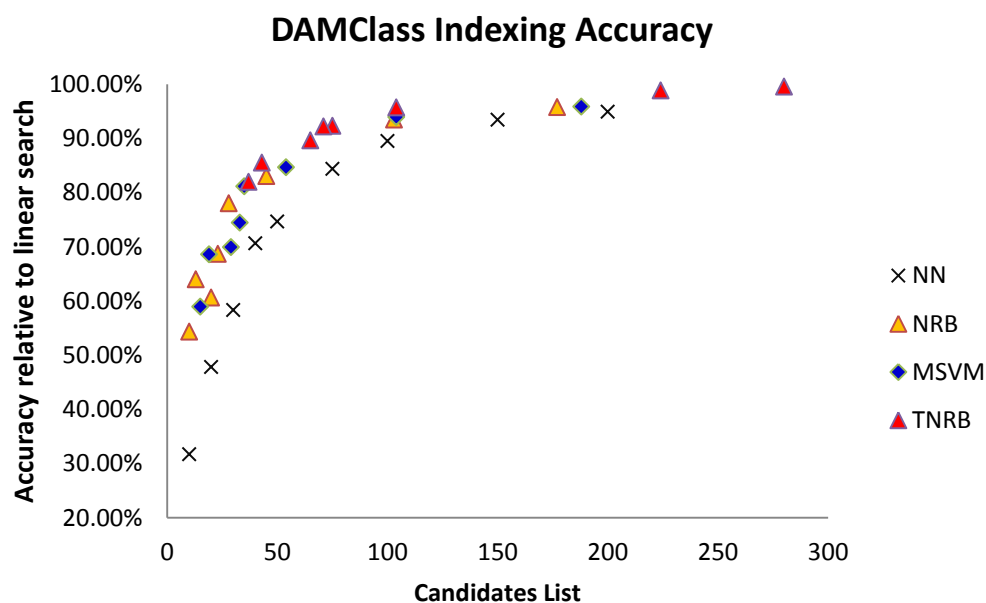


Figure 5. DAMClass Indexing Accuracy

4. Conclusion

We presented a novel direct access strategy based on classification for speaker indexing. Our experiment result shows that this proposed model can improve the data access time of the speaker identification system. The baseline speaker identification system uses GMM algorithm for speaker modeling and training and EM algorithm for speaker matching.

We use our own dataset which is in Bahasa and Hyke dataset which is in English to evaluate the performance of our direct access strategy. Based on the experiment results, DAMClass with Multiclass SVM Mapping strategy gives better performance than the Range Based Indexing. The Normalized Range Based Indexing accuracy is 91.05% relative to linear search or sequential access method and the access time is 16 times faster than the sequential access method. Pitch is the best direct access facility in this paper. The indexing accuracy by pitch is 95.89% in Multiclass SVM Mapping strategy which outperformed the flatness, brightness, and roll off.

Modifying the lower and upper bound of DAMClass strategy increases the indexing accuracy. However the number of speaker models in the candidate list increases compared to the fixed lower and upper bound. The experiment confirms the existence of a trade-off between accuracy and speed in the direct access method. The optimum trade-off point of DAMClass in

indexing accuracy is 95.74% and the number of speaker models in the candidate list is 104. It means that DAMClass can speed up the data access time (compared to sequential access method) by 7-8 times. Larger number of speaker models for each speaker can improve indexing accuracy of DAMClass.

From the experiments in this paper, we can conclude that the DAMClass is a robust and stable direct access strategy. This strategy can also be applied in a voice biometric system, such as access control or speaker diarization. There are a couple of issues in DAMClass that needs further investigation. First, how to determine the other direct access facilities in the form of audio features that are really discriminative, permanent, and has a structure that is easily accessible. Second, how to implement an approach other than statistical approaches, such as syntactic and semantic approaches in the direct access method in particular audio data for the speaker identification system.

References

- [1] Heryanto H, Akbar S, Sitohang B. *Direct Access in Content-Based Audio Information Retrieval: A State of The Art and Challenges*. IEEE International Conference of Electrical Engineering and Informatics (ICEEI). Bandung. 2011; 2: 644-649.
- [2] Heryanto H, Akbar S, Sitohang B. *A New Direct Access Framework for Speaker Identification System*. IEEE International Conference on Data and Software Engineering (ICODSE). Bandung, 2014; 1: 7-11.
- [3] Kwon S, Narayanan S. Unsupervised Speaker Indexing Using Generic Models. *IEEE Transactions on Speech and Audio Processing*. 2005; 13(5): 1004-1013.
- [4] Schmidt L, Sharifi M, Moreno IL. *Large-Scale Speaker Identification*. IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP). Florence. 2014; 1: 1669-1673.
- [5] Indrawan G, Sitohang B, Akbar S. Review of Sequential Access Method for Fingerprint Identification. *TELKOMNIKA*. 2012; 10(2): 335-342.
- [6] Indrawan G, Sitohang B, Akbar S. Fingerprint Direct-Access Strategy Using Local-Star-Structurebased Discriminator Features: A Comparison Study. *International Journal of Electrical and Computer Engineering (IJECE)*. 2014; 4(5): 817-830.
- [7] Ning W. Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features. *IEEE Transaction on Audio, Speech, and Language Processing*. 2011; 19(1): 196-205.
- [8] Hosseinzadeh D, Krishnan S. *Combining Vocal Source and MFCC Features for Enhanced Speaker Recognition Performance Using GMMs*. IEEE 9th Workshop on Multimedia Signal Processing. Crete. 2007; 1: 365-368.
- [9] Karpov E. Real-Time Speaker Identification. Master Thesis. Joensuu: Post Graduate Department of Computer Science, University of Joensuu; 2003.
- [10] Reynolds DA, Rose RC. Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models. *IEEE Transactions on Speech and Audio Processing*. 1995; 3(1): 72-83.
- [11] Chen WC, Hsieh CT, Hsu CH. Robust Speaker Identification System Based on Two-Stage Vector Quantization. *Tamkang Journal of Science and Engineering*. 2008; 11(4): 357-366.
- [12] Lartillot O, Toivainen P, Eerola T. *A Matlab Toolbox for Music Information Retrieval*. University of Jyvaskyla. Finlandia. 2007.
- [13] Giannakopoulos T. *Some Basic Audio Features*. Department of Informatics and Telecommunications, University of Athens. Greece. 2010.
- [14] Maltoni D, Maio D, Jain AK, Prabhakar S. *Handbook of Fingerprint Recognition*. London: Springer. 2009.
- [15] Reda A, Panjwani S, Cutrell E. *Hyke: A Low-Cost Remote Attendance Tracking System for Developing Regions*. Proceedings of the 5th ACM workshop on Networked systems for developing regions. New York. 2011; 1: 15-20.
- [16] Dijk ET, Jagannathan SR, Wang D. *Voice-based Human Recognition*. Eindhoven University of Technology. 2011.