Exploiting Signal Linear Trend in the Time Domain to Enhance Speech Feature

HAO DUC DO
Computing Fundamental Department, FPT University, Ho Chi Minh City 70000, Vietnam
e-mail: haod3@fpt.edu.vn

This work was supported by Vingroup Joint Stock Company through the Ph.D. Scholarship Program of Vingroup Innovation Foundation (VINIF), Institute of Big Data, under Grant VINIF.2021.TS.120.

ABSTRACT Speech feature extraction usually begins with transforming the signal from the time domain to the frequency domain via integral transforms. Although human speech contains sine-shape waves and linear trends in the time domain, the linear trend is usually ignored. This research proposes two new methods to strengthen the speech feature vector using these unused elements. Before transforming the speech frames into the frequency domain, we use linear regression to identify each speech frame’s linear trend or linear envelope in the time domain. Then we remove the impact of that trend to normalize the signal and emphasize the stationary elements in the signal. The proposed feature vector includes parameters from the linear envelope and the conventional vectors as a spectrum or MFCC. Our new features not only emphasize the stationary property in the speech signal, but also improve the result for speech recognition in terms of error rate. Experimental results demonstrate that the impact of the linear envelope is significant and the linear envelope subtraction is a meaningful stage.

INDEX TERMS Speech feature, envelope subtraction, linear regression, speech recognition, integral transform.

I. INTRODUCTION Naturally, human speech is created by the vibration of the vocal tract [1], so it is naturally the stationary signal. Hence, the speech behaviors are presented as more informative in the frequency domain than in the time domain. In most of the research on speech processing, the Fourier transform and its variants are used to convert the speech from the time domain into the frequency domain. With this approach, some well-known speech features were introduced, such as spectral, cepstral [2], [3], [4], Mel cepstral [5], [6], MFCC [7], [8]. Because the human brain understands the properties of speech in the frequency domain, these features are widely used and gain many possible results.

When the speech is released out of the mouth, it is not pure stationery because the tongue, lips, and the outside environment impact it. This leads to the fact that the signal cannot be presented perfectly in the frequency domain, or the Fourier transform cannot analyze the signal into frequency elements.

Fourier transform, naturally, is an approach to separate a mixture into its periodic elements, which are the sinusoidal waveforms. When the speech signal contains at least one non-stationary element, for instance, a linear element, the Fourier transform’s separation process is inefficient, and then the results contain much noise.

This research first shows a non-stationary element in human speech and then suggests a new approach to design a new and more effective feature representation. We mainly focus on the non-stationary signal caused by linear function because this kind of signal commonly exists in everyday speech datasets. These signals include the envelopes, which present the main trends in the signals, and the details, which present the preliminary information. We apply a linear regression algorithm to identify the signal envelope in the first step. We then integrate these parameters into the feature vectors. After that, the vector representing speech is extended and holds more meaningful information.

Our experimental results have demonstrated two crucial observations. Firstly, compared to the accuracy of the speech recognition model for the same kinds of features,
the performance gained by features with envelope trend is better. We get the best results when combining the MFCC with the signal envelope via our proposed ES MFCC method. Secondly, the results are equivalent to many speech features, so we can conclude that our proposed methods are effective and stable in applications. 

The main contributions of this research are in three aspects:

- Demonstrating the negative impact of the speech signal linear envelope in the feature extraction process and the quality of the extracted feature.
- Proposing two new algorithms for speech signal feature extraction and then applying them to speech recognition with an improved result.
- Demonstrating and analyzing the contribution of each main phase in the proposed algorithms, mainly focusing on getting coefficients for the linear envelope and envelope subtraction.

The remaining of this work is organized as follows. Section II briefly mentions some related speech feature extraction methods. This section concerns the conventional extracting process, which motivates us to conduct this research. Section III demonstrates the appearance and the negative impact of a new element, named linear envelope, on the signal transform process. Section IV shows our proposed method, including two algorithms, to extract the efficient feature vector to capture the envelope information. Two main experiments in section V are the experimental demonstration of our observations and suggestion. Next, section VI discusses the problem and contributions of this research in detail. The paper ends with section VII, summarising and concluding the proposed approach.

II. RELATED WORKS

The speech signal is one of the first and most common researched objects of artificial intelligence and machine learning. Many features were proposed for speech research during the development of approaches and applications. This section would like to look at some widely used features briefly.

A. MAIN METHODOLOGIES IN SPEECH FEATURE EXTRACTION

The first group of speech signals relates to prosodic features [9]. This group focuses on increasing or decreasing the tone and pitch of human speech. These features can be applied to phonemes, words, phrases, or whole sentences. In signal processing, the most popular prosodic features include the pitch or frequency \( f_0 \), energy, and duration of the signal [10], [11], [12].

The vibration of the vocal tracts creates frequency \( f_0 \). This value presents most human speech’s content, prosody, and many acoustics properties. Mutual with its values, the shift of frequency \( f_0 \) during the time can reflect some critical information that is meaningful in some applications. On the other hand, signal energy is helpful in many cases. Because the energy relates to the amplitude of the speech sound, it directly impacts the human brain’s perception of sounds. Energy also depends on the creation of human speech, such as emotion or health, being advantageous for many applications related to speech. Similar to energy, sound duration reflects the general features of human speech. In this context, sound duration means the proportion of speech signal with the whole signal. A signal with a low duration is a speech with many walls of silence, while a signal with a higher duration implies a continuous speech with fewer silence ranges. This feature is also impacted by emotion or health as energy.

