Query by Example of Speaker Audio Signals using Power Spectrum and MFCCs

Pafan Doungpaisan[1] **and Anirach Mingkhwan**[2]
[1,2]Faculty of Information Technology, King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800, Thailand
[2]Faculty of Industrial Technology and Management, King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Road, Wongsawang, Bangsue, Bangkok 10800, Thailand

Article Info

Article history:
Received: Mar 13, 2017
Revised: Aug 9, 2017
Accepted: Aug 22, 2017

Keyword:
Speaker identification
Acoustic signal processing
Content-Based Audio Retrieval
Speaker Recognition
Database Query Processing

ABSTRACT

Search engine is the popular term for an information retrieval (IR) system. Typically, search engine can be based on full-text indexing. Changing the presentation from the text data to multimedia data types make an information retrieval process more complex such as a retrieval of image or sounds in large databases. This paper introduces the use of language and text independent speech as input queries in a large sound database by using Speaker identification algorithm. The method consists of 2 main processing first steps, we separate vocal and non-vocal identification after that vocal be used to speaker identification for audio query by speaker voice. For the speaker identification and audio query by process, we estimate the similarity of the example signal and the samples in the queried database by calculating the Euclidian distance between the Mel frequency cepstral coefficients (MFCC) and Energy spectrum of acoustic features.
The simulations show that the good performance with a sustainable computational cost and obtained the average accuracy rate more than 90%.

1. INTRODUCTION

The Internet has become a major component of their everyday social lives and business. Another important use of the internet is search engine technology. Though they rarely give it a moment has thought, the search engines that help them navigate through the many of information, web pages, images, video files, and audio recordings found on the World Wide Web have become important. Search engine technology is develop over 20 years ago [1][2]. It has changed how we get information at school, collage, work and home. A search engine is an information retrieval system designed to search for information on the World Wide Web. The search results are generally as search engine results pages (SERPs). A search engine results page, or SERP, is the web page that appears in a browser window when a keyword query is put into a search field on a search engine.The information may be a mix of text, web pages, images, video, and other types of files. Some search engines also mine data available in databases. Search engines also maintain real-time information by running an algorithm on a web crawler. It would be easy to find if to search by entering keywords. However, if we want to used search engine to search image or sound. It will make more difficult and more complicated.

1.1. Content-based image retrieval or Reverse image search engines

Content-based image retrieval or Reverse image search engines are those special kind of search engines where you dont need to input any keyword to find pictures [3][4][5][6]. Instead, we have to put a picture and the search engine finds the images similar to you entered. Thus, you can get to know everything you wish to, just with the help of one picture. Practical uses for reverse image search include

- Searching for duplicated image or content.
- Locating the source information for an image.

- Ensuring compliance with copyright regulations.
- Finding information about unidentified products and other objects.
- Finding information about faked images.
- Finding higher resolution versions of images.

There are three types of Content-based image retrieval or Image Search Engines such as Search by Meta-data, Search by Example, and Hybrid Search.

1.1.1. Search by Meta-data

Search by Meta-data: Metadata is data that summarizes basic information about image, which can make finding and working with particular instances of data easier. For example, author, file size, date created and date modified. All are examples of very basic document metadata.

A famous Search Engines such as Google are presented with a text box that you type your keywords into, and click buttons: Google Search. Manually typing in keywords and finding interrelated results. In fact, a meta-data image search engine is only marginally different from the text search engine mentioned above. A search by meta-data image search engine rarely examines the actual image itself. Instead, it relies on textual clues. These searches can come from a variety of sources.The two main methods of Search by Meta-data are Manual Annotations and Contextual Hints.

1.1.2. Search image by Example image

Search image by Example image: Search image by Example image, we can used Google or TinEye. Instead, we can build a search by example image search engine. These types of image search engines try to quantify the image itself and are called Content Based Image Retrieval (CBIR) systems. An example image would be to characterize the color of an image by the standard deviation, mean, and skewness of the pixel intensities in the image. By given a dataset of images, we would compute these moments over all images in our dataset and store them on disk. The next step is called indexing. When we quantify an image, we are describing an image by extracting image features. These image features are an abstraction of the image and used to characterize the content of the image. Lets pretend that we are building an image search engine for Twitter.

1.1.3. Hybrid Approach

Hybrid Approach: An interesting hybrid approach is Twitter. Twitter allows you to include text and images with your tweets. Twitter lets you used hashtags to your own tweets. We can used the hashtags to build a search by meta-data image search engine and then analyzed and quantified the image itself to build a search by example image search engine. From this concept, we would be building a hybrid image search engine that includes both keywords and hashtags with features extracted from the images.

1.2. Content-based audio retrieval or audio search engines

Content-based functionalities aim at finding new ways of querying and browsing audio documents as well as automatic generating of metadata, mainly via classification. Query-by-example and similarity measures that allow perceptual browsing of an audio collection is addressed in the literature and exist in commercial products, see for instance: www. findsounds.com, www.soundfisher.com. There are three types of Content-based audio retrieval Such as Search by search from text, search from image and search from audio.

1.2.1. Audio search from text or Search by Meta-data

Audio search from text or Search by Meta-data: Text entered into a search bar by the user is compared to the search engine's database. Matching results are accompanied by a description or Meta-data of the audio file and its characteristics such as sample frequency, bit rate, type of file, length, duration, or coding type. The user is given the option of downloading the resulting files. On other hand, Keywords are generated from the analyzed audio by using speech recognition techniques to convert audio to text. These keywords are used to search for audio files in the database such as Google Voice Search.

1.2.2. Audio search from image

Audio search from image: The Query by Example (QBE) system is a searching algorithm that uses Content-based image retrieval (CBIR). Keywords are generated from the analyzed image. These keywords are used to search for audio files in the database. The results of the search are displayed according to the user preferences regarding to the type of file such as wav, mp3, aiff etc.

1.2.3. Audio search from audio

Audio search from audio: In audio search from audio, the user must play the audio of a song either with a music player, by singing or by humming to the microphone. Then, an audio pattern is derived from the audio waveform and a frequency representation is derived from its Discrete Fourier Transform. This pattern will be matched with a pattern corresponding to the waveform of sound files found in the database. All those audio files in the database whose patterns are similar to the pattern search will be displayed as search results.

The most popular of an audio search from audio is an audio fingerprint [7][8][9]. An audio fingerprint is a content-based compact signature that summarizes an audio files. Audio fingerprinting technologies have recently attracted attention since they allow the monitoring of audio independently of its format and without the need of watermark embedding or meta-data. Audio fingerprinting, also named as audio hashing, has been well-known as a powerful technique to perform audio identification and synchronization. The figure 1 describe a model of audio fingerprint, An audio fingerprint involves two major steps: fingerprint or voice pattern design and matching search. While the first step concerns the derivation of a compact and robust audio signature, the second step usually requires knowledge about database and quick-search algorithms [10].

