International Journal of Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2024.8.3

Volume 10, Issue 8
August - 2024

Modulation and Energy Components for Speaker Recognition System

Tariq A. Hassan[1], & Salam A. Hussein[2*]

[1,2]Department of Computer Science, College of Education

Mustansiriyah University, Baghdad

Iraq

ABSTRACT

This paper presents a model use the modulation and energy components for speaker recognition application, that is mainly follows the short-term scenario in speech signal processing, and also introduce a parameter combination that includes the instantaneous components and the energy parameters. This will describe the importance of short-term speech analysis in estimating the modulation parameters and the role of the instantaneous energy in estimating the speaker-dependent parameters. Simply, the short-term scenario is used to, first; avoid the silent and background noise speech portions that present in speech signals, and also to benefit from the stationary concept of the short-term processing in the speech signal. The energy components, on the other hand, are adopted purely in many speech parameterisation models, such as, linear predictive coding (LPC) and Mel-frequency cepstral coefficients (MFCCs). The main ideal of our mixture parameter (or MFCC/AM-FM model) is to determined the extent of these components to contribute together in extracting the parameters that are more related to the speaker more than anything else presented in the speech signal. We evaluated both models using the text-dependent and text-independent speech corpora. The accuracy results show that the frame-based AM-FM model achieve better performance comparing with the traditional structure of the AM-FM modulation model (the model presented in [1]). The MFCC/AM-FM parameters, on the other hand, perform much better, in terms of text-dependent, comparing with the AM-FM parameters and the MFCC parameters. In the case of the text-independent, however, the MFCC/AM-FM model provide better results than the MFCC features but less performance comparing to the AM-FM modulation parameters.

Key Words: AM-FM model, linear predictive coding (LPC), MFCC Features, Signal modulation.

1. INTRODUCTION

In most cases, speaker identification and verification usually constitute the main activities of all speaker recognition models. Basically, speaker identification is the task of assigning an unknown voice to one of the speakers known by the system, while speaker verification refers to the case of accepting or denying a claim to be one particular speaker [2]. Whether the application was identification or verification, the most important part in constructing a speaker model is to find measurable quantities that minimise intra-speaker variability and, at the same time, maximise inter-speaker variability [3]. These materials are then used as speech parameters in a process of code generating in speaker identification or verification tasks.

One recent example of those parameters are the signal modulation components, i.e., the instantaneous amplitude (envelope) and instantaneous frequency (phase). In fact, these parameters came to be an alternative to source-filter model parameters, namely, linear predictive coding (LPC), mel-frequency cepstral coefficients (MFCCs) and perceptual linear prediction (PLP) coefficients. As it is well-known, the source-filter model majorly relies on the energy components of the speech signal, and ignores any other speech components especially those related to the phase parameter [4], [5].

Practically, the AM-FM modulation parameters are actually the only ones that describe the speech signal in terms of its energy information, represented by the instantaneous amplitude, and the its phase information represented by the instanta-neous frequency. The representing of the speech signals by its envelope and phase comes from the desire to

understand the speech signal processing function performed by the auditory periphery, particularly the cochlea. AM and FM signals provide important parameters in understanding the information transmission in the nervous system [6]. The first applying of these parameters in speaker identification is presented by [1]. The modulation model adopted in is actually based on the multiband demodulation analysis (MDA) described in [7] as a method for tracing the speech formants frequency and bandwidths. This model, however, adopted, what we call it as, segment-based AM-FM modulation processing. This is, simply, dealing with the speech signal as one segment, no matter how long this signal is or the amount of low-energy periods that come along with the speech. In this paper, we suggest a new style in dealing with the speech signal in terms of the modulation components estimating. The model that we suggest is the frame-based AM-FM modulation model. This model, however, is simply presented in order to avoid the contribution of the low-energy periods in modulation parameter estimation, and, at the same time, exploit the stationary properties of the speech signal within short time periods. This is archived by cutting the signal into fixed length frames and then applying the filtering and the modulation parameter estimation on each frame of the speech.

In addition to that, we also suggest to use , what we called, the instantaneous energy parameters along with the modulation components. This will give us a chance to appreciate to what extent that the signal energy around each filter channel can contribute along with the instantaneous parameters in identifying the speaker identity.