Many researchers combined the features with some statistics in some works related to prosodic features to form the completed features vector for speech [13]. These values included the mode frequency (the frequency with the highest energy) and the ranges of frequency corresponding with 25%, 50%, 75% energy of the signal. These approaches were usually applied in machine learning models, including SVM, HMM, or GMM, to gain acceptable results.

The second group mainly applies the integral transform to represent the data into a new space with a more informative form [14]. This group significantly improved by adding a proportional function into the transform kernel. Many well know in this group include Wavelet, Chirplet, and some extensions of Fourier transform [15], [16]. They would choose the most suitable kernel for the application based on the data differences and particular purposes.

In essence, after being processed by an integral transform, the signal spectrum is kept transforming into a more informative form. This is because many prosodic features, such as prosody, gender, dialect, and emotion, significantly depend on the spectral features of the speech signal. Analyzing the speech spectrum helps to effectively model and exploit the speech data. One of the first approaches to analyze the spectral is to convert the spectrum to Mel-scale via Mel Frequency Cepstral Coefficients (MFCC) [17]. On the other hand, many similar approaches were usually used, including Linear Prediction Cepstral Coefficients (LPCC), Log-Frequency Power Coefficients (LFPC), or Gammatone Frequency Cepstral Coefficients (GFCC) [18], [19]. These methods share the same ideas of separating and grouping the meaning in the frequency domain. The differences make these features strong in a different context, and each is only suitable for a group of particular applications.

The third group to represent speech features is to describe the signal quality. The most common factors of this group are jitter, shimmer, and Harmonic-to-Noise Ratio (HNR) [20]. Jitter and shimmer present the fluctuating ranges of frequency \( f_0 \) and signal amplitude from one cycle to the next. Hence, the meaning that these two factors hold is signal stability. HNR implies the ratio of noise in the speech signal. Some other features presenting the signal quality include Normalized Amplitude Quotient (NAQ), Quasi Open Quotient (QQQ), Maxima Dispersion Quotient (MDQ), and Parabolic Spectral Parameter (PSP) [21].
In addition, this group also includes many works which show the properties of speech sounds such as harsh, tense, breathy, modal, whisper, creaky, lax-creaky [22]. These factors have a close relationship with emotion or health. This is based on the properties above relating to the combination of frequency $f_0$ and the energy in the signal.

The fourth group turns around Teager Energy Operator (TEO features). This group includes the speech features researched by Teager [23] and some variations proposed by Kaiser [24]. TEO features use some nonlinear functions to approximate the effects caused by articulating organs (teeth, lip, tongue) to the airflow. Then the parameters of the approximation function are used as the feature vector for the released speech sound. After 2000s, some extended feature were proposed such as TEO-FM-Var, TEO-Auto-Env, TEO-CB-Auto-Env [25], [26], [27]. Finally, because of the similarity between these features and some other groups, such as energy or spectrum, TEO features were usually combined with many other vectors to form the final feature vector to improve the performance of the processing model.

The final and newest group uses a deep neural network as a feature extractor. The model belonging to this group is fed by a raw speech signal called waveform, then processed through many layers to return the vector [28], [29], [30]. Although the meaning of the vector extracted by a neural network is not interpretive (by a human), it plays a positive role in improving the whole model’s performance. This can explain why a neural network can extract the idiosyncratic vector in a hidden space. This feature cannot be visible in the data space but is distinguished in the hidden space. This behavior uses features extracted by neural networks to help present data in a better space.

### B. SPEECH FEATURE REPRESENTATIONS: ADVANTAGES AND DISADVANTAGES

Throughout history, many speech feature extractors have been invented and used. However, only a few of them continue being used in research and industry.

The prosodic features, signal quality features, and TEO features do not belong to the mainstream speech feature development. These features are good at computational complexity and contain a piece of fixed information. For example, $f_0$, a well-known representative from the prosodic group, presents the fundamental frequency of the speech signal. This feature could be good in a particular application but not good enough for speech signal understanding in general. Similarly, energy, HNR, or TEO operator just provide a piece of information encoded in human speech. That is why these features could not stand alone in speech processing and are usually used as an additional feature to enhance the other.

In the era of deep learning, many neural network architectures were proposed for speech processing and understanding. Some of them, such as Tacotron or Wave2Vec, have become state-of-the-art in TTS and ASR tasks. Although gaining a high performance in both computing and accuracy, deep neural networks have two main problems. The first is that the data volume demand for training is enormous. For instance, in speech recognition, the HMM-GMM model needs around 200 hours to gain a good model, but Wave2Vec or Jasper needs at least ten times that volume to gain the same accuracy. The large volume of data also requires a vast computing system to train the network, so small research groups have many difficulties reproducing the report results. The second problem of a deep neural network is understandable or interpretable. While $f_0$ or total energy has a close relationship with the nature and origin of speech, deep neural networks are black boxes and cannot be understood. Although gaining many good results, based on the above properties, in-depth features from neural networks are disbelieved by many researchers.

The remaining group, integral transform, is a classic approach but is still commonly used in modern research. Spectrogram and MFCC have become two of the most widely used feature representations in current research. These methods have a low-cost computation and are simple to implement. The main strength of integral transform origins from its mathematical foundation. Spectrogram or MFCC transforms the signal from the time domain into another domain named the frequency domain and applies some techniques to enhance the feature. The signal can show its properties, structure, and energy distribution in this new domain. This frequency domain provides a better view of the speech signal with more directions and more information.

### C. THE IDEA WITH SIGNAL ENVELOPE SUBTRACTION

It is easy to observe that after 2010, the amount of research related to building a new feature extractor or new data representation was reduced. One of the most significant causes is the popularity of the deep neural network model with the ability to learn how to represent the raw data in a better form. Although some researchers can reach state-of-the-art solutions, the model and the features of a deep learning model are non-interpretable. It is challenging to make sure that the current best solution can work well in the future. No one can state the reliability of a deep model and its extracted vector. This leads to the demand that the traditional approach, for example, the integral transform, should be improved and extended to establish a more effective extractor.

