Language and Text-Independent Speaker Recognition System Using Energy Spectrum and MFCCs

Pafan Doungpaisan[1(✉)] and Anirach Mingkhwan[2]

[1] Faculty of Information Technology,
King Mongkut's University of Technology North Bangkok, Bangkok, Thailand
pafan@kmutnb.ac.th
[2] Faculty of Industrial Technology and Management, King Mongkut's
University of Technology North Bangkok, Prachinburi, Thailand
Anirach.M@fitm.kmutnb.ac.th

Abstract. Speaker identification, especially in critical environments, has always been a subject of great interest. In this paper, we present a language and text independent speaker identification algorithm that able to automatically identify a speaker in an audio signal with noise or real environment sound in background. The method is inspired by using a pairing of Energy spectrum and MFCCs audio feature techniques generated from base on Discrete Fourier transform (DFT). After that the audio feature extracted in real time was compared with a Euclidean Distance to measures of different between speakers to obtain the most likely speakers. The Energy spectrum feature is adopted to supplement the MFCC features to yield higher recognition accuracy for speaker identification sound.

The proposed technique is test with 30 different speakers in three languages. The experimental result on speaker identification algorithm using an Energy spectrum and MFCCs features with Euclidean Distance can effectively identify speaker in noise or real environment sound in background with a language and text independent more than 83%. Notably, our approach is not language-specific; it can identify speaker in more than one language.

Keywords: Speaker identification · Energy spectrum · MFCCs

1 Introduction

Speech is the product of a complex behavior conveying different speaker specific nature that are potential sources of complementary information. Historically, speech signal processing and analysis has attracted wide consideration. Especially by using varied applications. For instance, automatic speaker recognition (ASR) have been research areas at least since earlier 70s [1]. Recently, voice has catches again researchers attention its usefulness in order to assess early vocal pathologies, neurodegenerative and mental disorders among others [2]. Progress achieved in these new applications have allowed for a better understanding of the resource of voice production, which have led to an improvement in speaker feature to solve the speaker recognition problem.

P. Meesad et al. (eds.), *Recent Advances in Information and Communication Technology 2017*, Advances in Intelligent Systems and Computing 566,
DOI 10.1007/978-3-319-60663-7_33

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 [3, 4] consist of Speaker Verification and Speaker Identification.

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 and Text-Dependent Speaker Verification. Text-Dependent Speaker Verification 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 Text-Dependent Speaker Verification, 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.

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.

This research, we have worked on language and text-independent speaker verification. Research interesting of speaker recognition such as. Research of Poignant, J. [5] used unsupervised way 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 divarication step but this source is too imprecise for having sufficient confidence. There propose two approaches for finding speaker identity based only on names written in the image track such as with the “late naming” and “Early naming”. These methods were tested on the REPERE corpus phase 1, containing 3 h 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 [6] focused on 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 −20 dB to +20 dB 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 distortion less 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 −20 dB across a number of the noise sources considered.

Research of Almaadeed, N. [7] 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 related system, 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 [8] 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 multi-condition 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., [9] present frameworks for privacy preserving 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 we formalize the privacy criteria for the speaker verification and

speaker identification problems and construct Gaussian mixture model-based protocols. We also report experiments with a prototype implementation of the protocols on a standardized dataset for execution time and accuracy.

Bhardwaj, S. [10] 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 wellknown models.

This paper proposes a speaker verification algorithm that able to automatically identify a speaker in an audio signal with noise or real environment sound in background. The method is inspired by using the Energy spectrum audio feature techniques generated from base on Discrete Fourier transform (DFT). The method is made of two phases: First, the characteristic of the user's voice is generated from components of sound. Second, the characteristic extracted in real time are compared with the Speaker sound using a Euclidean Distance to measures of different between speakers to obtain the most likely speakers.

The rest of the paper is organized as follows. The detail of our proposed algorithm described in Sect. 2. Experimental results showed in Sect. 3 and Sect.4 concludes paper.

2 Methodology

Figure 1 shows a Content-based Speaker Identification Framework. The method is in-spired by using a concatenation of the Energy spectrum and MFCCs features. First, of speaker was extracted without needing a filtering phase. All audio windows were extracted comprehensive characteristic of speaker sound are belonging to two components of sound consist of Energy spectrum and MFCCs feature to yield higher recognition accuracy for speaker identification sound.