In the next sections, we will first give a brief of the AM-FM modulation model as presented in [1], then we move to explain our model in estimation the modulation components. After that, we will describe the mixture model (the MFCC/AM-FM model) and give the formula to achieve such combination. Finally, we will present our proposed model and its performance compared with the segment-based AM-FM modulation model presented in [1].

2. THE SEGMENT-BASED AM-FM MODEL

AM-FM speech signal modulation is a procedure that is used to decompose a speech signal into its modulated compo-nents of envelope (instantaneous amplitude) and phase (instantaneous frequency). Subject to nonlinear and time-varying phenomena during speech production (the frequency content changes with time) the AM-FM modulation model comes to seek and exploit the rich set of time-varying frequencies inherent in the speech signal and employed them in many speech signal applications. For example, in [7] AM-FM approach is adopted in the context on formant tracking while in [8] the model is employed in the speech recognition application. In [1] and [9] the AM-FM model is adopted for the purpose of speaker identification, both works used speech signal in the parameters encoding for speaker identification.

In order to characterise a modulation component for a speech signal, a single-valued frequency signal must be generated by breaking the signal down into its frequency components. To do so, a filterbank consist of set of bandpass filters are adopted. The idea is to perform the method of multi band demodulation analysis (MDA) as its described in [7]. We used the Gabor bandpass filters because they are optimally compact and smooth in both the time and frequency domains [10]. This characteristic guarantees accurate amplitude and frequency estimates in the demodulation stage [7]. The analytic signal is then constructed for each bandpass output waveform; it is a transformation of the real signal into the complex domain and it is adopted because it permits the characterisation of the real input in terms of instantaneous amplitude and frequency [11]. Mathematically, given a real input signal s[n], its analytic signal can be computed as

$$sa[n] = s[n] + jH[s[n]] = s[n] + js\hat{}[n]; \qquad (1)$$

where the quadrature signal s^[n] is the Hilbert transform of speech signal s(t). From the analytic and speech signal the phase (t) can be calculated as:

$$\phi[n] = \arctan\left(\frac{\hat{s}[n]}{s[n]}\right).$$

The instantaneous frequency (IF) of the signal can be computed directly from the phase as:

$$f[n] = \frac{1}{2\pi}.\frac{d\phi[n]}{dt}.$$

DOI: 10.31695/IJASRE.2024.8.3

The instantaneous amplitude is computed as:

$$\hat{a}[n] = \sqrt{s^2[n] + \hat{s}^2[n]}.$$

To get the proper values of instantaneous frequency, instantaneous amplitude and instantaneous frequency are combined together to obtain a mean-amplitude weighted short-time estimate Fi of the instantaneous frequency for each bandpass filter output (waveform) [7].

$$F_i = \frac{\int_{n_0}^{n_0+\tau} [f[n].\hat{a}^2[n]]dt}{\int_{n_0}^{n_0+\tau} [\hat{a}^2[n]]dt},$$

where τ is the selected length of the time-frame. The adoption of a mean amplitude weighted instantaneous frequency in is motivated by the fact that it provides more accurate frequency estimates and is more robust for low energy and noisy frequency bands when compared with an unweighted frequency mean [7]. The computation of the short-time instantaneous frequency for each signal leads to the extraction of the speaker descriptors set that used as a system parameter.

This model, however, does not follow any principles in speech signal precessing, it is simply constructed in a way that allows to use the MDA method to generate the pyknogram parameters (originally presented by [7] as a time-frequency representation of the speech signal) and used them for speaker identification. Also, Although is quite obvious that the contribution of the low-energy period have detrimental effect in many speech signal analysis models, however, the authors in did not show that the contributing of the low-energy periods can affect the recognition performance.

Therefore, our aim is to modify the processing structure of the AM-FM model and makes it more appropriate with the general structure of speech signal processing usually adopted in previous models, such as the source-filter model. The next section will explain the main steps of our suggested frame-based AM-FM modulation model.