Excepting the vector from a deep neural network, most of the speech features are established as follows:

- Applying an integral transform to transform the speech signal into another space.
- Designing some extended operators to process data in the transformed spaces such as Mel scale, Filter bank, or Statistics.

The feature-extracting process always begins with an integral transform. This means that the suitability of the integral transform to the speech signal defaults as right. The problem is whether or not the integral transform should be applied to the signal without any involvement. An integral transform is a traditional tool in mathematics with hundreds
of years of history, so it is hard to modify or extend this side. On the other hand, due to the popularity of human speech, the variety of speech signals is wide too. There can be a significant ratio of human speech distorted. These data cannot be transformed directly via some traditional transforms. They should be pre-processed or enhanced before passing into the transformation process.

III. IMPACTS OF NON-STATIONARY PROPERTY TO SIGNAL PRESENTATION IN THE FREQUENCY DOMAIN

A. HUMAN SPEECH AND ITS PRESENTATION IN THE FREQUENCY DOMAIN

Human speech comes from the vibration of the vocal tract, so it can be seen as a stationary signal [31], [32]. This motivates many researchers that use integral transforms to analyze speech signals into a collection of basis components such as sine shape waves. Transform methods represent the signal into a new form called spectrum or representation in the frequency domain. Because most of the essential properties of human speech are presented clearly in this new domain, it has become one of the most commonly used forms to present the speech signal. With a problem needing stationary in a short range, for example, a speech recognition task, this approach promotes efficiency because it specifies the contribution of each sine wave in the speech.

B. NON-STATIONARY SIGNAL IN THE FREQUENCY DOMAIN

On the other hand, many applications related to the change of speech tone of other properties, such as gender recognition, age estimation, emotion recognition, and health status prediction, need to be exploited more than using a collection of sine waves. Besides, places and manners of articulation impact the airflow when producing speech, transforming the speech signal into a non-periodic signal. It can be assumed that the unique properties of recognizing the genders, ages, dialects, and emotions are encoded in these non-repeated elements.

Mathematically, with a non-repeated signal, applying an integral transform to a signal to form an effective feature vector may not be actual. This method can cause some loss in information at the feature vectors, mainly the information presenting the change in speech tone.

A repeated signal is usually decomposed as a sum of sine functions. To explore the impact of a non-repeated element on the transformation process of a signal, let us consider the simplest case where the signal is the sum of a harmonic function and a linear function as below:

\[ f(x) = \sin(3x) + x, \quad x \in [0, 2\pi] \]  

The function \( f(x) \) is illustrated in figure 1 with blue. Generally, the signal main trend is a line, presented by red, while the signal detail is a sinusoidal wave, presented by orange. If we apply an integral transform to \( f(x) \), the coefficients in the frequency domain can be some noise values. Supposing that we use \( b(x) = \{b_1(x), b_2(x), b_3(x), b_4(x), b_5(x)\} \) = \{\sin(x), \sin(2x), \sin(3x), \sin(4x), \sin(5x)\} as the basic vector set for destination space. The transform is designed as follows:

\[ c = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(x)b(x)dx \]  

or:

\[ c_i = \frac{1}{2\pi} \int_{0}^{2\pi} f(x)b_i(x)dx, \quad i \in \{1, 2, 3, 4, 5\} \]  

The corresponding coordinate \( c \) of \( f(x) \) in this space is computed via equation 3, and then the result is \( c = (c_1, c_2, c_3, c_4, c_5) = (-1, \frac{1}{3}, \frac{1}{6}, \frac{1}{2}, \frac{1}{3}) \). This 5-dim vector presents the results of five projections of \( f(x) \) to the set of five basic vectors. Let \( f_1(x) = \sin(3x) \) and \( f_2(x) = x, f(x) = f_1(x) + f_2(x) \). In \( f(x) \), stationary property originates from \( f_1(x) \) while the linear property of \( f_2(x) \) causes the non-periodic property for \( f(x) \). Back to the nature of \( f(x) \), the stationary element mainly occurs at frequency 3Hz, corresponding with \( f_1(x) = \sin(3x) \). Function \( f(x) \) does not contain any elements related to frequency 1Hz, 2Hz, 4Hz, or 5Hz. On the other hand, all values in the coordinate vector \( c \) are non-zero. This fact means the linearity of \( f_2(x) = x \) creates the noise values for the coordinate values at the basic vector corresponding with frequency 1Hz, 2Hz, 4Hz, or 5Hz.

To verify the rightness of the feature vector \( c \), it would be transformed into the time domain via the inverse transform. Basically, the transform below is a way to separate the function \( f(x) \) into the five elements \( b_i(x) \) and the coordinates \( c_i \) are the five corresponding weights. So the calculation process to reconstruct the function \( f(x) \) is described as follows:

\[ \tilde{f}(x) = \sum_{k=1}^{5} c_k \sin(kx), \quad x \in [0, 2\pi] \]  

The illustration for the function \( \tilde{f}(x) \) is shown in figure 1 with green. The difference between function \( \tilde{f}(x) \) and \( f(x) \) is significant. This fact demonstrates that the transform cannot process the non-periodic signal well. Function \( f(x) \) contains only one sine-shape wave, \( f_1(x) = \sin(3x) \), and a linear function \( f_2(x) = x \). Non-zero values in \( c_1, c_2, c_4 \) and \( c_5 \) are all caused by function \( f_2(x) = x \). Because these values are not true, the quality of feature vector \( c \) is downgraded a lot.