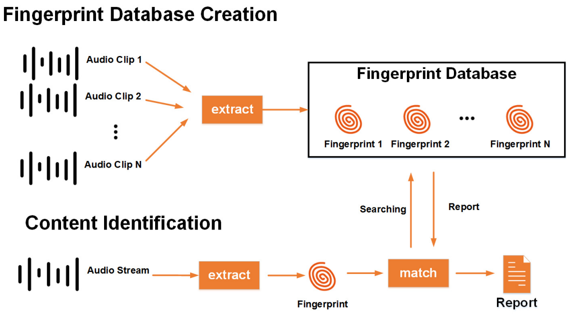

Figure 1. State-of-the-art audio fingerprinting algorithms.

For example of Audio Fingerprinting application are Shazam (http://www.shazam.com/) and SoundHound (http://www.soundhound.com/) [11][12].

A human listener can only identify a piece of music if he has heard it before, unless he has access to more information than just the audio signal. Similarly, fingerprinting systems require previous knowledge of the audio signals in order to identify them, since no information other than the audio signal itself is available to the system in the identification phase. Therefore, a musical knowledge database must be built. This database contains the fingerprints of all the songs the system is supposed to identify. During detection, the fingerprint of the input signal is calculated and a matching algorithm compares it to all fingerprints in the database. The knowledge database must be updated as new songs come out. As the number of songs in the database grows, memory requirements and computational costs also grow; thus, the complexity of the detection process increases with the size of the database. This technique is useful but it has its limitations.

- Audio fingerprint cannot find a live version of music because different key or tempo.
- Audio fingerprint cannot find a cover version because different instruments.
- Audio fingerprint cannot find a hummed version of music because single melody.
- Audio fingerprint unable to find music if a singer sings with text independent.

Unfortunately, in the present audio search engine cannot be search human voice in the database of speakers by example of Spoken audio signals. From this problem, this paper proposes a method for query by example of Spoken Audio signals by using Speaker identification algorithm.

2. LITERATURE REVIEWS (SPEAKER VERIFICATION AND IDENTIFICATION)

Speaker identification is one of the main tasks in speech processing. In addition to identification accuracy, large-scale applications of speaker identification give rise to another challenge: fast search in the database of speakers. Research about Speaker recognition, there are two different types of Speaker Recognition [13][14] consist of Speaker Verification and Speaker Identification. Speaker Verification is the process of accepting or rejecting the identity mention of a speaker. Speaker Identification is the process of determining which registered speaker provides a given utterance. In the speaker identification task, a voice of an unknown speaker is analyzed and then compared with speech samples of known speakers. The unknown speaker is identified as the speaker whose model best matches the input model. There are two different types of speaker identification consist of open-set and closed-set.

Open-set identification similar as a combination of closed-set identification and speaker verification. For example, a closed-set identification may be proceed and the resulting ID may be used to run a speaker verification session. If the test speaker matches the target speaker, based on the ID returned from the closed-set identification, then the ID is accepted and it is passed back as the true ID of the test speaker. On the other hand, if the verification fails, the speaker may be rejected all-together with no valid identification result. Closed-set identification is the simpler of the two problems. In closed-set identification, the audio of the test speaker is compared against all the available speaker models and the speaker ID of the model with the closest match is returned. In closed-set identification, the ID of one of the speakers in the database will always be closest to the audio of the test speaker; there is no rejection scheme.

Speaker verification is the process of verifying the claimed identity of a speaker based on the speech signal from the speaker call a voiceprint. In speaker verification, a voiceprint of an unknown speaker who claims an identity is compared with a model for the speaker whose identity is being claimed. If the match is good enough, the identity claim is accepted. A high threshold reduces the probability of impostors being accepted by the system, increasing the risk of falsely rejecting valid users. On the other hand, a low threshold enables valid users to be accepted consistently, but with the risk of accepting impostors. In order to set the threshold at the optimal level of impostor acceptance or false acceptance and customer rejection or false rejection. The data showing impostor scores and distributions of customer are needed.

There are two types of speaker verification systems: Text-Independent Speaker Verification (TI-SV) and Text-Dependent Speaker Verification (TD-SV). TD-SV requires the speaker saying exactly the enrolled or given password. Text independent Speaker Verification is a process of verifying the identity without constraint on the speech content. Compared to TD-SV, it is more convenient because the user can speak freely to the system. However, it requires longer training and testing utterances to achieve good accuracy and performance. Using audio in identifying the many factors involved. Factors within the characteristics of human sound and technologies related to the acquisition of sound. Factors within the characteristics of human such as.

- The context of a public speaking experience is the environment or situation in which the speech occurs [14].
- Each person has a unique communication style such as Pitch, Tone or Timbre, Melody, Volume or Intensity and Rhythm [15][16][17].
- Emotional speech or the mood of the people such as angry, sad, fearful and fun.
- Physiological, sometimes people have an illness or in a state of alcohol or drugs. Which influences the sound [17].
- Counterfeiting or Disguise sound, Sometimes speaker changed my own voice, so changed from the original, whether it is higher or lower. Speaking to the rhythm that influences the characteristics of sound [18].
- Factors within the technologies related to the acquisition of sound such as. The quality of the microphone (Microphone) or the equipment used to record which greatly affects the quality of the sound. The microphones each have different features. The frequency response or the microphone's sensitivity to sound from various directions [18][19][20].
- Environment in sound recording such as the noise from the environment, the distance from the microphone to record sound [21][22].

- The basics of digital audio: sample rate, bitrate, and how analog signals are represented digitally.

This research, we have worked on text-independent speaker verification. Research interesting of speaker recognition such as. Research of Poignant, J. [23] used an unsupervised approach to Identifying speakers in TV broadcast without biometric models. Existing methods usually use pronounced names, as a source of names, for identifying speech clusters provided by a speaker diarisation step but this source is too imprecise for having sufficient confidence. Existing methods propose two approaches for finding speaker identity based only on names written in the image track. With the late naming approach, there propose different propagations of written names onto clusters. Our second proposition, Early naming, modifies the speaker diarisation module by adding constraints preventing two clusters with different associated written names to be merged together. These methods were tested on the REPERE corpus phase 1, containing 3 hours of annotated videos. With the late naming system reaches an F-measure of 73.1%. With the early naming improves over this result both in terms of identification error rate and of stability of the clustering stopping criterion. By comparison, a mono-modal, supervised speaker identification system with 535 speaker models trained on matching development data and additional TV and radio data only provided a 57.2% F-measure.

