Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.Doi Number

A New Efficient Feature Extraction Technique Based on Voiceless Sound Features for Speaker Dependent Keyword Spotting

Yong Jie Hong1.

1AUO Corporation, No. 23, Lixing Rd., East Dist., Hsinchu City 300, Taiwan, ROC

Corresponding author: First A. Author (e-mail: Jay.YJ.Hung@auo.com).

ABSTRACT This paper is devoted to developing an efficient method for speaker-dependent keyword spotting. To this end, we have defined a new set of feature coefficients called formant-based frequency cepstral coefficients (FFCCs) to account for both voiced and voiceless sound features. FFCCs address the limitations of traditional Mel frequency cepstral coefficients (MFCCs), which emphasize voiced features while underrepresenting voiceless features. In keyword spotting, the underrepresentation of voiceless features often leads to keyword misclassification. Unlike previous studies that used a single filter bank to capture speech features, we have proposed a formant-based filter bank (FFB) to enhance speech characteristics according to the formant distribution of each frame. We have evaluated the performance of the keyword spotting system based on FFCCs using the RSR2015 database and similarly pronounced words.

INDEX TERMS **Feature extraction, Keyword spotting, Mel frequency cepstral coefficient, Speech processing.**

1. INTRODUCTION

The development of speech recognition technology has a long history. With the advent of smart living technologies, speaker-dependent keyword spotting has received unprecedented attention in various applications [13] [14]. In keyword spotting, the features extracted from keywords have a significant role in determining the recognition rate. Hence, the need for an efficient feature extraction technique in keyword spotting systems is inevitable. Existing advanced feature extraction techniques are primarily based on feature coefficients derived from human auditory system characteristics. For example, Mel-frequency cepstral coefficients (MFCCs) [10] [19], gammatone frequency cepstral coefficients (GFCCs) [20] [24], modified Mel-frequency cepstral coefficients (MMFCCs) [3], and power-normalized cepstral coefficients (PNCCs) [21] [12] fall into this category. These feature extraction techniques share a common characteristic: the filter banks used in these algorithms place greater emphasis on the voiced features of the speech signal. Voiced features contain a large number of speaker voiceprints, making them more distinguishable than voiceless features.

However, the absence of voiceless features can easily result in keyword misclassification. For instance, in the words 'stop' and 'top,' the voiced features are highly similar, and it is the voiceless features that provide the key distinction. In addition, some literatures used inverse MFB to extract audio features and named it as inverse Mel filter bank (IMFB). The resulting frequency cepstral coefficients (IMFCCs) outperform the traditional MFCCs by considering high-frequency features and form a complementary relationship with the original MFCCs. Based on such complementary relationship, the classification algorithm was performed separately by using MFCCs and IMFCCs as features. Then, the classification results were finally determined by weighted fusion [4, 6, 7]. Here, we have noticed that IMFCCs method has two drawbacks:

•Defects in the IMFB structure: The IMFB focuses on features ranging from 7000 to 8000 Hz. However, the main feature of the voiceless sound is not fixed in this range. It often distributes between 4000 and 8000 Hz according to different speakers and different keywords. In other words, IMFB cannot accurately reinforce the main features of the voiceless sound.

•Cannot be used as independent sound feature to perform classification: It is aforementioned that the voiced features contain more distinguishing voice prints. IMFCCs itself focuses on voiceless sounds and disregards voiced features. Therefore, the usage of IMFCCs alone as a feature to perform keyword spotting will result in a poor recognition rate.

To address the limitations of MFCCs, particularly their underrepresentation of voiceless features, researchers have explored various alternative approaches. One such technique is the Linear Frequency Cepstral Coefficients (LFCCs), which employs a uniform filter bank across both low and high frequencies. However, this approach somewhat diminishes the emphasis on voiced features. Consequently, when processing speech predominantly composed of voiced sounds, LFCCs may underperform compared to MFCCs, especially in applications like speaker recognition, where capturing voiced features is critical for achieving high recognition accuracy [16] [22].