On the other hand, let us consider a traditional Fourier transform method to compare the reconstructed result with the original signal.

Figure 2 shows the spectra of a sine-shape signal, a linear signal, and the summing-up signal. From these sub-figures, their properties are shown as follows:

- Sine signal \( y = \sin(3x) \): the most information concentrates around the value \( \frac{1}{3\pi} \). This value is the frequency of the stationary signal above.
- Linear function \( y = x \): the most coefficients gather the value 0. This is led by the fact that the signal is pure non-stationary. A linear line is a periodic function that does not repeat anywhere, so small non-zero frequency coefficients are.
H. D. Do: Exploiting Signal Linear Trend in the Time Domain to Enhance Speech Feature

FIGURE 1. The time domain representation for four signals: sine wave, linear line, sum-up signal, and reconstructed signal via the inverse Fourier transform. It can be observed that the sum-up signal mainly depends on the linear function, while the reconstructed signal distributes similarly with the sine shape.

- Mixture signal $y = \sin(3x) + x$: The spectrum is generally similar with the spectrum of $y = x$. The primary distributions are high at zero and reduce smoothly when the frequency increases. The distinguished property here is that two values near 0.5 are nearly equal. Because this signal contains $y = \sin(3x)$, which gets the highest spectrum value at frequency $\frac{3}{2\pi} \sim 0.5$, the coefficients around frequency 0.5 are nearly stable.

On the other hand, figure 3 presents the spectrum of the reconstructed signal with the Inverse Fourier Transform. There are two important observations with this spectrum:
- The highest coefficient is more similar to the linear function than the sine function.
- The stationary information, which encoded in the sine wave, are mostly disappeared.

These two observations lead the spectrum of the sum-up signal to be distorted and biased to the linear elements. Moreover, the meaningful information in the repeated elements is ignored too. This motivates us to analyze the linear trends in the signal and their impacts on the spectrum.

C. SPEECH SIGNAL LINEAR ENVELOPE IN THE TIME DOMAIN

As briefly mentioned in the introduction, human speech is not a pure stationary signal. This causes a big problem with the rightness of the extracted feature vector. Most of the current speech feature extraction processes, from simple features such as spectral or cepstral to more complicated features such as Mel-cepstral or MFCC, always start with an integral transform to represent the speech into the frequency domain. The complicated features are only the post-processing or a chain of transforms resulting from the beginning integral transform. The problem is whether the final feature vector is correct if the first transform’s quality is not guaranteed. The answer is naturally not. A non-repeated signal, after an integral transform with a sine-shape basis, can contain noise and loose information in the frequency domain.

The human brain is most sensitive to the frequency of speech. This means that the most critical information humans receive and understand is the frequency or periodic property. In many cases, when linear elements impact the speech, as in figure 1, the coordinate of speech in the frequency domain is noise. If we pass these coefficients to a machine learning model, it cannot process well. So one of the main questions in this research is what the linear elements are and how to process them. These elements are described in the definition 1.

Definition 1: Linear envelope of a signal is the main trend of this signal in the time domain. It is also the line that approximates the signal with the least error.

In figure 1, the signal $f(x) = \sin(3x) + x$ with blue has the linear envelope presented by red, or $f(x) = x$. The red line does not fit with all blue signal values but reflects the main trend of the signal. Next, more practically, because the linear envelope’s nature is the line approximating the signal, it can also be identified as in the definition 2.

Definition 2: Linear envelope of a signal is represented as the linear function, which is the result of the linear regression with the signal as the input.

The definition 2 provides two crucial aspects for the envelope of a signal. Firstly, the envelope depends on the signal, so it depends on many signal properties, including the signal length and its values. This means that the linear envelope of the whole signal can be different from the first and second halves of the signal. Secondly, because the signal is the time series data and mainly stationary data, the linear envelope is only meaningful when the slope of the envelope line is big enough. With a signal $f(t)$ in the time domain, these two aspects of an informative linear envelope can be formed as a constraint in the equation 5.

\[
y = at + b = \text{LinearRegression}(f(t))a \geq \theta, \quad \theta > 0 \quad (5)
\]
Equation 5 can be rewritten as follows:

\[ LE(f(t)) = \text{LinearRegression}(f(t)) = at + b \]  

with:

\[ \text{slope}(LE(f(t))) = a \]  

and:

\[ \text{bias}(LE(f(t))) = b \]  

With definition 2, the identification mark for the linear envelope is the whole signal or a segment of signal tends to increase or decrease during its whole length with a stable slope. A representative for the envelope is illustrated in figure 4.

Figure 4 shows a speech signal in a real audio file. The duration of the audio is over 1.6s. The subfigure 4a presents the waveform for the whole signal, while the subfigure 4b shows only the first 0.5s. Although the whole signal is mainly distributed as a horizontal trend in these figures, some segments tend to be slope distribution, which is highlighted with green and red colors. The enlarged illustration clearly shows the sloping trend and the linear envelope of the segment. This is solid evidence that the role of the linear envelope
for speech signals is significant and exists in standard datasets.

**D. FREQUENCY OF NON-STATIONARY SIGNAL AND THE DEPENDENCY OF THE LINEAR ENVELOPE TO THE SIGNAL LENGTH**

Generally, the linear envelope hardly exists in a long signal, but it is usually distributed in a short frame. Let us define a new concept to describe the internal stability of a speech signal as follows:

**Definition 3:** Linear Envelope Rate (LER) of a signal with a threshold $\theta$ is the ratio between the number of frames with a slope higher than $\theta$ and the total frames in the signal:

$$\text{LER}(\text{signal}, \theta) = \frac{\#\text{frame, slope}(\text{LE(frame)}) > \theta}{\#\text{frame}}$$  \hspace{1cm} (9)

LER with a threshold $\theta$ presents the ratio of the frames with the high slope and the total frames during the whole signal. Of course, these values can vary for both the different speech data and the value for $\theta$.