Research of M. K. Nandwana [24] used an unsupervised approach for detection of human scream vocalizations from continuous recordings in noisy acoustic environments. The proposed detection solution is based on compound segmentation, which employs weighted mean distance, T2-statistics and Bayesian Information Criteria for detection of screams. This solution also employs an unsupervised threshold optimized Combo-SAD for removal of non-vocal noisy segments in the preliminary stage. A total of five noisy environments were simulated for noise levels ranging from -20dB to +20dB for five different noisy environments. Performance of proposed system was compared using two alternative acoustic front-end features (i) Mel-frequency cepstral coefficients (MFCC) and (ii) perceptual minimum variance distortionless response (PMVDR). Evaluation results show that the new scream detection solution works well for clean, +20, +10 dB SNR levels, with performance declining as SNR decreases to -20dB across a number of the noise sources considered.

Research of Almaadeed, N. [25] is to investigate the problem of identifying a speaker from its voice regardless of the content. In this study, the authors designed and implemented a novel text-independent multimodal speaker identification system based on wavelet analysis and neural networks. The system was found to be competitive and it improved the identification rate by 15% as compared with the classical MFCC. In addition, it reduced the identification time by 40% as compared with the back-propagation neural network, Gaussian mixture model and principal component analysis. Performance tests conducted using the GRID database corpora have shown that this approach has faster identification time and greater accuracy compared with traditional approaches, and it is applicable to real-time, text-independent speaker identification systems.

Research of Xiaojia Zhao [26] investigates the problem of speaker identification and verification in noisy conditions, assuming that speech signals are corrupted by environmental noise. This paper is focused on several issues relating to the implementation of the new model for real-world applications. These include the generation of multicondition training data to model noisy speech, the combination of different training data to optimize the recognition performance, and the reduction of the model's complexity. The new algorithm was tested using two databases with simulated and realistic noisy speech data. The first database is a redevelopment of the TIMIT database by rerecording the data in the presence of various noise types, used to test the model for speaker identification with a focus on the varieties of noise. The second database is a handheld-device database collected in realistic noisy conditions, used to further validate the model for real-world speaker verification. The new model is compared to baseline systems and is found to achieve lower error rates.

Pathak, M.A. and Raj, B., [27] present frameworks for speaker verification and speaker identification systems, where the system is able to perform the necessary operations without being able to observe the speech input provided by the user. In this paper formalize the privacy criteria for the speaker verification and speaker identification problems and construct Gaussian mixture model-based protocols. The proposed also report experiments with a prototype implementation of the protocols on a standardized dataset for execution time and accuracy.

Bhardwaj, S. [28] presents three novel methods for speaker identification of which two methods utilize both the continuous density hidden Markov model (HMM) and the generalized fuzzy model (GFM), which has the advantages of both Mamdani and Takagi-Sugeno models. In the first method, the HMM is utilized for the extraction of shape-based batch feature vector that is fitted with the GFM to identify the speaker. On the other hand, the second method makes use of the Gaussian mixture model (GMM) and the GFM for the identification of speakers. Finally, the third method has been inspired by the way humans cash in on the mutual acquaintances while identifying a speaker. To see the validity of the proposed models [HMM-GFM, GMM-GFM, and HMM-GFM (fusion)] in a real-life scenario, they are tested on VoxForge speech corpus and on the subset of the 2003 National Institute of Standards and Technology evaluation data set. These models are also evaluated on the corrupted VoxForge speech corpus by mixing with different types of noisy signals at different values of signal-to-noise ratios, and their performance is found superior to that of the well-known models.

Abrham Debasu Mengistu and Dagnachew Melesew Alemayehu [29] presented the implementation of Text Independent Amharic Language Speaker Identification using VQ (Vector Quantization), GMM (Gaussian Mixture Models), BPNN (Back propagation neural network), MFCC (Mel-frequency cepstrum coefficients), GFCC (Gammatone Frequency Cepstral Coefficients). For the identification process, speech signals are collected from different speakers including both sexes; for our data set, a total of 90 speakers speech samples were collected, and each speech have 10 seconds duration from each individual. From these speakers, 59.2%, 70.9% and 84.7% accuracy are achieved when VQ, GMM and BPNN are used on the combined feature vector of MFCC and GFCC.

Wajdi Ghezaiel1, Amel Ben Slimane and Ezzedine Ben Braiek [30] proposed to extract minimally corrupted speech that is considered useful for various speech processing systems. In this paper, there are interested for co-channel speaker identification (SID). There employ a new proposed usable speech extraction method based on the pitch information obtained from linear multi-scale decomposition by discrete wavelet transform. The idea is to retain the speech segments that have only one pitch detected and remove the others. Detected Usable speech was used as input for speaker identification system. The system is evaluated on co-channel speech and results show a significant improvement across various target to Interferer Ratio (TIR) for speaker identification system.

Syeiva Nurul Desylvia [31] presented the implementation of Text Independent Speaker Identification. In this research, speaker identification text independent with Indonesian speaker data was modelled with Vector Quantization (VQ). In this research VQ with K-Means initialization was used. K-Means clustering also was used to initialize mean and Hierarchical Agglomerative Clustering was used to identify K value for VQ. The best VQ accuracy was 59.67% when k was 5. According to the result, Indonesian language could be modelled by VQ. This research can be developed using optimization method for VQ parameters such as Genetic Algorithm or Particle Swarm Optimization.

Hery Heryanto, Saiful Akbar and Benhard Sitohang [32] 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 voices 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 speakers 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.

3. RESEARCH METHOD

This paper presents an audio search engine that can retrieve sound files from a large files system based on similarity to a query sound. Sounds are characterized by speech templates derived from MFCC and Power spectrum. Audio similarity can be measured by comparing templates, which works both for simple sounds and complex audio such as music.

Development in speech technology [33][13] has been inspired by the reason that people desire to develop mechanical models that permits the emulation of human verbal communication capabilities. Speech processing allow computer to follow voice commands and different human languages. A number of relevant tasks For example Source Identification, Automatic Speech Recognition, Automatic Music Transcription, Labeling/Classification/Tagging, Music/Speech/Environmental Sound Segmentation, Sentiment/Emotion Recognition, Common machine learning techniques applied in related fields such as image processing and natural language processing.

Figure 2 describe a model of audio recognition system that represents different stages of a system including pre-processing, feature extraction, classification and language model [13]. The pre-processing transforms the input signal before any information can be extracted at feature extraction stage.

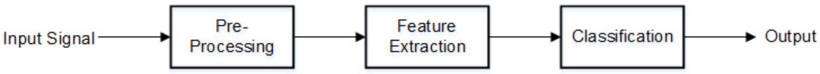

Figure 2. State-of-the-art Audio Classification

vectors must be robust to noise for better accuracy [34][35].The features extraction [35][36][37] is the most important part of a recognizer. If the features are ideally good, the type of classification architecture wont have much importance. On the opposite, if the features cannot discriminate between the concerned classes, no classifier will be efficient, as advanced as it could be. In practical situations, features always present some degrees of overlap from one class to the other class. Therefore, it is worth using good and adapted classification architectures. Feature Extraction for Classification such as Linear prediction coefficients: LPC, Cepstral coefficient, Mel frequency cepatral coefficients: MFCC, Cepstral meansubstraction: CMS and Post filtered cepsturm: PFL.