In Fig. 2, The Mel-frequency cepstral coefficients (MFCCs) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. These features are typically obtained by first applying a Fourier transform to short-time window segments of audio signals followed by further processing to derive the features of interest. Some commonly used ones include the MFCC [11]: After taking the FFT of each short-time window, the first step in MFCC calculation is to obtain the mel filter bank outputs

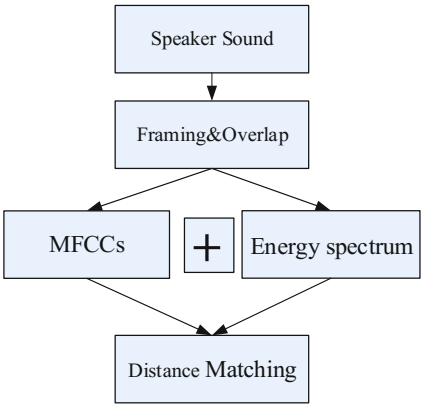

Fig. 1. Content-based Speaker Identification Framework

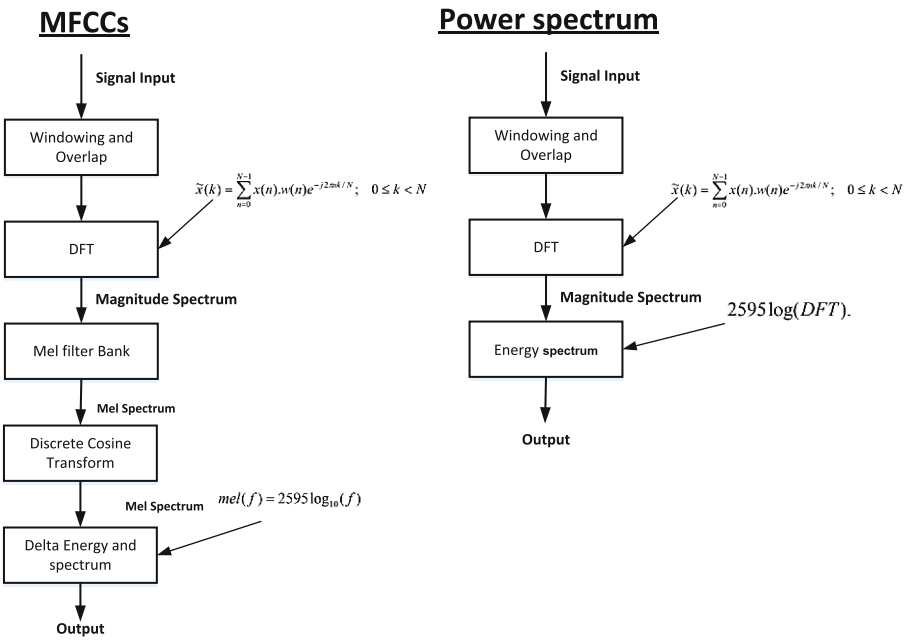

Fig. 2. To calculate the energy spectrum (power spectrum) and calculated MFCCs

by mapping the powers of the spectrum onto the mel scale, using 23 triangular mel filter bank, and transformed into a logarithmic scale, which emphasizes the low varying frequency characteristics of the signal. Typically, 13 Mel-frequency cepstral coefficients are then obtained by taking the discrete cosine transform (DCT).

Figure 2, 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.

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

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

3 Experimental Evaluation

3.1 Data Collection

Audio data used for this experiment included 303 files, total length of 130 h or 7855 min. Sound clips 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 audio files consist of 30 people in three languages with varied environments sound in background including the meeting rooms of various sizes, office, construction site, television studio, streets, parks, the International Space Station. All downloaded video files was used Pazera Audio Extractor to extract audio tracks from video. 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.

3.2 Measure of Similarity

The purpose of a measure of similarity is to compare two vectors and compute a single number that evaluates their similarity. Euclidean distance 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 defined as follows:

$$\left|d_j - d_k\right| = \sqrt{\sum_{i=1}^{n} \left(d_{i,j} - d_{i,k}\right)^2} \tag{2}$$

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.

Here is an algorithm step by step on how to use Euclidean Distance to measures a different between speaker:

1. Calculate the distance between the query instance and all the training samples each category Y.
2. Sort the distance and determine nearest samples based on the minimum distance.
3. Gather the category Y of the minimum distance nearest samples.
4. Use simple majority of the category of nearest samples as the prediction value of the query instance.

4 Result

In each experiment, we performed 50 runs of the 5-fold cross-validation to obtain statistically reliable results. The mean recognition rate was calculated based on the error average for one run on test set. We examined the performance of the Energy spectrum and Mel Frequency Cepstral Coefficents (MFCCs) as described in Sect. 2, a concatenation of the Energy spectrum and MFCCs to form a feature vector as showing in Fig. 2. For statistically reliable results, we compare the overall recognition accuracy using an Energy spectrum and MFCCs with a variety of Distance Measure algorithm as shown in Table 1.