Figure 5 presents the distribution of LER in dataset TIMIT. From this figure, there are three meaningful aspects:

• The small threshold $\theta$, the big LER
• The short frame, the big LER
• The common and meaningful ranges, around threshold $\theta = 0.01$ and frame_length $= 0.02$, are valuable (because LERs are not small numbers).

**E. NON-STATIONARY SPEECH SIGNAL IN SOME STANDARD DATASETS**

Speech is usually considered a periodic signal because of its origin. Remarkably, human speech is generated by the
vibration of the vocal tract. Since vibration is a stationary process, speech naturally is a periodic waveform. Nevertheless, this statement is not always the truth. In figure 4, some segments in the speech signal are not stationary and are mainly distributed as a linear function. When the airflow is released out of the human mouth, it is affected by organs such as the uvular, velum, palate, alveolar ridge, tooth, lips, and tongue. In friction with the articulating organs, the properties of airflow change; hence, the periodical of the vibration cannot be held ultimately. This is a universal phenomenon, so it can occur in many kinds of data.

Because the root cause of the linear envelope in the speech signal is the effects from the organs when they work non-stably, these effects can exist in many datasets depending on the content, emotion, language, age, gender, etc. of the speakers.

Table 1 shows the popularity of linear envelopes for two common speech datasets. The values in the table are computed as follows:

$$Popularity(\theta) = \frac{\#sample, LER(sample, \theta) > 0}{\#sample}$$  \hspace{1cm} (10)

All these values in table 1 are non-zero means. The linear trend in the speech signal is not a rare phenomenon but also widely exists in many datasets, with many applications for many languages. If some unique properties of this envelope are exploited, the solution’s performance can be improved overall.

F. SIGNAL DISTORTION OF HUMAN SPEECH

LE can occur in many datasets and languages, but if the impact on the signal is not too significant, it cannot significantly affect the whole system. This section provides a simple tool to measure the interference of the linear elements to the signal.

The total distortion caused by the linear elements is defined in the equation as follows:

$$SDR = \frac{P_{signal}}{P_{distortion}}$$  \hspace{1cm} (11)

The corresponding equation with decibel scale is:

$$SDR_{dB} = 10 \log_{10} \left( \frac{P_{signal}}{P_{distortion}} \right)$$  \hspace{1cm} (12)

In the equation above, the term $P_{signal}$ is computed by:

$$P_{signal} = \int_{t_0}^{t_1} |f(t)|^2 dt = \sum_{t_0}^{t_1} |f(t)|^2$$  \hspace{1cm} (13)

And the $P_{distortion}$, which reflects the power of the linear elements in the whole signal, is described as follows:

$$P_{distortion} = \int_{t_0}^{t_1} |a \cdot t + b|^2 dt = \sum_{t_0}^{t_1} |a \cdot t + b|^2$$  \hspace{1cm} (14)

With SDR equations, it is easy to compute the total change caused by the linear envelope of the linear elements to the speech data. If this value is too much, the signal can be changed a lot, and then the quality of speech feature or speech spectral cannot be guaranteed.

G. DISTORTION IN SPECTRAL REPRESENTATION CAUSED BY LINEAR ENVELOPE

Let $ax + b$ and $\sin(nx)$ denote the linear envelope and a harmonic function. When we compute the integral transform with signal, the dot product for these two functions is identified by:

$$\langle ax + v, \sin(nx) \rangle = \int_{0}^{2\pi} (ax + b)\sin(nx)dx$$  \hspace{1cm} (15)

$$\langle ax + v, \sin(nx) \rangle = \frac{-2\pi a}{n}$$  \hspace{1cm} (16)

The above result shows the impact of the linear envelope on the coefficients in the frequency domain divers due to the distribution of function 16. Because $\frac{-2\pi a}{n}$ is an inverse function with variable $n$, the primary impact range of the linear envelope concentrate on the low frequency. With the high frequency, the increase of $n$ leads to the value of $\langle ax + v, \sin(nx) \rangle$ reduced significantly. In [33], Do et al. showed that with the threshold for computing error under $10^{-8}$, the impacted range of spectrum ends at 6300 Hz.

H. NEGATIVE IMPACT OF THE LINEAR ENVELOPE ON THE WHOLE SYSTEM

Although existing everywhere and affecting many approaches, the problem of LE is much more severe with the deep learning approach. Many deep models mainly depend on RNN to take advantage of its ability to connect the results at many previous steps to process at the current step. If we use RNN or its variants to recognize the speech in figure 3, at the
highlight segment, the feature vector with the noises caused by the linear function cannot be recognized well, and then all the segments after cannot be recognized too. This means that before extracting a feature or transforming a signal, it is necessary to normalize the input signal by removing the linear element.

IV. PERIODICALIZING SPEECH SIGNAL WITH ENVELOPE SUBTRACTION

A. NON-ORTHOGONAL TRANSFORM AND ITS APPLICATION TO AVOID THE LOSS OF INFORMATION IN THE FREQUENCY DOMAIN

1) THE IDEA

One of the first approaches to this problem is using another basis vector set for the frequency space. Although this method can present enough vital information for the recognition model, it still causes the problem that many coefficients are noise. This problem is similar to the example in figure 1. Although these elements have no relevance, the inner product between a basis vector and the speech signal is non-zero. The non-zero is caused by the negative effect from the linear envelope of a speech signal.