3. THE FRAME-BASED AM-FM MODEL

This model basically follows the processing style adopted in most speech signal analysis systems. In this model, instead of dealing with the whole signal as one segment (the demodulation method adopted in [1]), the speech signal is initially divided into fixed-length frames. These frames are actually characterised by their own energy which is, in this case, the parameter that we need to decide which frames will be contributed in the demodulation processing.

To be more clear, each frame is regraded as one limited-length speech signal (one unit) and the modulation components extracted from one frame are unrelated to the other frame of the signal. This means, each frame will represent a unique case in the feature vector in terms of the modulation parameters.

In fact, this approach is very similar to the standard procedure used in MFCC encoding, the main difference is the output of each channel is a single-band signal (single-valued frequency signal), and also no discrete cosine transform (DCT) will be applied on the filter channels outputs. From this signal, however, the modulation components (the instantaneous amplitude and frequency) will be estimated and, from these components, the mean amplitude weighted instantaneous frequency (equ. 5) is calculated.

Figure 1 explain the block diagram of our main step of the frame-based AM-FM modulation model. The signal is first divided into fix-length frames (25ms in our experiment with 10ms overlap) then energy value of each frame is calculated. If the energy value is equal to, or more than, the threshold (0.1 in our experiment) then the segment is taken, otherwise it is regarded as a silence segment and no more process is done on it. A high emphasis filter (1 0:98Z 1) is applied to the accepted segments (in order to balance the signal spectrum) and a 25 ms Hamming window is applied to the emphasised speech every 10 ms.

The speech segments are then passed through a set of Gabor bandpass filters (30 channels in our experiment) with center frequencies that are mel-spaced on the frequency axis with critical bandwidth computed through the equation:

$$Bw(n) = 25 + 75\ 1 + 1{:}4(f\ c(n)=1000)2\ \ 0{:}69 : \qquad (6)$$

The idea of non-uniform frequency analysis comes from the actual processing of the speech signal by the basilar membrane inside the human ear. In this case, the frequency responses overlap significantly since points on the basilar membrane cannot vibrate independently of each other. The critical bandwidth, on the other hand, has been defined and measured using a variety of methods, showing that the effective bandwidths are constant at about 100 Hz for center frequencies below 500 Hz, and with a relative bandwidth of about 20% of the center frequency above 500 Hz [12].

DOI: 10.31695/IJASRE.2024.8.3

For each bandpass filter channel output, an analytic signal is constructed by Hilbert transform of real-valued signal, the instantaneous amplitude and frequency are then computed directly through the equations (3). Instantaneous amplitude and frequency are combined together to obtain a mean amplitude weighted instantaneous frequency and encoded them as parameters for speaker identification.

After the demodulation analysis, the feature vectors of each speaker can be represented by an N F matrix, where N corresponds to the number of Gabor filters used in the filterbank and F corresponds to the number of frames that have energy equal to or more than the energy threshold of 0.01. For the decision making step, GMM classification is adopted with the number of component densities in the mixture model chosen depending on the identification task (text-dependent or -independent).