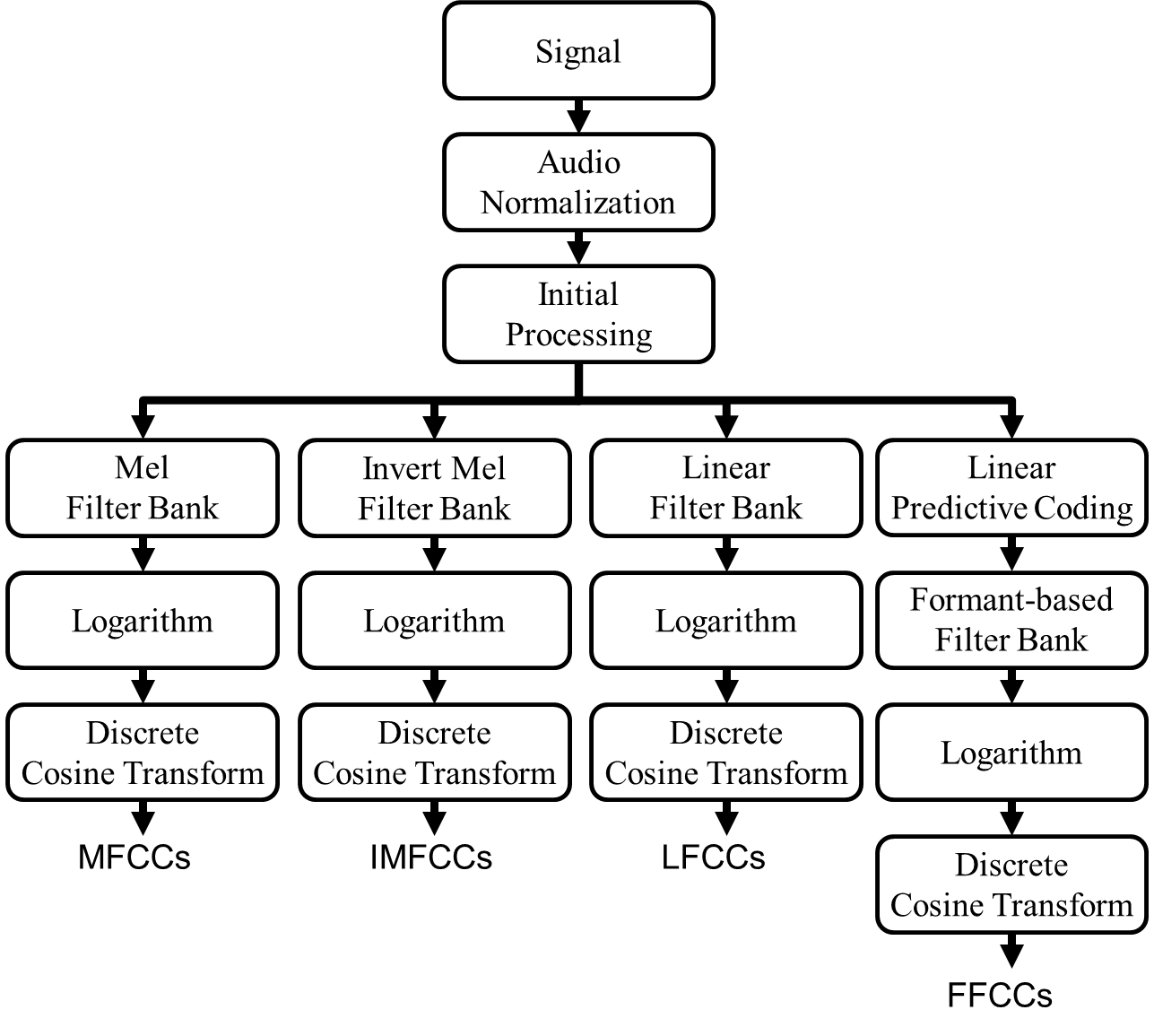
The primary objective of this paper is to enhance the discriminative capability of features when different speakers articulate the same keyword or when the same speaker produces different keywords, thereby facilitating more effective speaker recognition. To achieve this, the paper introduces an innovative audio feature extraction technique: Formant-based Frequency Cepstral Coefficients (FFCCs).

Formants, as the primary frequency components of speech signals, are intrinsically linked to the shape of the speaker's vocal tract and the manner of speech production. They reflect the physiological characteristics of the speaker, such as the length and shape of the vocal tract, leading to distinct formant patterns when different speakers produce identical speech sounds. Moreover, when a speaker articulates different keywords, the formant distribution adapts to the phonetic characteristics of the language. These properties make formants a crucial element in speaker recognition, as they effectively capture the unique attributes of individual speakers [9] [1] [5].

The rationale for leveraging formants in speaker recognition lies in their stability and high discriminative power. Formants not only differentiate speakers but also encapsulate the speaker's unique vocal traits. The proposed FFCCs method builds upon this principle by amplifying the formant distribution within each frame, thereby accentuating the personalized characteristics of speech and improving speaker recognition accuracy.

STRUCTURE OF FFCCs

Traditional feature extraction techniques enhance features within fixed frequency ranges, while the Formant-based Filter Bank (FFB) dynamically adjusts the frequency ranges for enhancement based on formants extracted using linear predictive coding (LPC), thus more effectively highlighting relevant features for different speakers and keywords. FIGURE 1 compares the structural differences between the proposed FFCCs and three traditional feature extraction techniques: MFCCs, IMFCCs, and LFCCs.