On the other hand, the quality of feature vector is not satisfied for the learning model. Let us consider the signal in the equation 1. Instead of using \([\sin(x), \sin(2x), \sin(3x), \sin(4x), \sin(5x)]\) as the basis vectors set, we use a set with six basis functions \([\sin(x), \sin(2x), \sin(3x), \sin(4x), \sin(5x), x]\). All five first values are equal with the result in equation 3. The sixth value, which corresponds with basis vector \(x\) is:

\[
c_6 = \frac{1}{2\pi} \int_0^{2\pi} f(x)dx = \frac{4\pi^2 - 1}{3}
\]

(17)

The result of the integral is non-zero means the value of the signal in the sixth dimension is non-zero. This fact implies that the linear element in the speech signal is one of the main properties of this signal. So, with this basis set, critical information is not lost. The disadvantage of this approach is that the basis set is not orthogonal because the inner product of the sixth vector with the others is non-zero:

\[
\int_0^{2\pi} x \sin(x)dx = -2\pi
\]

(18)

or:

\[
\int_0^{2\pi} x \sin(nx)dx = \frac{-2\pi}{n}, n > 0
\]

(19)

Because the basis vectors are not orthogonal, the information provided by the dimensions of the space is overlapped. Many pieces of research show that the optimal features must present unique and independent information in each element in the vector. Although the information is not optimal in organizing, the meaning provided by the LE is significant and can help the model a lot in terms of improvement.

2) THE ALGORITHM

In this section, we use Fourier transform to illustrate the proposed method. The kernel for the Fourier transform is the complex exponential function as follows:

\[
K(\omega, t) = e^{-j\omega t}
\]

(20)

So the traditional form for the Fourier transform is defined by:

\[
F(\omega_k) = \int_{-\infty}^{\infty} f(t)K(\omega_k, t)dt = \int_{-\infty}^{\infty} f(t)e^{-j\omega_k t}dt, \quad k \in N^+
\]

(21)

In the proposed algorithm 1, we append two dimensions to the frequency domain to describe the linear envelope for the signal. They include one function for the slope and one function for the bias of the LE line. Two corresponding kernels are:

\[
K_{\text{slope}}(t) = t
\]

(22)

\[
K_{\text{bias}}(t) = 1
\]

(23)

The coordinates in these two dimensions are computed by:

\[
dim_{\text{slope}} = \int_{-\infty}^{\infty} f(t)K_{\text{slope}}(t)dt = \int_{-\infty}^{\infty} f(t)dt
\]

(24)

\[
dim_{\text{bias}}(\omega_k) = \int_{-\infty}^{\infty} f(t)K_{\text{bias}}(t)dt = \int_{-\infty}^{\infty} f(t)dt
\]

(25)

After all, the final feature vector includes three elements: two scalars for dimensions corresponding with the slope and bias of the linear envelope, and the last element is the coordinates vector of the traditional transform. The detailed computing process is presented in algorithm 1.

Algorithm 1 Emphasizing Speech Feature by Combining With Signal Linear Envelope

0: Input: audio signal \(f(t)\), feature extractor Extractor (default: Fourier Transform)
0: Output: Feature map Feature
0:                
0: frames ← \(f(t)\)
0: \(n \leftarrow \text{length(frames)}\)
0: Feature ← \text{array[]}
0: \textbf{for} \(i \in [1 : n]\) \textbf{do}
0: \(f_i(t) \leftarrow \text{frames}[i] \ast \text{Windowfunction}\)
0: \(a_i \leftarrow \int_{-\infty}^{\infty} f_i(t)dt\)
0: \(b_i \leftarrow \int_{-\infty}^{\infty} (t \ast f_i(t))dt\)
0: \(F_i \leftarrow \text{Extractor}(f_i(t))\)
0: \(\text{Feature}[i] \leftarrow (a_i, b_i, F_i)\)
0: \textbf{end for}
0: \textbf{return} Feature
0: = 0

The algorithm we have presented is an extension of the Fourier transform. With the other integral transform, such as Wavelet, Laplace, or Melin, we can extend the traditional algorithms by adding two coordinates corresponding with the
slope and bias of the linear envelope of the signal while the remaining of main algorithms are not changed. The feature vector is longer than the original vector with these two values for the LE.

3) ALGORITHM COMPLEXITY

The computational process in Algorithm 1 has a low cost.

In comparison with the traditional Fourier Transform, Algorithm 1 uses two more steps for computing LE coefficients $a_i, b_i$. Because $a_i \leftarrow \int_{-\infty}^{\infty} f_i(t) dt$ and $b_i \leftarrow \int_{-\infty}^{\infty} (t \ast f_i(t)) dt$, they can approximate by a summary of a sequence using the discrete form of integral. In this case, because $f_i(t)$ is only a signal frame, the cost of computing in a frame is too low. For instance, with a 25ms frame of 16.000Hz signal, $f_i(t)$ is presented by 400 data points. The program only uses a loop processor to compute these 400 points to return $a_i$ or $b_i$.

B. IDENTIFYING THE LINEAR ENVELOPE OF THE SPEECH SIGNAL VIA LINEAR REGRESSION

1) THE IDEA

To identify the linear envelope for $f(x)$, we need to compute $a, b$ such that $f(x)$ is as close to $ax + b$ as possible. Let $g(x) = ax + b$, it means that:

$$a, b = \arg \min_{a,b} \int (f(x) - g(x))^2 dx$$

(26)

In the integral $I = \int (f(x) - g(x))^2 dx$, element $f(x)$ is the available value, $I$ can be seen as a linear function with parameters $a$ and $b$. To compute the optimal value for $a$ and $b$, we need to solve the equation $I' = 0$.