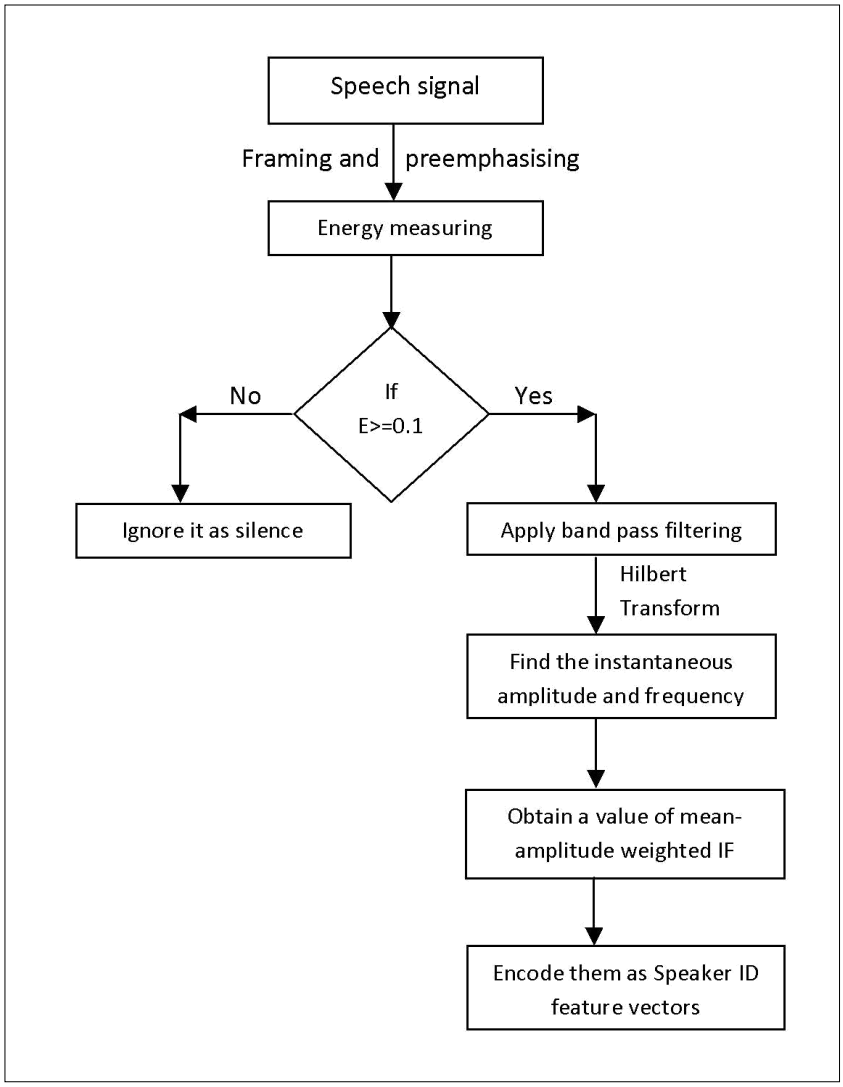

Figure 1. Block diagram of our suggested method.

4. THE FRAME-BASED AM-FM MODEL AND MFCC FEATURES COMBINATION

In an attempt to generate a new set of speaker identification features, [13] adopted the AM-FM model along with LPC analysis to produce a feature set derived from both models. The LPC/AM-FM model, as they called it, is used to extract features that are described to be robust over a degraded channel. In this work, the LPC analysis is applied to estimate the location of the speech formants. This estimates the spectral envelope of the speech signal, and ignores the source features present in the speech. The AM-FM parameters are used to generate three waveforms for each single-band signal. These waveforms represent; the energy, the envelope (the instantaneous amplitude), and the phase (the instantaneous frequency) for each formant estimated in the previous step.

DOI: 10.31695/IJASRE.2024.8.3

In our work however, we also try to benefit from the idea of mixing features from more than one model of the speech signal in an attempt to improve the performance of the speaker identification system. Our proposed idea is to make some type of mixture between the spectrum parameters that are usually adopted in MFCC and LPC analysis and the modulation parameters of the AM-FM model.

In this approach, which we call the MFCC/AM-FM model, instead of only estimating the modulation components of each single-band signal (frame) the energy components are also computed. The both parameters (the modulation and the energy parameters) are then computed together to generate one parameter values represent the effect of the modulation parameters and the energy around each filter channel in the system.

The discrete cosine transform (DCT) is applied in order to obtain the instantaneous parameters affected cepstral coef-ficients, which is simply represent three combined components (the instantaneous amplitude, the instantaneous frequency and the signal energy at one specific time). That is, the speech signal will be represented by an N C matrix of features, where N is the number of channels of the Gabor filterbank and C is the number of DCT coefficients. Mathematically, the MFCC/AM-FM process can be described as:

$$En_i = \sum_{t=1}^{\tau} s_n(t)^2$$

$$EnF_i = En_i \times \left[\frac{\int_1^{\tau} [f(t).\hat{a}^2(t)]dt}{\int_1^{\tau} [\hat{a}^2(t)]dt} \right]$$

5. SPEECH DATABASE

Two speech corpora were used to evaluate our proposed model. The first is the text-dependent British Telecom Millar database, specifically designed and recorded for text-dependent speaker recognition research. This data set consists of 60 (46 males and 14 females) native English speakers saying the digits one to nine, zero, nought and oh obtained in five different sessions over a time span of about three months. Each speaker in the corpus participated in five recording sessions. The first and second recording sessions (10 repetitions) were used for training while the rest of the data (15 repetitions) were used for testing.