**FIGURE 1. Comparison of MFCC, IMFCC, LFCC, and FFCC in terms of structure.**

1. AUDIO NORMLIZATION

Usually even if the same person speaks the same keyword, the volume level will not be the same each time. This difference in volume level adversely affects the accuracy of the speaker dependent keyword spotting. Therefore, in order to reduce the difference in volume between the speech signals, we perform audio normalization, which primarily evaluates the histogram of the signal amplitude and regards the position of 70% accumulation as the reference value of the signal amplitude. Then, the amplitude scaling is set as the ratio of expected value and reference value. The equation is shown below, where represents the sound signal and t represents time.

|  |  |
| --- | --- |
|  | (1) |

1. INITIAL PROCESSING

In the initial processing step, we have filtered out the ultra-low frequency noise caused by microphone. The pre-emphasis filter is [11]. Moreover, a short-time Fourier transformation (STFT) was performed using a 32ms Hamming window, with frames overlap of 16ms [2].

FORMANT-BASED FILTER BANK

This paper studies a voiced and voiceless sound feature extraction approach where the filter placement is determined by distribution of energy in the short-term spectrum. More precisely, in each time frame, this is achieved by estimation of formant information using linear prediction.

1. FORMANT EXTRACTION THROUGH LPC

In this method, we first compute the linear predictive coefficients using the LPC method, which models the speech signal by approximating the vocal tract's behavior. These coefficients represent the spectral envelope of the speech signal and provide valuable information about the resonant frequencies of the vocal tract. Once the linear predictive coefficients are obtained, they are used to derive the LP smoothed spectrum, which is essentially a smoothed version of the speech signal's power spectrum. The LP smoothed spectrum highlights the spectral peaks, which correspond to the formant frequencies in the speech signal. These peaks are important markers for identifying the resonant frequencies of the vocal tract. Formants are then extracted by detecting these peaks from the LP smoothed spectrum, as shown in FIGURE 2. The frequency and amplitude of the peaks provide key information about the phonetic characteristics of speech, such as vowel sounds, and are crucial for speech analysis and synthesis.



**FIGURE 2. Flow chart of extracting formants.**

The short-time transfer function of the vocal tract filter, obtained using the All-pole model, is given by the following equation, where G is the gain factor, P is the order of linear prediction, and is set of LPC coefficients.

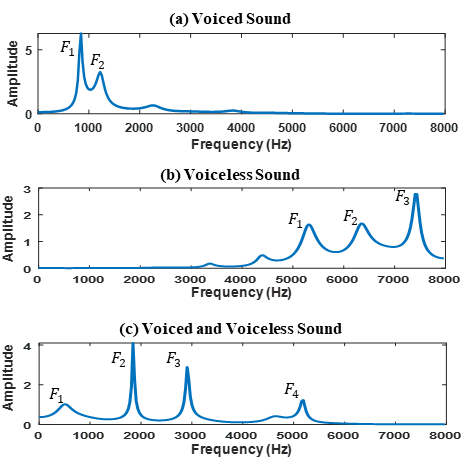
|  |  |
| --- | --- |
|  | (2) |

Each frame is processed by LPC to obtain the formant frequency distribution, which provides important information about the vocal tract resonances. Through experimental results, it has been found that the frequency distribution of the formants can be roughly divided into three categories based on their distinct characteristics. These categories are defined according to the frequency ranges and patterns observed in the formants.

The first category consists of formants below 4000 Hz, indicating that the audio signal in the current frame contains only voiced characteristics. This suggests that the speech signal is produced with vocal cord vibration, resulting in resonant frequencies that fall within this range. This category is often associated with vowels and other voiced sounds, as illustrated in FIGURE 3 (a).

The second category consists of formants above 4000 Hz, signifying that the speech signal in the current frame contains only voiceless characteristics. This typically corresponds to sounds that do not involve vocal cord vibration, such as certain consonants like fricatives and stops. The higher frequency formants reflect the noise components of these sounds, as shown in FIGURE 3 (b).

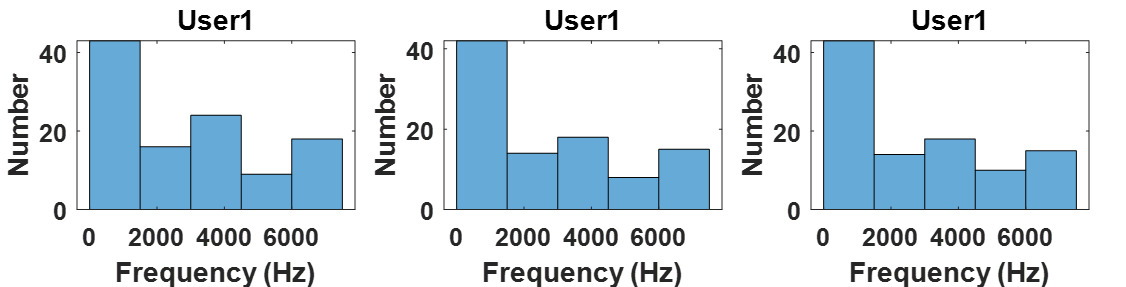
The third category includes formants both below and above 4000 Hz, which implies that the audio signal in the current frame contains a mixture of both voiced and voiceless characteristics. This scenario is often observed in consonant-vowel transitions, where both types of sounds occur in close succession, as depicted in FIGURE 3 (c).