With discrete data such as digital signals, linear regression is a much easier approach. Assuming that in each signal segment, there are $n$ points $f(x_0), f(x_1), \ldots, f(x_{n-1})$. Solving the linear regression problem means identifying $a, b$ such that:

$$a, b = \arg \min_{a,b} \sum_{i=0}^{n-1} (f(x_i) - (ax_i + b))^2$$

(27)

From this equation, $a, b$ is computed by:

$$a = \frac{\sum_{i=0}^{n-1} x_i f(x_i) - n \bar{x} \bar{f}}{\sum_{i=0}^{n-1} x_i^2 - n \bar{x}^2}$$

$$b = \bar{f} - a \bar{x}$$

(28)

with:

$$\bar{f} = \frac{\sum_{i=0}^{n-1} f(x_i)}{n}$$

$$\bar{x} = \frac{\sum_{i=0}^{n-1} x_i}{n}$$

(29)

After computing $a, b$ with the equation 28, the function $g(x) = ax + b$ can be considered as the linear envelope for the signal segment. There are two important things about the linear envelope:

- Just some, not all, signal segments have the linear envelope. The others have the horizontal envelope or $a = 0$.
- Linear envelope only exists in the short segment. With the long segments or the whole signal, because of the limit of signal amplitude, $a$ tends to 0.

2) ALGORITHM FOR SPEECH FEATURE EXTRACTION

Traditionally, a speech signal is transformed to show its unique features. This process usually uses Fourier transform and then applies more advanced techniques. In this work, with the motivation from the linear envelope presented in section IV-B, this process is changed, and its pseudo-code is presented in algorithm 2.

Algorithm 2 Emphasizing Speech Feature With Signal LinearEnvelope Subtraction

0: Input: audio signal $f(t)$, feature extractor $Extractor$ (default: MFCC)
0: Output: Feature map $Feature$
0: ————
0: $frames \leftarrow f(t)$
0: $n \leftarrow length(frames[n]$
0: $Feature \leftarrow array[n]$
0: for $i \in [1 : n]$
0: $f_i(t) \leftarrow frames[i] \ast Windowfunction$
0: $(a_i, b_i) \leftarrow LinearRegression(f_i(t))$
0: $f_i(t) \leftarrow f_i(t) - (a_i t + b_i)$
0: $F_i \leftarrow Extractor(f_i(t))$
0: $Feature[i] \leftarrow (a_i, b_i, F_i)$
0: end for
0: return $Feature = 0$

In this pseudo-code, $a, b$ are the slope and bias values for the linear envelope. With step 6, the process of subtracting the signal with its linear envelope can be interpreted as a horizontalizing step. This interference removes the negative impacts of the linear envelope on the frequency coefficients. The noises can be reduced, but the main frequency elements or speech formants are also enhanced.

3) ALGORITHM COMPLEXITY

Similar to Algorithm 1, the computational cost for Algorithm 2 is insignificant.

In Algorithm 2, the additional cost comes from the step $(a_i, b_i) \leftarrow LinearRegression(f_i(t))$. In this step, linear regression is computed via equation 28. Both $a_i$ and $b_i$ are computed by $O(n)$ algorithms. In equation 28 and 29, the main computation costs come from computing $\bar{f}$ and $\bar{x}$. If the frame length is 25ms, corresponding with 400 points or $n = 400$, the cost for computing $(a_i, b_i)$ is too low.

V. EXPERIMENTS

A. AIMS, SCOPE, AND MODEL IN THE EXPERIMENTS

This experiment aims to show the difference in recognition results between our proposed algorithms and the conventional approaches. Particularly, we compared the accuracy of
some recognition tasks with many features using the same recognition method. Generally, we considered three groups of features:

- **Baseline group**: spectrogram, cepstrogram, and MFCC.
- **Group 1**: Algorithm 1 and its variations. This group includes features enhanced with linear envelope coefficients $a$, $b$.
- **Group 2**: Algorithm 2 and its variations. The features include linear envelope coefficients $a$, $b$ and the features extracted from the signal after subtracting the envelope.

To show the improvement of the proposed methods, we conducted some comparisons in speech recognition and speaker demographic recognition. Speech recognition is one of the most complicated tasks, requiring a lot of data, while demographic recognition, including gender, dialect, etc. recognition is more straightforward. These tests also need a diversity of audio lengths for recognition. As presented in figure 5, LE has a close relationship with the audio length; these experiments are used to verify whether LE is a good augmentation technique in general or not.

We used the DeepSpeech2 [36] framework with a deep neural network for all extraction methods to process the feature vectors. DeepSpeech2 in this research is a simplified form of the original network, including three 1D-convolutional layers, six LSTM layers, and one FC layer. We used RELU as the activation function for all layers. For the speech recognition task, the objective function was CTC loss. In the other tasks, MSE loss was used as the function for minimization.

In the beginning, the whole network was initialized using the Xavier method. Then the network was trained with gradient descent using batch normalization for 20 epochs for each task. Finally, we used the model to test with 10% remaining data (we use 80% data for training and 10% data for validation) to get the predicted results.

### B. DATASETS

We used TIMIT [34] dataset for English and VIVOS [35] dataset for Vietnamese. TIMIT includes 6300 audio files with 5.4 hours in total duration, while VIVOS length is 15.7 hours with 12420 audio files. They are standard datasets for speech recognition and are widely used in academics. Each dataset archives a wide range of phonemes in phonetics and semantics meanings. This is why they are usually used in academia to verify new ideas.