Table 1. Summarized the average accuracy of all Distance Measure.

Feature	Distance measure	Accuracy (%)
Energy spectrum + MFCCs	Euclidean	81.62
	Cityblock	80.40
	Cosine	46.22
	Correlation	44.47
MFCCs	Euclidean	74.40
	Cityblock	72.62
	Cosine	59.36
	Correlation	54.73
Energy spectrum	Euclidean	78.27
	Cityblock	77.98
	Cosine	43.98
	Correlation	44.64

From results in Table 1, by using concatenation of the Energy spectrum and MFCCs was performed better recognition accuracy than using an Energy spectrum or MFCC only for all Distance Measure algorithm. The highest performance of Energy spectrum and MFCC was obtain by using Euclidean Distance 81.62%.

Next, we examined the performance of concatenation of the Energy spectrum and MFCCs with another feature vector such as Spectral centroid, Discrete Fourier Transform, Haar Discrete Wavelet Transform, Mel Frequency Cepstral Coefficents

Table 2. Summarized the average accuracy of Euclidean Distance Measure and all Feature.

Feature	Accuracy (%)
Energy spectrum + MFCCs	81.62
Energy spectrum	78.27
Spectral centroid	12.03
Discrete Fourier Transform	62.36
Haar Discrete Wavelet Transform	16.75
Mel Frequency Cepstral Coefficents (MFCCs)	74.40
RollOff	11.40
Root Mean Square (RMS)	26.22
Linear prediction (LP)	48.26
perceptual linear prediction (PLP) coecients	24.02
partial correlation coefficients (PARCORs)	68.18

(MFCCs), RollOff, Root Mean Square (RMS), Linear prediction (LP), perceptual linear prediction (PLP) coecients and partial correlation coefficients (PARCORs). We comparable performance to another feature extraction method on a similar task. The result was show in Table 2.

From results in Table 2, by compare accuracy all feature with Energy spectrum and MFCC. The Energy spectrum and MFCC was show the best accuracy.

5 Summary

The paper reports a concatenation of the Energy spectrum and MFCCs a small set of time–frequency features, which is flexible, intuitive and physically interpretable. A combination of Energy spectrum and MFCC features can identification a speaker sounds in real environment and improve the overall performance.

The experimental results show promising performance in identifying a different audio speaker and shows comparable performance to another feature extraction method on a similar task. By using Energy spectrum and MFCC was show the best accuracy when compare accuracy all feature. Notably, our approach is not language-specific; it can identify speaker in more than one language.

References

1. Rosenberg, A. E.: Automatic speaker verification: a review. In: Proceedings of the IEEE, pp. 475–487 (1976)
2. Gómez Vilda, P., Rodellar Biarge, V., Nieto Lluis, V., Muñoz Mulas, C., Mazaira-Fernández, L.M., Martínez Olalla, R.: Characterizing neurological disease from voice quality analysis. Cognit. Comput. **5**(4), 399–425 (2013)
3. Furui, S.: Digital Speech Processing: Synthesis, and Recognition. CRC Press, New York (1989)
4. Hansen, J.H.L.: Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech Commun. **20**(1–2), 151–173 (1996)
5. Poignant, J., Besacier, L., Quénot, G.: Unsupervised speaker identification in TV broadcast based on written names. IEEE Trans. Audio Speech Lang. Process. **23**(1), 57–68 (2015)
6. Nandwana, M.K., Ziaei, A., Hansen, J.H.L.: 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, South Brisbane (2015)
7. Almaadeed, N., Aggoun, A., Amira, A.: Speaker identification using multimodal neural networks and wavelet analysis. IET Biom. **4**(1), 18–28 (2015)
8. Zhao, X., Wang, Y., Wang, D.: Robust speaker identification in noisy and reverberant conditions. In: 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3997–4001, Florence (2014)
9. Pathak, M.A., Raj, B.: Privacy preserving speaker verification and identification using gaussian mixture models. IEEE Trans. Audio Speech Lang. Process. **21**(2), 397–406 (2013)
10. Bhardwaj, S., Srivastava, S., Hanmandlu, M., Gupta, J.R.P.: GFM-based methods for speaker identification. IEEE Trans. Cybernet. **43**(3), 1047–1058 (2013)
11. Rabiner, L., Juang, B.H.: Fundamentals of Speech Recognition. PTR Prentice Hall, Englewood Cliffs (1993)