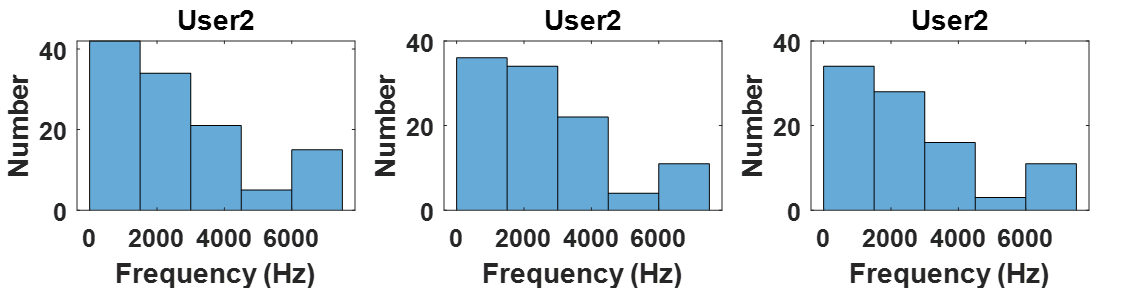
**FIGURE 3. Three types of formant frequency distribution.**

When comparing the histograms of different speakers, we observe that the formant frequency distributions for the same speaker using the same keyword exhibit high consistency, whereas significant differences exist between different speakers. This finding further supports our hypothesis that the formant distribution trends remain stable for the same speaker under identical keyword conditions and can serve as a distinguishing feature for speech recognition or speaker identification.

As illustrated in FIGURE 4 and FIGURE 5, the results show that the distances between histograms of the same speaker are relatively small, whereas those between different speakers are significantly larger. This further validates the feasibility of using a formant-based filter bank for speech feature analysis.



**FIGURE 4. Formant distribution of the keyword “Stop” spoken three times by user 1.**



**FIGURE 5. Formant distribution of the keyword “Stop” spoken three times by user 2.**

As shown in FIGURE 6, each histogram represents the result of ten sound files. These histograms show that the formant frequency distribution is very similar for the same keyword spoken by the same person. However, different people saying the same keyword results in different formant frequency distribution.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Speaker A | Speaker B | Speaker C |
| Glee |  |  |  |
| Rave |  |  |  |
| Stop |  |  |  |

**FIGURE 6. LTF distribution of three keywords spoken ten times by three speakers.**

1. Composition of FFB

FFB is based on the concept that the same speaker and the same keyword share similar formant frequencies. Nolan and Grigoras [18] once remarked that formant-based information is among the most effective for speaker identification. Other researchers [23] [8] also utilized formant characteristics to successfully identify speakers. All of the above evidence confirms that formant frequency distribution remains consistent when the same speaker utters the same keyword. When designing the curve, we define the bell curve using the following Gaussian function:

|  |  |
| --- | --- |
|  | (3) |

Where μ represents the frequency of the second formant, defines the width of the bell shape, which is fixed at 2 kHz in this work, and k denotes the frequency. The aforementioned and σ can be adjusted based on the reader's specific application requirements. Establish the center frequencies () for each filter according to the following equation:

|  |  |
| --- | --- |
|  | (4) |

Where and are respectively the minimum and maximum frequencies of the filter bank, is 1 Hz, is 8000 Hz; and  with the total number of filters N = 80. Finally, Eq. (3) is used to modulate the amplitudes of the triangular band-pass filters, as expressed in Eq. (5).

|  |  |
| --- | --- |
|  | (5) |

After processing the three types of formant frequency distributions in FIGURE 3 through FFB, the resulting frequency response is shown in FIGURE 7.







**FIGURE 7. The frequency response of the FFB.**

1. Composition of FFB

The spectrum of each frame must be enhanced using FFB to obtain its corresponding FFB spectrum. This enhanced spectrum is then processed via discrete cosine transform (DCT) to extract FFCCs. FIGURE 9 presents the pseudocode for obtaining FFCCs.

一張含有 文字, 螢幕擷取畫面, 字型, 數字 的圖片

自動產生的描述

**FIGURE 8.** **FFCCs pseudocode.**

EXPERIMENTAL RESULTS

In this study, we have used the dynamic time warping (DTW) algorithm [17] to perform the speaker dependent keyword spotting based on MFCCs, LFCCs, IMFCCs, and FFCCs respectively, to compare the performance of these four feature extraction methods. When an unknown voice signal is to be identified, its feature coefficients are first extracted and then followed by feature matching with the voice templates in the database based on DTW technique. Assuming that the shortest matching distance is lower than the threshold we have set, the corresponding template is treated as the final target. On the contrary, if it is greater than the threshold, the voice signal is regarded as unknown. It is to be noted that the filter banks for the four feature extraction methods are composed of 80 filters, resulting in the first 13 cepstral coefficients per frame.