The second data set is the NTIMIT corpus [14]. A subset of the NTIMIT database is used in our work. This subset (38 speakers from the New England dialect region) contains 10 sentences of read speech for each speaker. The male subset contains 24 speakers, while the female subset contains 14 speakers.

6. RESULTS

To evaluate how the framing of the speech signals affects the speaker identification performance, we compare the per-formance of our proposed method (frame-based AM-FM model) with the traditional AM-FM modulation model presented in [1]. The two approaches are evaluated using both the text-dependent (BT-Millar) and text-independent (NTIMIT) speech corpora.

In the case of text-dependent evaluation, the assessment includes the original, clean tokens of the BT-Millar database and also noisy versions of these corpus(to which Gaussian white noise has been added at 40% SNR. This is done because noisy speech is more realistic for an application, but also explicitly to see the effect of noise on both performances. In the case of text-independent evaluation, however, we use the telephone-quality NTIMIT corpus as it stands, because realistic noise is already presented in the speech.

In both cases, GMM is used with a diagonal covariance matrix. This choice is based on the empirical evidence that diagonal matrices outperform full matrices and the fact that the probability density modeling of an Mth-order full covariance matrix can equally well be achieved using a diagonal covariance lager-order mixture [15]. the number of

DOI: 10.31695/IJASRE.2024.8.3

component densities in the mixture model is chosen to be between (7-15) Gaussians. Our experiment shows that increasing the number of Gaussians beyond this values does not improve the performance of the classifier with AM-FM features.

In terms of the instantaneous frequency parameters, the text-dependent speaker identification results are shown in Figures 2, 3. These Figures are describe the recognition results of both frame-based and segment-based demodulation approaches for text-dependent speaker identification using both the clean and the noisy speech data of the Millar corpus.

From these figures, we can notice that the frame-based AM-FM model can achieve very much better performance compared with the segment-based model. Applying a one-tail t-test of significance (two-sample, equal variance) to the results for all 60 speakers yields, in case of clean database, a probability of p = 3:28 10 13, while in the case of noisy database a probability of p = 3:21 10 12 that the differences in mean accuracy between our method and the segment-based method.

Figure 4 shows the recognition results of the text-independent speaker identification of the frame-based model and segment-based model. As in the case of text-dependent identification, our proposed method out-performs Grimaldi and Cummins' method. A one-tail t-test of significance (two-sample, equal variance) applied to the results for all 38 speakers gives a probability of p = 1:12 10 6 that the differences in mean accuracy between our method and the segment-based method.

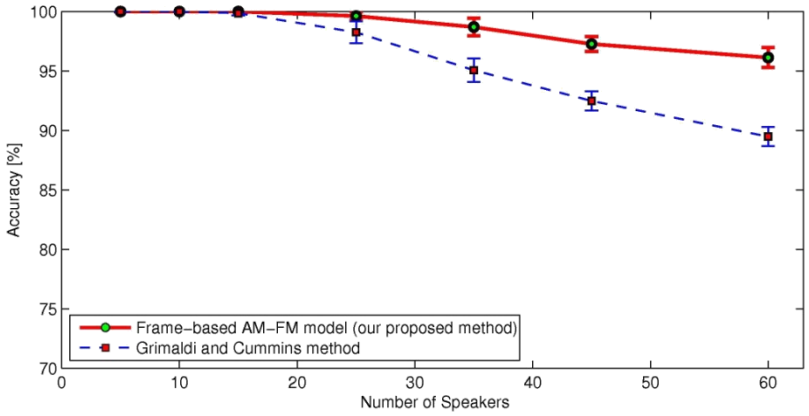

Figure 2. The proposed AM-FM frame-based

Figure 2. Text-dependent speaker identification accuracy results of clean Millar database with mel-scaled centre frequency and bandwidth and frequency range of 0–8 kHz. Error bars are standard deviations across the 10 folds of training/test data.