Classification stage recognize using extracted features and language model where Language Model contains syntax related to language responsible that helps classifier to recognize the input. In Pattern Classification problems, the goal is to discriminate between features representing different classes of interest. Based on learning behavior, classifiers can be divided into two groups: classifiers that use supervised learning (supervised classification) and unsupervised learning (unsupervised classification).

In supervised classification, we provide examples of the correct classification or a feature vector along with its correct class to teach the classifier. Based on these examples, which are commonly termed as training samples, the classifier then learns how to assign an unseen feature vector to a correct class. Examples of supervised classifications include Hidden Markov Model (HMM), Gaussian Mixture Models (GMM), K- Nearest Neighbor (k-NN), Support Vector Machine (SVM), Artificial Neural Networks (ANN), Bayesian Network (BN) and Dynamic Time Wrapping (DTW) [38][39][40][41][42][43].

In unsupervised classification or clustering [38], there is neither explicit teacher nor training samples. The classification of the feature vectors must be based on similarity between them based on which they are divided into natural groupings. Whether any two feature vectors are similar depends on the application. Obviously, unsupervised classification is a more difficult problem than supervised classification and supervised classification is the preferable option if it is possible. In some cases, however, it is necessary to use unsupervised learning. For example, this is the case if the feature vector describing an object can be expected to change with time. Examples of unsupervised classifications include k-means clustering, Self-Organizing Maps (SOM), and Linear vector Quantization (LVQ) [43][44][45].

Classifiers can also be grouped based on reasoning process as probabilistic and deterministic classifiers. Deterministic reasoning classifiers classify sensed data into distinct states and produce a distinct output that cannot be uncertain or disputable. Probabilistic reasoning, on the other hand, considers sensed data to be uncertain input and thus outputs multiple contextual states with associated degrees of truthfulness or probabilities. Decision of the class type to which the feature belongs is made based on the highest probability.

Due to limitation of Audio fingerprint concept, Audio fingerprint cannot used to find audio files if a speaker spoken with text independent. So with the technical limitations of Audio fingerprint, the lack of flexibility in the search for Audio information and cannot be applied to other types of search such as voice search. Therefore, this research was also interested in Speaker Identification concept to be applied to the speaker voice retrieval system. The operating system can process as follows.

3.1. Feature extraction.

Feature extraction is the process of computing a compact numerical representation that can be used to characterize a segment of audio. The research uses Mel Frequency Cepstral Coefficients analysis that based on Discrete Fourier transform (DFT) and Energy spectrum as show in Figure 3.

The use of MFCCs coefficients is common in automatic speech recognition (ASR), although 10-13 coefficients are often considered to be sufficient for coding speech [38]. A subjective pitch is present on Mel Frequency Scale to capture important characteristic of phonetic in speech. MFCC [38][39] is based on human hearing perceptions that cannot perceive frequencies over 1Khz. Figure 3 shows the process of creating MFCCs features. The first step is to segmenting the audio signal into frames with the length with in the range is equal

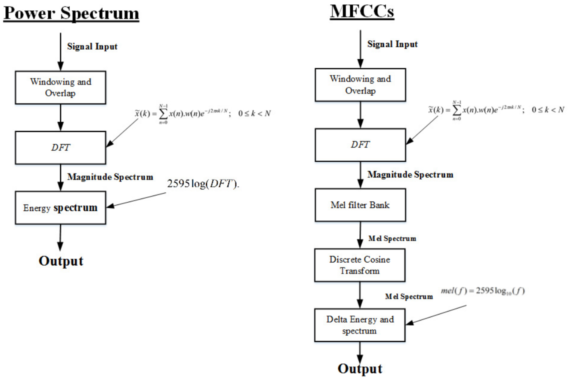
Power Spectrum
Signal Input
Windowing and Overlap
$\widetilde{x}(k)=\sum_{n=0}^{N-1}x(n).w(n)e^{-j2\pi nk/N};\quad 0\le k<N$
DFT
Magnitude Spectrum
$2595\log(DFT).$
Energy spectrum
Output
MFCCs
Signal Input
Windowing and Overlap
$\widetilde{x}(k)=\sum_{n=0}^{N-1}x(n).w(n)e^{-j2\pi nk/N};\quad 0\le k<N$
DFT
Magnitude Spectrum
Mel filter Bank
Mel Spectrum
Discrete Cosine Transform
Mel Spectrum
$mel(f)=2595\log_{10}(f)$
Delta Energy and spectrum
Output

Figure 3. Calculate the energy spectrum (Power spectrum) and MFCCs

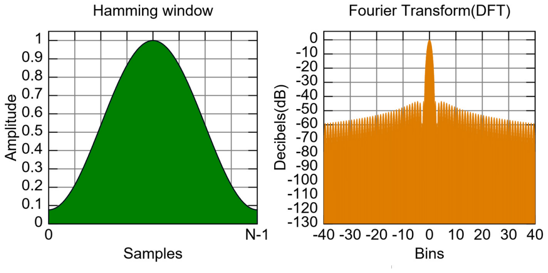

Figure 4. Create a (N-1)-point Hamming window and Display the result.

to a power of two, usually by applying Hamming window function as show in Figure 4. The next step is to take the Discrete Fourier Transform (DFT) of each frame. The next step is Mel Filter Bank Processing. The frequencies range in DFT spectrum is very wide and voice signal does not follow the linear scale. The next step is Discrete Cosine Transform. This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency Cepstrum Coefficient (12 cepstral features plus energy).

The process of creating Energy spectrum features. The first step is to segmenting the audio signal into frames with the length with in the range is equal to a power of two, usually by applying Hamming window function. The next step is to take the Discrete Fourier Transform (DFT) of each frame. The next step is to take the power of each frames, denoted by P(k), is computed by the following equation 1.

$$P(k) = 2595 \times log(DFT) \tag{1}$$

The result of P(k) is called Energy spectrum.