In this experiment, we use a sampling frequency of 16 kHz with a frame length of 512 samples and a frame shift of 256 samples. The detection error tradeoff (DET) curve graphically illustrates the relationship between false rejects and false accepts as the decision threshold varies across the genuine and imposter match score distributions.

We determine the equal error rate (EER) using the DET curve. The EER represents the point at which the false accept rate and the false reject rate are equal. Therefore, we use the decision threshold corresponding to the EER as the optimal decision threshold for DTW.

1. Speech Corpus

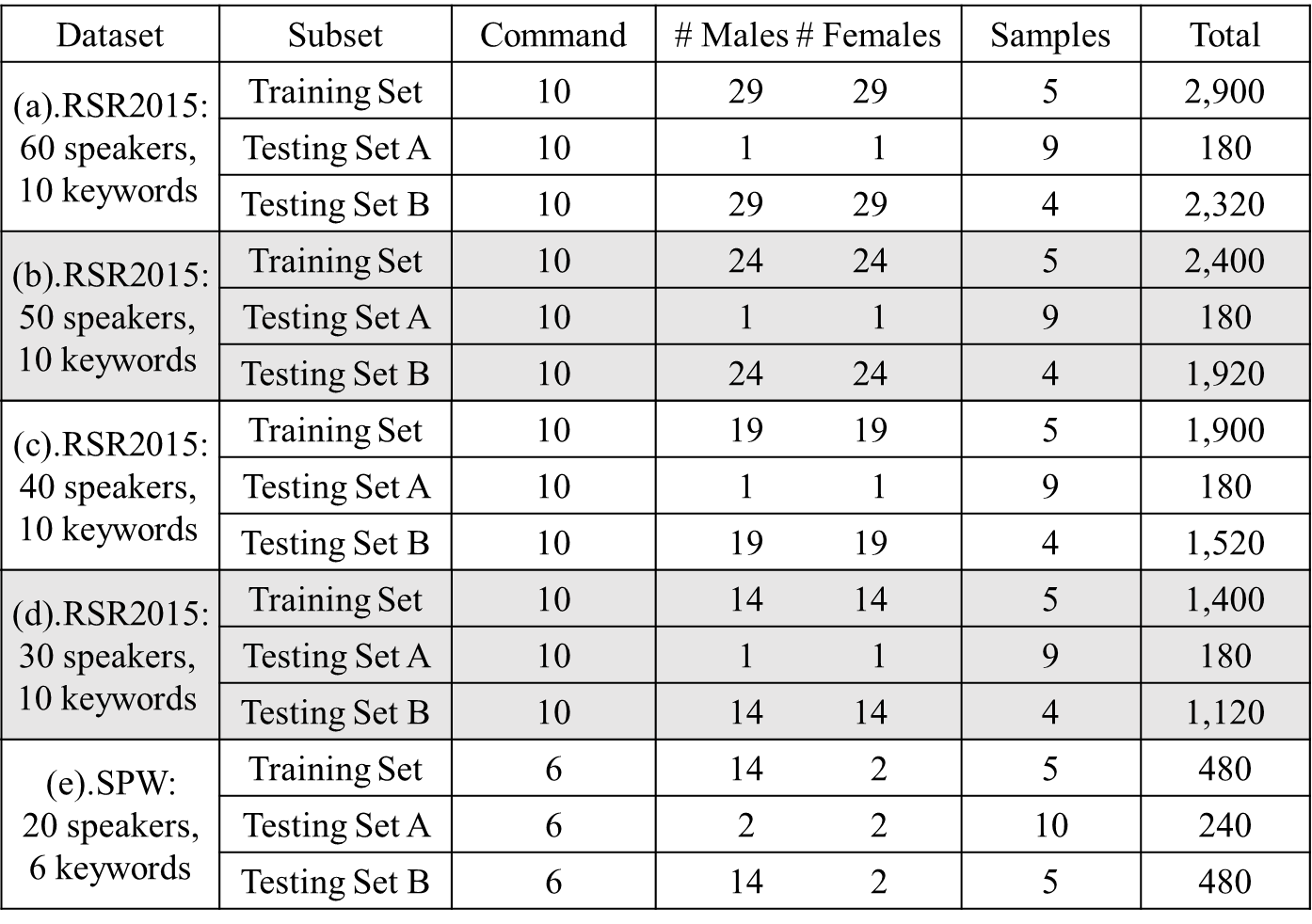
The performance comparison of MFCCs, LFCCs, IMFCCs, and FFCCs was carried out based on the Similarly Pronounced Words (SPW) database and the RSR2015 database [15] in this work. The SPW database consists of voice samples recorded in our laboratory, containing speech data from 20 speakers (6 females and 14 males) uttering 6 keywords, with each keyword having 10 voice samples per speaker. The duration of each keyword is at most 1 second, and the sampling rate is 16 kHz.