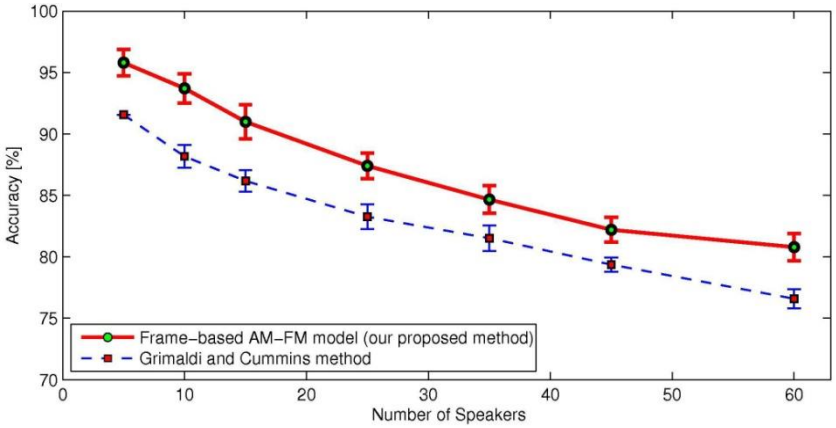

Figure 3. Grimaidi method

Figure 3. Text-dependent speaker identification accuracy results of noisy Millar database with mel-scaled centre frequency and bandwidth and frequency range of 0–8 kHz. Error bars are standard deviations across the 10 folds of training/test data.

DOI: 10.31695/IJASRE.2024.8.3

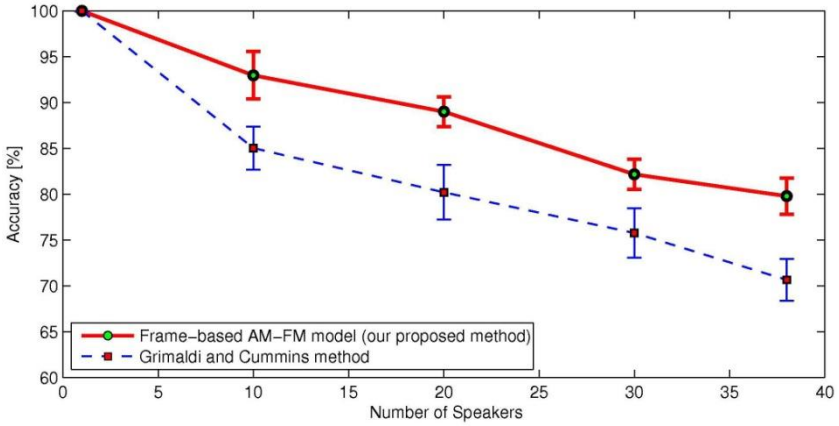

Figure 4. Cummins method

Figure 4. Text-independent speaker identification accuracy results of telephone quality NTIMIT database with mel-scaled centre frequency and bandwidth and frequency range of 0–8 kHz. Error bars are standard deviations across the 10 training/test folds.

Figures 5, 6, 7 are presenting the MFCC/AM-FM model performance in terms of text-dependent (clean and noisy) and text-independent speaker identification. The figures show the recognition results for three models parameters (instantaneous frequency, instantaneous amplitude, and the MFCC coefficients). From these figures we can see that the performance of the MFCC/AM-FM parameters are varying depends on the speech data type (dependent and independent). In the context of text-dependent (clean and noisy), it is clear that our MFCC/AM-FM method achieves better performance than traditional MFCC method. Applying a one-tail t-test of significance (two sample, equal variance) to the results for all 60 speakers yields a probability of p = 0:006 (clean), p = 9:9 10 9 (noisy) that the differences in mean accuracy between our MFCC/AM-FM and the MFCC method. In the context of text-independent, however, the AM-FM model gives better results comparing with the other models. This might, as we think, there is a much bigger performance penalty to introducing MFCC parameters for the text-independent NTIMIT database than for the text-dependent Millar database. Hence, we might expect the use of MFCCs along with AM-FM parameters to perform less well than the frame-based AM-FM model, which uses modulation parameters only.