3.1. Measure of similarity

The purpose of a measure of similarity is to compare two vectors and compute a single number which evaluates their similarity. In other words, the objective is to determine to what extent two variables co-vary, which is to say, have the same values for the same cases. Euclidean distance is most often used to compare profiles of respondents across variables. For example, suppose our data consist of demographic information on a sample of individuals, arranged as a respondent-by-variable matrix. Each row of the matrix is a vector of m numbers, where m is the number of variables. We can evaluate the similarity or the distance between any pair of rows. Euclidean Distance is the basis of many measures of similarity and dissimilarity is Euclidean distance. The distance between vectors X and Y is defined as follows:

$$|d_j - d_k| = \sqrt{\sum_{i=1}^{n}(d_j - d_k)^2} \tag{2}$$

In other words, Euclidean distance is the square root of the sum of squared differences between corresponding elements of the two vectors. Note that the formula treats the values of X and Y seriously: no adjustment is made for differences in scale. Euclidean distance is only appropriate for data measured on the same scale. As you will see in the section on correlation, the correlation coefficient is related to the Euclidean distance between standardized versions of the data.

3.2. Content-Based Retrieval of Spoken Audio Step

This section discusses the methodology used in our proposed techniques. It includes the description of the experiment setup, the comparative study method and the implementation details. In a speaker voice retrieval system consisting of two stages as showing in Figure 5. The first stages

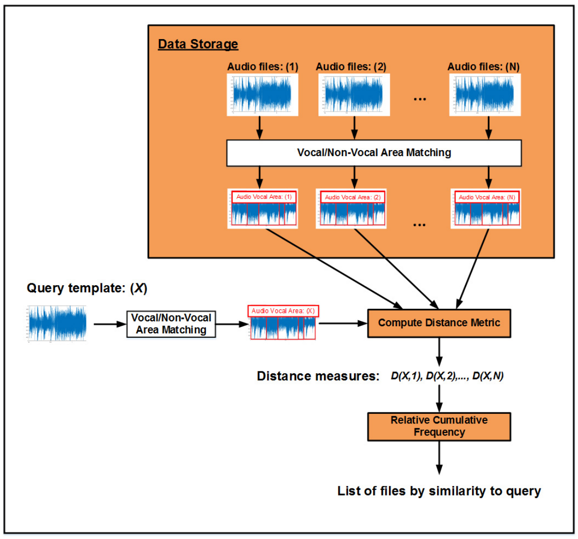

Figure 5. The process of Content-Based Retrieval of Spoken Audio.

The first step, this research store an audio files in voice retrieval system. The step we identify audio areas where is vocal or non-vocal area in each audio files. Because we used only vocal area in each audio files for speaker voice retrieval system. Vocal sound is a type of sound performed by one or more speaker, with or without noises and instrumental accompaniment.

In this step, we used Euclidean Distance to measures a vocal or non-vocal area in each audio files. By using vocal or non-vocal template. The vocal template is consists of singing and speech both men and women length approximately 10 minutes. The non-vocal template is consisting of varied environments sound in background including the meeting rooms of various sizes, office, construction site, television studio, streets, parks, the International Space Station etc. The non-vocal template length approximately 10 minutes.

1. Read the audio files to extract vocal and non-vocal area.
2. Convert audio data to Energy spectra (Energy spectrum) and Mel Frequency Cepstral Coefficents (MFCCs) with windows of 512 as described in Section 4.2 (a concatenation of the Energy spectrum and MFCCs to form a longer feature vector as showing in Figure 6).
3. Calculate the distance between the query-instance of audio files with all samples vector in vocal and non-vocal template.
4. Sort the distance and determine nearest samples based on the minimum distance each sample windows.
5. Use simple majority of the category was choosing from a distance at least to prediction value of the query instance as vocal or non-vocal.
6. Reject the non-vocal vector, leaving only vocal vector of each audio file and storage files to voice retrieval system.

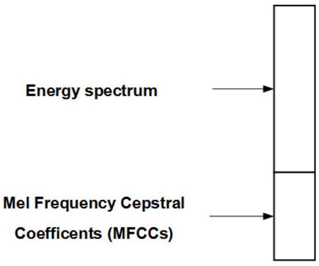

Figure 6. Forming the input vector of energy spectrum (Power spectrum) and MFCCs.

After the procedure, a data in audio files leaving only vocal vector of each audio file. Which is prepared for use in a speaker voice retrieval system. Following a speaker voice retrieval Process according to Figure 7. The second step is the retrieval procedure, here is an algorithm step by step on how to used Euclidean Distance to retrieval human voice in each audio files The goal of this step is to find out target and reject not target files.This research used a relative frequency distribution to decide whether or not target class. A frequency distribution shows the number of elements in a data set that belong to each class. In a relative frequency distribution, the value assigned to each class is the proportion of the total data set that belongs in the class. Heres a formula for calculating the relative frequency of a class:

$$RELATIVEFREQUENCYOFACLASS = \frac{classfrequency}{n} \tag{3}$$

Class frequency refers to the number of observations in each class; n represents the total number of observations in the entire data set.

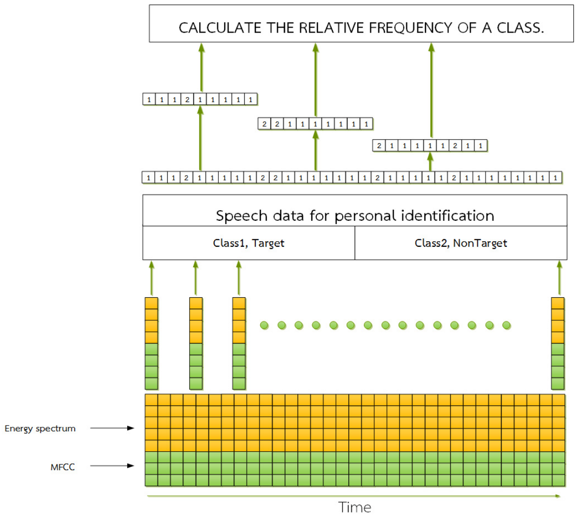

Figure 7. The process of relative frequency and searching for Speaker voice.

1. Read a Voice files of the target person convert to Energy spectrum and Mel Frequency Cepstral Coefficents (MFCCs) and assigned to target class template. The target class template is length approximately 5-10 minutes. Afterthat , Read a Voice file of the non-target person and convert to Energy spectrum and Mel Frequency Cepstral Coefficents (MFCCs). Then assigned to non-target class template. The non-target class template is length approximately 5-10 minutes.

2. To that end, first read vocal files in voice retrieval system and split into frames with predefined duration. Then, each frame further split into N non-overlapping segments, where N called frame size. Afterwards, segments in each frame measure distance between the target class template or non-target class template by Euclidean distance. Used a minimum distance each sample windows to choose the target class or non-target class.

3. After that, we used relative frequency distribution to increase classification accuracy whether target class or non-target class. We calculate frequency distribution every two seconds of vocal files. The frequency distribution refers to the number of observations in each target class; n represents the total number of observations in the entire windows every two seconds of vocal files. A short period of time two seconds derived from the concept of Words per minute (WPM)[46]. If the relative frequency of Target Class is higher than 0.7 percent, we decide this time has a Target Class voice.