The six keywords are “glee,” “sleep,” “rave,” “serve,” “stock,” and “top.” Among these, the voiced features of the paired keywords (“glee,” “sleep”) and (“rave,” “serve”) are highly similar, meaning that only their unvoiced features differ. For the paired keywords (“stock,” “stop”), both voiced and unvoiced features exhibit high similarity. We leveraged the strong similarity between these paired keywords in the SPW database to evaluate the performance of different feature extraction methods.

This study uses 30 keywords of RSR2015 database group II [22], which are used to control smart home appliances. Table I depicts the details of the training sets and testing sets of RSR2015 and SPW. For instance, in dataset (a) in, we have used half of the 24 female and 24 male voices as training set, and the other half of the voices as testing set B. In addition, the voices of remaining 2 speakers who are not included in the speaker template are treated as interference voices, named testing set A.

TABLE I

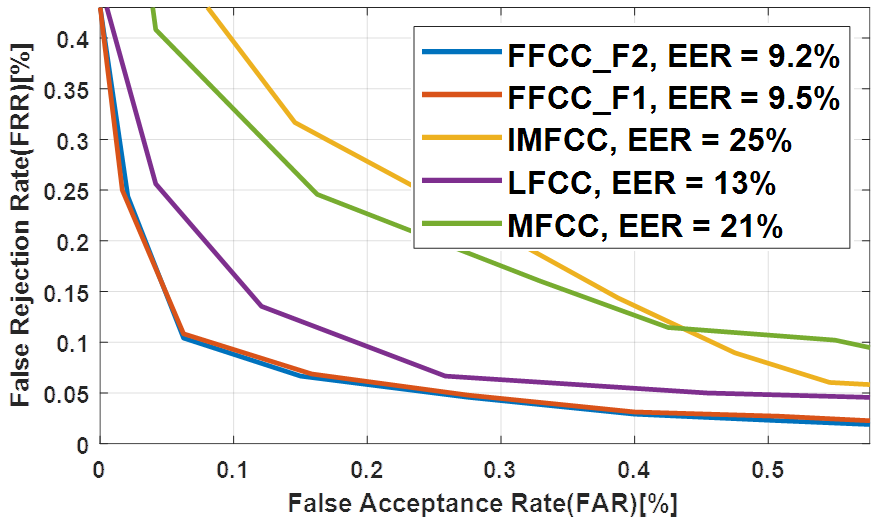
Details of the training sets and testing sets of RSR2015 and SPW



1. SPW Database

From the DET curves, it is evident that the Equal Error Rate (EER) of FFCCs significantly outperforms other feature extraction methods. This indicates that FFCCs offer more accurate and reliable performance in speaker identification tasks. In comparison, the EER of the second-best method, LFCCs, is 3.8% higher than that of FFCCs, further underscoring the superior performance of FFCCs. These results highlight the advantages of using FFCCs, particularly in scenarios where precise speaker recognition is critical.

From FIGURE 9, we can observe the bell curve of FFCCs, showing the EER differences between the first and second formants. Although the FFCCs using the second formant perform better than those using the first formant, the difference in EER is not significant. The EER of FFCCs (first formant) is only 0.3% higher than that of FFCCs (second formant). This may be because both the first and second formant frequencies play a crucial role in speech recognition and speaker identification. When the speech signal within a frame contains both voiced and voiceless features, the bell curve of the first formant tends to shift below 1000 Hz. As a result, the weight enhanced by the voiceless features may be too low, which could explain why the EER of the second formant is slightly better than that of the first formant.