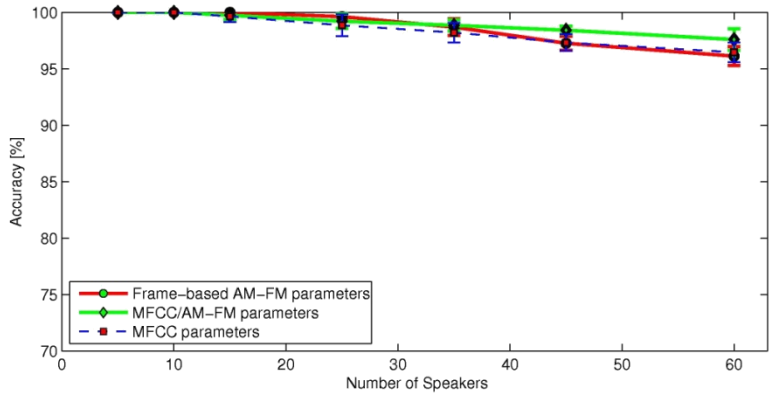

Figure 5. MFCC/AM-FM method

Figure 5. Performance of the MFCC/AM-FM method relative to the frame-based AM-FM and MFCC approaches for text-dependent identification using clean Millar speech with mel-scaled centre frequency and bandwidth and frequency range of 0–8 kHz.

DOI: 10.31695/IJASRE.2024.8.3

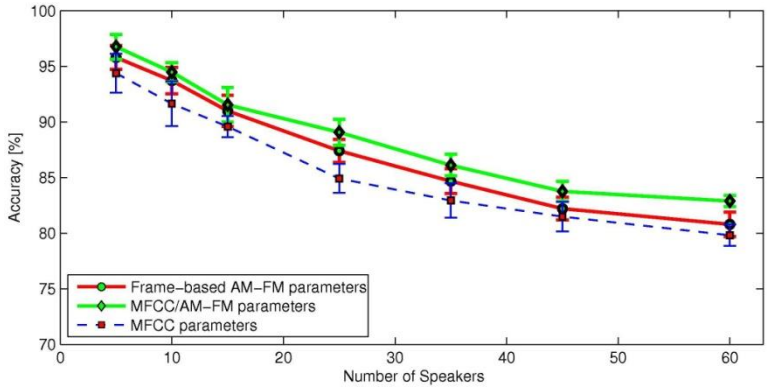

Figure 6. Frame-based AM-FM and MFCC

Figure 6. Performance of the MFCC/AM-FM method relative to the frame-based AM-FM and MFCC approaches for text-dependent identification using noisy Millar speech with mel-scaled centre frequency and bandwidth and frequency range of 0–8 kHz.

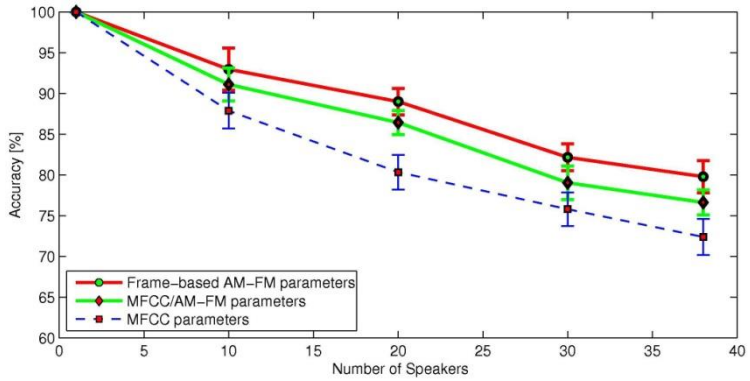

Figure 7. AM-FM, MFCC parameters

Figure 7. Text-independent speaker identification results for frame-based AM-FM parameters, MFCC parameters, and the MFCC/AM-FM parameters using the NTIMIT database with mel-scaled centre frequency and bandwidth and frequency range of 0–8 khz.