4. EXPERIMENTAL EVALUATION.

4.1. Data Collection

This research dataset consists all of 60 people in four languages including Thai, English, Chinese and Japanese for testing. Each languages consist of 15 people. An Audio data used for this experiment more than 3,000 files, total length of 165 hours or 9,900 minutes. Sound files was take from two different sources, the teachings of the MIT OpenCourseWare (http://ocw.mit.edu/courses/audio-video-courses/) and YouTube website (https://www.youtube.com/). All test set are consist varied environments sound in background including the meeting rooms of various sizes, office, construction site, television studio, streets, parks, the International Space Station etc.

All downloaded video files was used Pazera Audio Extractor to extract audio tracks from video file. All audio files after extracted are code in the Wave Files (for uncompressed data, or data loss) Mono Channel and sample rate at 11,025 Hz. We chose this sample rate because the human range is commonly given as 20 to 20,000 Hz, though there is considerable variation between individuals, especially at high frequencies, and a gradual loss of sensitivity to higher frequencies with age is considered normal. In most of the experiment was searched all people in dataset. For this research, we evaluate the performance using the following conditions.

1. For each of the following arguments, we considered correct.
   (a) Speaker verication is decision to accept result, if target voice in audio files.
   (b) Speaker verication is decision to reject result, if target voice not in audio files.
2. For each of the following arguments, we considered mistake.
   (a) Speaker verication is decision to accept result, if target voice not in audio files.
   (b) Speaker verication is decision to reject result, if target voice in audio files.

5. RESULT AND ANALYSIS

Figure 8 shows the summary results for search human voice in the database of speakers by example of audio signal. It can be predicted that accept and rejected with identification result, if target voice in or not in audio files with performance more than 90% in all varying the people. As you can see, the power spectrum (Energy spectrum) and Mel Frequency Cepstral Coefficents (MFCCs) can used together for retrieval human voice and achieved maximum accuracy at 95.00%. Therefore, it was concluded that the power spectrum (Energy spectrum) and Mel Frequency Cepstral Coefficents (MFCCs) can used together to applied to search voice in the database of speakers by example of audio signal.

5.1. Searching on variety languages

Human Beings have a unique sound to their voice. We each have a unique voice because so many factors work together to produce that voice. Your voice starts down in your lungs, where air is exhaled to create an airstream in the trachea and across the larynx, which is often called the voice box. This research, we hypothesized that human voice not depends on language ability. We can use Several languages in searched. This research conducted several experiment to show the hypothesis is true.

1. Experiment 1 used Japanese and English voice in searched.
   (a) Using Japanese and English voice searching files from Japanese and English voice.
   (b) Using Japanese and English voice searching files from Japanese voice.

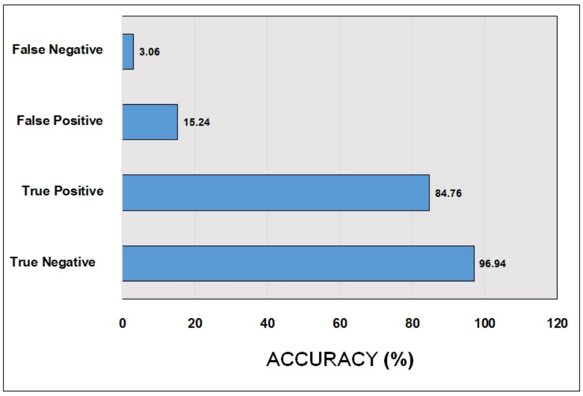

Figure 8. Shows the summary results for search human voice in the database of speakers by example of audio signal by using the power spectrum and Mel Frequency Cepstral Coefficents (MFCCs).

(c) Using Japanese and English voice searching files from English voice.

(d) Using Japanese voice searching files from Japanese and English voice.

(e) Using Japanese voice searching files from Japanese voice.

(f) Using Japanese voice searching files from English voice.

(g) Using English voice searching files from Japanese and English voice.

(h) Using English voice searching files from Japanese voice.

(i) Using English voice searching files from English voice.

2. Experiment 2 used Thai and English voice in searched.

   (a) Using Thai and English voice searching files from Thai and English voice.

   (b) Using Thai and English voice searching files from English voice.

   (c) Using Thai and English voice searching files from Thai voice.

   (d) Using Thai voice searching files from Thai and English voice.

   (e) Using Thai voice searching files from English voice.

   (f) Using Thai voice searching files from Thai voice.

   (g) Using English voice searching files from Thai and English voice.

   (h) Using English voice searching files from English voice.

   (i) Using English voice searching files from Thai voice.

3. Experiment 3 used Chinese and Thai voice in searched.

   (a) Using Thai and Chinese voice searching files from Thai and Chinese voice.

   (b) Using Thai and Chinese voice searching files from Chinese voice.

   (c) Using Thai and Chinese voice searching files from Thai voice.

   (d) Using Thai voice searching files from Thai and Chinese voice.

   (e) Using Thai voice searching files from Chinese voice.

   (f) Using Thai voice searching files from Thai voice.

   (g) Using Chinese voice searching files from Thai and Chinese voice.

   (h) Using Chinese voice searching files from Chinese voice.

   (i) Using Chinese voice searching files from Thai voice.

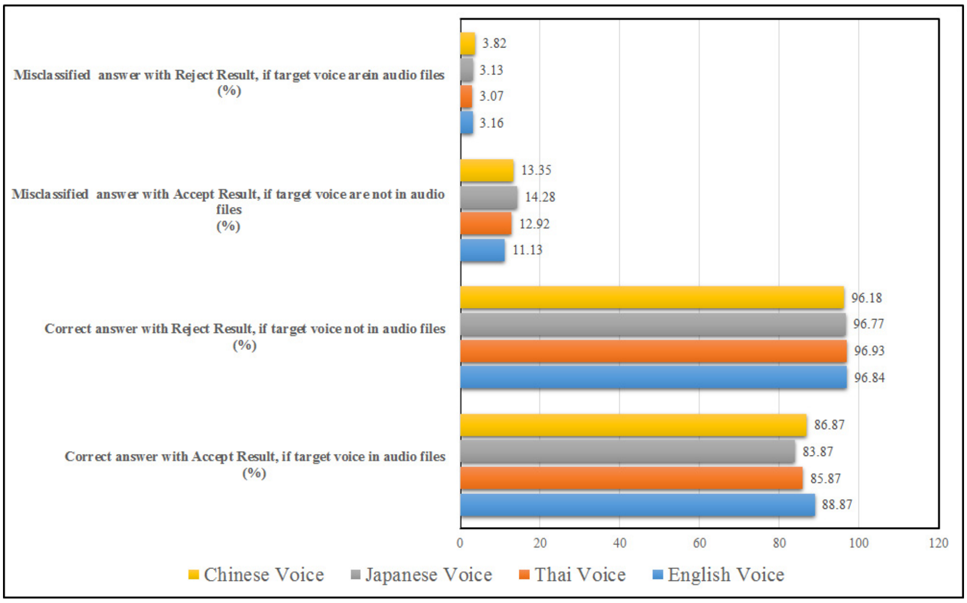

Figure 9. Shows the summary results for Searching human voice on variety languages in the database of speakers by example of audio signal by using the power spectrum (Energy spectrum) and Mel Frequency Cepstral Coefficents (MFCCs).