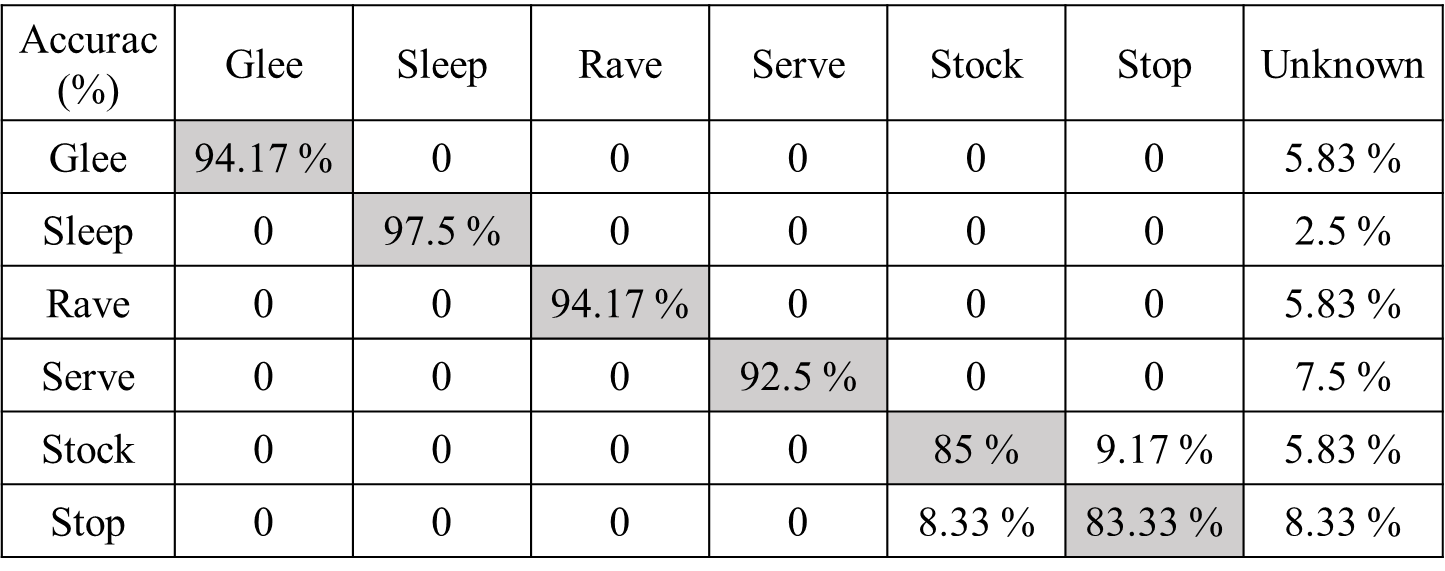
**FIGURE 9. DET curve.**

As previously mentioned, the voiced features of the SPW keywords are very similar, with differences primarily found in the voiceless features. Therefore, compared to MFCCs, which emphasize voiced features, the LFCCs method, which gives equal weight to both voiced and voiceless features, exhibits a better overall recognition rate. Furthermore, while IMFCCs focuses more on unvoiced features, it tends to overlook the crucial voiced features, resulting in a comparatively lower accuracy than other methods.

Finally, we analyze the false positive errors and false negative errors for FFCCs. The testing set of SPW dataset contains 120 speech samples for each keyword, as shown in Table I. The confusion matrix shown in TABLE II is the experimental result of recognizing keywords using FFCCs. It can be seen from this table that FFCCs has good discrimination for keywords with similar features. For example, “glee” and “sleep”, and “rave” and “serve” were not misclassified. Only the keywords “stock” and “stop” are misclassified in few instances. The reason behind this is that some speakers did not pronounce the sound of “k” in stock and “p” in stop clearly.

TABLE II

Confusion matrix of recognizing keywords using FFCCs (second formant)



1. RSR2015 Database

In this subsection, the results of testing with RSR2015 database group II are presented. TABLE IV shows the EER of the four feature extraction methods considered in this work. When we analyze the table, we can find that as the number of testers increases, EER gradually increases. Furthermore, from the results of this experiment depicted in TABLE IV, it is evident that the FFCCs method with significantly lower EER is better than other feature extraction methods.

TABLE IV

EER for four feature extraction methods

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| EER (%) | MFCC | LFCC | IMFCC | FFCC  (F1) | FFCC  (F2) |
| 30 speakers,  10 keywords | 22.14 | 24.61 | 39.41 | 19.87 | **18.78** |
| 40 speakers,  10 keywords | 22.77 | 25.5 | 40.29 | 20.11 | **18.81** |
| 50 speakers,  10 keywords | 24.54 | 26.83 | 41.4 | 21.26 | **19.84** |
| 60 speakers,  10 keywords | 25.72 | 28.49 | 43.47 | 21.91 | **20.57** |

The keywords used in RSR2015 have almost no similar features, and most of them are voiced. Therefore, the experimental results show that MFCCs focusing on voiced features is better than LFCCs, whereas the recognition rate of IMFCCs focusing on unvoiced is the lowest.

From the results obtained from keyword spotting experiments conducted using SPW and RSR2015 datasets as part of this work, it can be found that the selection of keywords will greatly affect the recognition rate. However, the FFCCs method proposed in this paper has strong robustness and hence it can be successfully applied to keyword spotting experiments having even keywords with similar features to achieve good recognition rate.

CONCLUSION

This study presents a new efficient feature extraction algorithm called FFCCs. Experiments carried out in this work based on SPW and RSR2015 database have verified that FFCCs have more comprehensive ability to deal with speaker dependent keyword spotting than MFCCs, IMFCCs and LFCCs. Keyword spotting based on FFCCs exhibited excellent recognition rate even when keywords with similar voiced features are used. Moreover, the experimental results show that keyword spotting based on FFCCs has comparatively very low false positive errors when keywords with similar voiced and voiceless features are used. This indeed verifies the robustness of the proposed FFCCs technique in keyword spotting applications.