7. CONCLUSION

From this work, we can conclude that the adopting of frame-based AM-FM modulation model can achieve two important points. First, this model allows only these parts of speech signal with reasonably high energy (0.01 in our experiment) to contribute in the modulation components estimating. This means, the will only deal with those parts that hold a useful information about the speaker. The second benefit is the importance of the short-time stationary assumption of the speech signal. Simply, the fixed-length frames are tending to be quasi-stationary in terms of its harmonics (excitation frequencies) and resonance (formant frequencies) [16, Sec. 5.3].

The accuracy results confirms the important role of the short-time stationary aspect of the speech signal in estimating the modulation parameters with high accuracy compared with long-term processing of speech. Also, the contribution of the energy parameter can improve the recognition performance, especially in the case of text-

DOI: 10.31695/IJASRE.2024.8.3

dependent identification. In the text-independent, however, it seems the the effect of the energy parameter does not help that much, but it still give us good results comparing with the traditional MFCC method.

REFERENCES

[1] M. Grimaldi and F. Cummins, "Speaker identification using instantaneous frequencies," IEEE Transactions on Audio, Speech, and Language Processing, vol. 16, no. 6, pp. 1097–1111, Aug. 2008.

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

[3] Tony Brian Alderman, Forensic speaker identification: a likelihood ratio-based approach using vowel formants, Lincom Studies in Phonetics, LINCOM, Munich, Germany, 2005.

[4] M. Foundez-Zanuy, S. McLaughlin, A. Esposito, A. Hussain, J. Schoentgen, G. Kubin, W. B. Kleijn, and P. Maragos, "Nonlinear speech processing: Overview and applications," Int. J. Control Intelligent System, vol. 30, no. 1, pp. 1–10, 2002.

[5] Mohammadi Zaki, J Nirmesh Shah, and Hemant A Patil, "Effectiveness of multiscale fractal dimension-based phonetic segmentation in speech synthesis for low resource language," in International Conference on Asian Language Processing (IALP), 2014. IEEE, 2014, pp. 103–106.

[6] S.M. Khanna and M.C. Teich, "Spectral characteristics of the responses of primary auditory-nerve fibers to frequency-modulated signals," Hearing Research, vol. 39, no. 1-2, pp. 159 – 175, 1989.

[7] Alexandros Potamianos, Ros Potamianos, and Petros Maragos, "Speech formant frequency and bandwidth tracking using multiband energy demodulation," J. Acoust. Soc. Amer, vol. 99, pp. 3795–3806, 1996.

[8] Niko Moritz, Jorn¨ Anemuller,¨ and Birger Kollmeier, "An auditory inspired amplitude modulation filter bank for robust feature extraction in automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, no. 11, pp. 1926–1937, 2015.

[9] Karthika Vijayan, Pappagari Raghavendra Reddy, and K Sri Rama Murty, "Significance of analytic phase of speech signals in speaker verification," Speech Communication, vol. 81, pp. 54–71, 2016.

[10] Ibrahim Missaoui and Zied Lachiri, Gabor Filterbank Features for Robust Speech Recognition, pp. 665–671, Springer International Publishing, Cham, 2014.

[11] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," IEEE Transactions on Speech and Audio Processing, vol. 8, no. 3, pp. 240–254, May 2000.

[12] E. Zwicker and H. Fastl, Psychoacoustics : facts and models, Berlin ; New York : Springer-Verlag, 1990.

[13] R. Jankowski, T. Quatieri, and D. Reynolds, "Measuring fine structure in speech: application to speaker identification," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP'95), May 1995, vol. 1, p. 325.

[14] C. Jankowski, A. Kalyanswamy, S. Basson, and J. Spitz, "NTIMIT: a phonetically balanced, continuous speech, telephone bandwidth speech database," White Plains, NY, apr 1990, vol. 1, pp. 109–112.

[15] D.A. Reynolds and R.C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," Speech and Audio Processing, IEEE Transactions on, vol. 3, no. 1, pp. 72–83, Jan 1995.

[16] J. D. Jr., J. Hansen, and J. Proakis, Discrete-Time Processing of Speech Signals, second ed., IEEE Press, New York, 2000.

DOI: 10.31695/IJASRE.2024.8.3