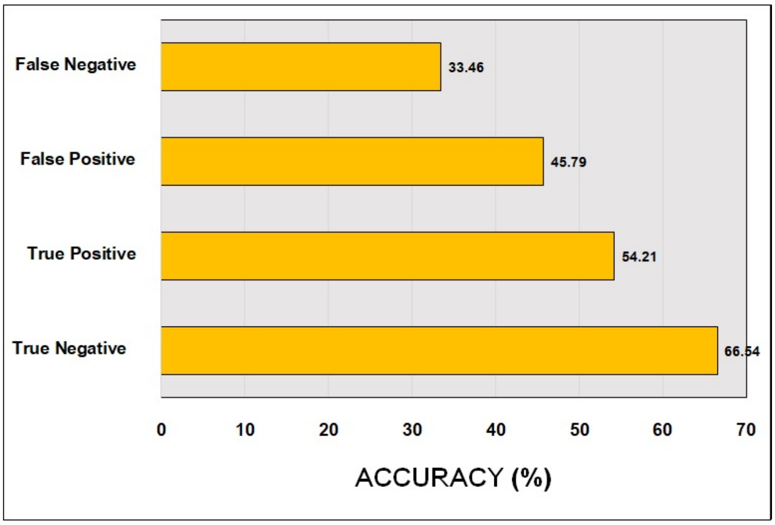

Figure 10. Shows the summary results for search human voice in the database of speakers by example of audio signal by using the Vector Quantization (VQ) and Mel Frequency Cepstral Coefficients (MFCC).

As shown in Figure 9, the results in search of Several languages by example of speech signal. The power spectrum (Energy spectrum) and Mel Frequency Cepstral Coefficents (MFCCs) can used together for retrieval human voice and achieved average accuracy over 85.00% with Speaker verication is decision to accept result, if target voice in audio files and average accuracy over 90.00% with Speaker verication is decision to reject result, if target voice not in audio files.

For reliability, we compare the experimental results of our approach and other researches with the work of Syeiva Nurul Desylvia [31] and Bhardwaj, S. [29]. The work of Syeiva Nurul Desylvia [31] , speaker identification text independent with Indonesian speaker data was modelled with Vector Quantization (VQ) and Mel Frequency Cepstral Coefficients (MFCC) is implemented for feature extraction. The work of Abrham Debasu Mengistu and Dagnachew Melesew Alemayehu [29] presented the implementation of Text Independent Amharic Language Speaker Identification. The algorithm for maximum efficiency was used BPNN (Back propagation neural network), MFCC (Mel-frequency cepstrum coefficients), GFCC (Gammatone Frequency Cepstral Coefficients).

Therefore, we implemented the methods of both of these works to compare the performance. As shown in Figure 8, Figure 10 and 11, the work of Syeiva Nurul Desylvia [31] and Bhardwaj, S. [29] gives a lower accuracy result when it applied to search of Several languages by example of speech signal.

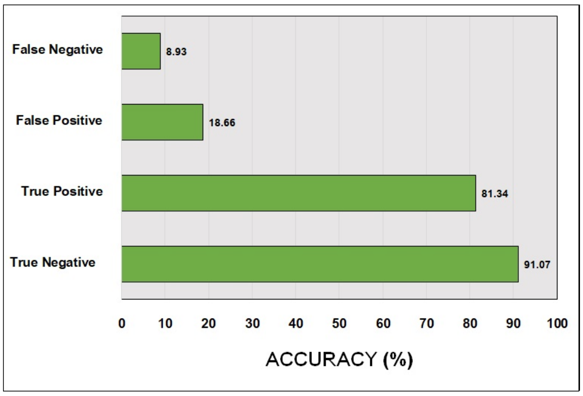

Figure 11. Shows the summary results for search human voice in the database of speakers by example of audio signal by using the BPNN (Back propagation neural network), MFCC (Mel-frequency cepstrum coefficients), GFCC (Gammatone Frequency Cepstral Coefficients)

6. CONCLUSION

This paper proposes a method for query by example of speaker audio signals by using Speaker identification algorithm. Query by example of multimedia signals aims at automatic retrieval of media samples from a database which independent the same speech. The method consists of two main processing first steps, we separate vocal and non-vocal identification after that vocal be used to speaker identification for audio query by speaker voice. For the speaker identification and audio query by process, we estimate the similarity of the example signal and the samples in the queried database by calculating the Euclidian distance between the Mel frequency cepstral coefficients (MFCC) and Energy spectrum of acoustic features. The simulations show that the good performance with low computational cost is obtained with the accuracy rate more than 90%.

ACKNOWLEDGEMENT

REFERENCES

[1] M. Azimzadeh, R. Badie, and M. M. Esnaashari, “A review on web search engines’ automatic evaluation methods and how to select the evaluation method,” in *2016 Second International Conference on Web Research (ICWR)*, pp. 78–83, April 2016.

[2] M. Marn, V. Gil-Costa, C. Bonacic, and A. Inostrosa, “Simulating search engines,” *Computing in Science Engineering*, vol. 19, pp. 62–73, Jan 2017.

[3] D. Boughaci, A. Kemouche, and H. Lachibi, “Stochastic local search combined with lsb technique for image steganography,” in *2016 13th Learning and Technology Conference (L T)*, pp. 1–9, April 2016.

[4] Z. Ji, Y. Pang, and X. Li, "Relevance preserving projection and ranking for web image search reranking," *IEEE Transactions on Image Processing*, vol. 24, pp. 4137–4147, Nov 2015.

[5] P. Huang, G. Liao, Z. Yang, X. G. Xia, J. Ma, and X. Zhang, "An approach for refocusing of ground moving target without target motion parameter estimation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, pp. 336–350, Jan 2017.

[6] C. W. Wang, C. W. Wang, and S. Shieh, "Probebuilder: Uncovering opaque kernel data structures for automatic probe construction," *IEEE Transactions on Dependable and Secure Computing*, vol. 13, pp. 568–581, Sept 2016.

[7] H. abuk and G. nce, "Commercial identification using audio fingerprinting," in *2015 23nd Signal Processing and Communications Applications Conference (SIU)*, pp. 427–430, May 2015.