REFERENCES

[1] N. Almaadeed, A. Aggoun, and A. Amira, "Text-independent speaker identification using vowel formants," *Journal of Signal Processing Systems,* vol. 82, pp. 345-356, 2016.

[2] M. N. Aulia, M. S. Mubarok, W. U. Novia, and F. Nhita, "A comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system," in *2017 5th International Conference on Information and Communication Technology (ICoIC7)*, 2017: IEEE, pp. 1-5.

[3] M. M. Azmy, "Gender of Fetus Identification Using Modified Mel-Frequency Cepstral Coefficients Based on Fractional Discrete Cosine Transform," *IEEE Access,* 2024.

[4] N. Boualoulou, B. Nsiri, and T. B. Drissi, "Comparative Evaluation of PD Detection Using Deep Learning on IMFCCs Extracted from VMD," *International Journal of Online & Biomedical Engineering,* vol. 20, no. 15, 2024.

[5] J. C. Cavalcanti, A. Eriksson, P. A. Barbosa, and S. Madureira, "Revisiting the speaker discriminatory power of vowel formant frequencies under a likelihood ratio-based paradigm: The case of mismatched speaking styles," *PloS one,* vol. 19, no. 12, p. e0311363, 2024.

[6] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *International Journal of Electronics and Communication Engineering,* vol. 2, no. 11, pp. 2554-2561, 2008.

[7] S. Chakroborty and G. Saha, "Improved text-independent speaker identification using fused MFCC & IMFCC feature sets based on Gaussian filter," *International Journal of Signal Processing,* vol. 5, no. 1, pp. 11-19, 2009.

[8] M. Chougala and S. Kuntoji, "Novel text independent speaker recognition using LPC based formants," in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 3-5 March 2016 2016, in ICEEOT, pp. 510-513, doi: 10.1109/ICEEOT.2016.7755666.

[9] G. Fant, "Acoustic theory of speech production Mouton," *The Hague,* 1960.

[10] A. Hamza *et al.*, "Deepfake audio detection via MFCC features using machine learning," *IEEE Access,* vol. 10, pp. 134018-134028, 2022.

[11] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, "An efficient MFCC extraction method in speech recognition," in *2006 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2006: IEEE, p. 4 pp.

[12] B. Iswanto, H. Hafizhahullah, H. Pardede, and A. Zahra, "The Power-Normalized Cepstral Coefficient (PNCC) for convolutional neural networks-based robust speech command recognition," in *Journal of Physics: Conference Series*, 2023, vol. 2596, no. 1: IOP Publishing, p. 012021.

[13] Y. Jia *et al.*, "The 2020 personalized voice trigger challenge: Open database, evaluation metrics and the baseline systems," *arXiv preprint arXiv:2101.01935,* 2021.

[14] W. Khan and R. Holton, "Decision Support System for Keyword Spotting Using Theory of Evidence," *International Journal of Computer and Electrical Engineering,* vol. 8, no. 1, p. 22, 2016.

[15] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communication,* vol. 60, pp. 56-77, 2014.

[16] H. Lei and E. Lopez, "Mel, linear, and antimel frequency cepstral coefficients in broad phonetic regions for telephone speaker recognition," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[17] M. Müller, "Dynamic time warping," *Information retrieval for music and motion,* pp. 69-84, 2007.

[18] F. Nolan and C. Grigoras, "A case for formant analysis in forensic speaker identification," *International Journal of Speech Language and the Law,* vol. 12, no. 2, pp. 143-173, 2007.

[19] S. Patnaik, "Speech emotion recognition by using complex MFCC and deep sequential model," *Multimedia Tools and Applications,* vol. 82, no. 8, pp. 11897-11922, 2023.

[20] A. Rasheed, "Analyzing the mfcc and gfcc to identify reverberation effects on the sound," *Al-Rafidain Engineering Journal,* vol. 29, no. 2, pp. 148-156, 2024.

[21] M. E. Safi and E. I. Abbas, "Isolated word recognition based on PNCC with different classifiers in a noisy environment," *Applied Acoustics,* vol. 195, p. 108848, 2022.

[22] H. Xing and P. C. Loizou, "Frequency shift detection of speech with GMMs and SVMs," in *2012 IEEE Workshop on Signal Processing Systems*, 2012: IEEE, pp. 215-219.

[23] X. Yixue and K. Jiangping, "Vocal tract characteristic on long-term formant distribution," in *Proceedings of 2012 2nd International Conference on Computer Science and Network Technology*, 2012: IEEE, in Int, pp. 207-211.

[24] D. ZHOU, Y. SHAO, H. ZHANG, H. LONG, and Y. PENG, "An improved GFCC algorithm for language recognition in noisy environments," *Journal of Yunnan University: Natural Sciences Edition,* vol. 46, no. 2, pp. 246-254, 2023.