[8] C. Ouali, P. Dumouchel, and V. Gupta, "Gpu implementation of an audio fingerprints similarity search algorithm," in *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pp. 1–6, June 2015.

[9] R. Sonnleitner and G. Widmer, "Robust quad-based audio fingerprinting," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, pp. 409–421, March 2016.

[10] J. S. Seo, M. Jin, S. Lee, D. Jang, S. Lee, and C. D. Yoo, "Audio fingerprinting based on normalized spectral subband moments," *IEEE Signal Processing Letters*, vol. 13, pp. 209–212, April 2006.

[11] A. L. chun Wang and T. F. B. F, "An industrial-strength audio search algorithm," in *Proceedings of the 4 th International Conference on Music Information Retrieval*, 2003.

[12] SoundHound, "Sound2sound (s2s) search science," jan 2014.

[13] A. N. Ince, *Digital Speech Processing: Speech Coding, Synthesis and Recognition*. Springer Publishing Company, Incorporated, 2014.

[14] J. H. Hansen, "Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition," *Speech Communication*, vol. 20, no. 1, pp. 151 – 173, 1996.

[15] X. Fan and J. H. L. Hansen, "Speaker identification within whispered speech audio streams," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 1408–1421, July 2011.

[16] C. Zhang and J. H. L. Hansen, "Whisper-island detection based on unsupervised segmentation with entropy-based speech feature processing," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, pp. 883–894, May 2011.

[17] M. Mehrabani and J. H. Hansen, "Singing speaker clustering based on subspace learning in the {GMM} mean supervector space," *Speech Communication*, vol. 55, no. 5, pp. 653 – 666, 2013.

[18] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary, and B. A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," in *1995 International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 329–332 vol.1, May 1995.

[19] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1435–1447, May 2007.

[20] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digit. Signal Process.*, vol. 10, pp. 42–54, Jan. 2000.

[21] R. C. Rose, E. M. Hofstetter, and D. A. Reynolds, "Integrated models of signal and background with application to speaker identification in noise," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 245–257, Apr 1994.

[22] Q. Jin, T. Schultz, and A. Waibel, "Far-field speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 2023–2032, Sept 2007.

[23] J. Poignant, L. Besacier, and G. Qunot, "Unsupervised speaker identification in tv broadcast based on written names," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 57–68, Jan 2015.

[24] M. K. Nandwana, A. Ziaei, and J. H. L. Hansen, "Robust unsupervised detection of human screams in noisy acoustic environments," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 161–165, April 2015.

[25] N. Almaadeed, A. Aggoun, and A. Amira, "Speaker identification using multimodal neural networks and wavelet analysis," *IET Biometrics*, vol. 4, no. 1, pp. 18–28, 2015.

[26] X. Zhao, Y. Wang, and D. Wang, "Robust speaker identification in noisy and reverberant conditions," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 836–845, April 2014.

[27] M. A. Pathak and B. Raj, "Privacy-preserving speaker verification and identification using gaussian mix-

ture models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, pp. 397–406, Feb 2013.

[28] S. Bhardwaj, S. Srivastava, M. Hanmandlu, and J. R. P. Gupta, "Gfm-based methods for speaker identification," *IEEE Transactions on Cybernetics*, vol. 43, pp. 1047–1058, June 2013.

[29] A. D. Mengistu and D. M. Alemayehu, "Text independent amharic language speaker identification in noisy environments using speech processing techniques," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 5, p. 109, Jan 2017.

[30] E. B. B. Wajdi Ghezaiel, Amel Ben Slimane, "On usable speech detection by linear multi-scale decomposition for speaker identification," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 6, p. 2766 2772, December 2016.

[31] B. P. S. Syeiva Nurul Desylvia, Agus Buono, "Modeling text independent speaker identification with vector quantization," *Institute of Advanced Engineering and Science*, vol. 15, p. 322 327, March 2017.

[32] B. S. Hery Heryanto, Saiful Akbar, "A new strategy of direct access for speaker identification system based on classification," *Institute of Advanced Engineering and Science*, vol. 13, p. 1390 1398, December 2015.

[33] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Processing Magazine*, vol. 32, pp. 74–99, Nov 2015.

[34] S. Roucos and A. Wilgus, "High quality time-scale modification for speech," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, vol. 10, pp. 493–496, Apr 1985.

[35] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, pp. 357–366, Aug 1980.

[36] H. H., "Perceptual linear predictive (plp) analysis of speech.," *The Journal of the Acoustical Society of America*, vol. 87, p. 1738, apr 1990.

[37] A. V. Oppenheim and R. W. Schafer, "From frequency to quefrency: a history of the cepstrum," *IEEE Signal Processing Magazine*, vol. 21, pp. 95–106, Sept 2004.

[38] Y. Zhang, Z. m. Tang, Y. p. Li, and B. Qian, "Ensemble learning and optimizing knn method for speaker recognition," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4, pp. 285–289, Aug 2007.

[39] C. Shahnaz and S. Sultana, "A feature extraction scheme based on enhanced wavelet coefficients for speech emotion recognition," in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, pp. 1093–1096, Aug 2014.

[40] M. R. Fallahzadeh, F. Farokhi, M. Izadian, and A. A. Berangi, "A hybrid reliable algorithm for speaker recognition based on improved dtw and vq by genetic algorithm in noisy environment," in *2011 International Conference on Multimedia and Signal Processing*, vol. 2, pp. 269–273, May 2011.

[41] N. S. Dey, R. Mohanty, and K. L. Chugh, "Speech and speaker recognition system using artificial neural networks and hidden markov model," in *2012 International Conference on Communication Systems and Network Technologies*, pp. 311–315, May 2012.

[42] A. Krause and H. Hackbarth, "Scaly artificial neural networks for speaker-independent recognition of isolated words," in *International Conference on Acoustics, Speech, and Signal Processing,*, pp. 21–24 vol.1, May 1989.

[43] K. Yu, J. Mason, and J. Oglesby, "Speaker recognition using hidden markov models, dynamic time warping and vector quantisation," *IEE Proceedings - Vision, Image and Signal Processing*, vol. 142, pp. 313–318, Oct 1995.

[44] D. S. Satish and C. C. Sekhar, "Kernel based clustering and vector quantization for speech recognition," in *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, pp. 315–324, Sept 2004.

[45] L. R. Rabiner, S. E. Levinson, and M. M. Sondhi, "On the application of vector quantization and hidden markov models to speaker-independent, isolated word recognition," *The Bell System Technical Journal*, vol. 62, pp. 1075–1105, April 1983.

[46] S. T.-K. K. D. I. S. Group, "Standardized assessment of reading speed: The new international reading speed texts irest," *Investigative Ophthalmology & Visual Science (IOVS)*, vol. 53, p. 1738, mar 2012.