In the speech recognition task, we used the original data form and ran the recognition model. On the other hand, in speaker demographic recognition, we made some modifications to gain more data. First, we separated each sample (audio file) into chunks of 1 second. Then we tagged these chunks with the same label as their source. This means that if we had one 4-second audio with the label “female”, we would gain four 1-second audio chunks with the same label. Speech recognition needs the whole audio signal to recognize the completed content, but with gender or dialect recognition, a chunk of 1 second is enough to identify who is speaking.

### C. EXPERIMENTAL RESULTS

#### 1) SPEECH RECOGNITION TASK

Generally, our proposed Algorithms 2 and its variations help the recognition model perform better than the group of Algorithm 1 and the baseline methods. The lowest Word Error Rate (WER) is gained by ES MFCC, or Algorithm 2 with extractor MFCC. This method yields 13.9% and 16.5% in English and Vietnamese, respectively.

On the other hand, Algorithm 1 is slightly better than the baseline group. This is caused by the fact that Algorithm 1 is the combination of spectrum and signal envelope coefficients. All information achieved in the spectrum is kept in the algorithm, and the improvement comes from the additional linear envelope parameters.

Each feature extractor should be analyzed to consider the improvements of Algorithm 1 and Algorithm 2 to the baseline methods. First, because the cepstrum is directly computed with the inverse Fourier transform of the logarithm spectrum, the spectrogram and the cepstrogram are similar and grouped. In this group, the impact of Algorithm 1 is not much. It can be observed in table 2 that the difference among methods in this group is not significant. On the other side, Algorithm 2 performs much better. The improvement of Algorithm 2 reaches around 4% for English and 10% for Vietnamese. These results are significantly better than around 1% by Algorithm 1. Finally, with MFCC, the predicted outputs form the same distribution. Algorithm 2 is the best, while the baseline gets the highest error rate. The only difference, in this case, is that the improvement of Algorithm 1 is more apparent. With MFCC, Algorithm 1 reduces 1.5% for English and 7.1% for Vietnamese, while Algorithm 2 gains better results with 2.1% for English and 10% for Vietnamese in terms of error rate reduction.

#### 2) SPEAKER DEMOGRAPHIC RECOGNITION

This experiment includes two sub-experiments with speaker gender recognition and speaker dialect recognition.

**Speaker gender recognition** is formulated as follows:

- **Input**: 1-second audio chunk
- **Output**: Male or Female

| Group       | Feature       | Size | English WER (%) | Vietnamese WER (%) |
|-------------|---------------|------|-----------------|-------------------|
| Baseline    | Spectrogram   | 128  | 22.1            | 29.0              |
|             | Cepstrogram   | 52   | 24.6            | 30.2              |
|             | MFCC          | 39   | 16.0            | 26.5              |
| Algo 1      | LE Spectrogram| 130  | 23.1            | 28.7              |
|             | LE Cepstrogram| 54   | 21.0            | 27.8              |
|             | LE MFCC       | 41   | 14.5            | 22.9              |
| Algo 2      | ES Spectrogram| 130  | 19.1            | 19.7              |
|             | ES Cepstrogram| 54   | 18.5            | 18.2              |
|             | ES MFCC       | 41   | 13.9            | 16.5              |
H. D. Do: Exploiting Signal Linear Trend in the Time Domain to Enhance Speech Feature

| Table 3. Speaker gender recognition (GR) and dialect recognition (DR) with different features. |
|---------------------------------------------|
| Group       | Feature     | Size | GR Error rate (%) | DR Error rate (%) |
|-------------|-------------|------|-------------------|-------------------|
| Baseline    | Spectrogram | 128  | 4.1               | 4.5               |
|             | Cepstragram | 52   | 3.4               | 4.2               |
|             | MFCC        | 39   | 3.2               | 2.2               |
| Algo 1      | LE Spectrogram | 130  | 3.1               | 4.5               |
|             | LE Cepstragram | 54   | 2.9               | 4.4               |
|             | LE MFCC     | 41   | 2.7               | 2.1               |
| Algo 2      | ES Spectrogram | 130  | 2.9               | 3.9               |
|             | ES Cepstragram | 54   | 2.9               | 3.9               |
|             | ES MFCC     | 41   | 2.6               | 1.8               |

This recognition task was then converted to a binary classification with two labels, male and female. We only used the TIMIT dataset for training and testing in this experiment. To create data for use, after separating all audio files, we randomly chose 10,000 1-second audio chunk with the label male and 10,000 audio chunk with the label female. The ratio for training:validating:testing was 80 : 10 : 10.

Similar to speaker gender recognition, the problem of speaker dialect recognition was converted into a classification task:

- Input: 1-second audio chunk
- Output: dialect label among New England, Northern, North Midland, South Midland, Southern, New York City, Western, Army Brat

This 8-class classification was processed in an unbalanced dataset. The chunks in each dialect were not too many and so diverse. We reduced the number of samples for each class/dialect to 2,000 samples. With the dialects containing insufficient samples, we duplicated their samples randomly to gain 2,000 samples. For training:validating:testing, the ratio was also 80 : 10 : 10, similar to gender recognition. The experimental results for these two tasks are in Table 3.

The result in table 3 reflects clearly the positive impact of our proposed method in terms of error rate. Both Algorithm 1 and Algorithm 2 receive better results than the baseline features. In gender recognition and dialect recognition tasks, the highest error rates come from the original spectrogram with 4.1% and 4.5%, while ES MFCC yields the lowest error rates with 2.6% and 1.8%, respectively.

In the gender recognition task, the best performances belong to emphasized MFCC. Particularly, ES MFCC in Algorithm 2 gains the lowest error rate with 2.6%, and LE MFCC in Algorithm 1 gains 2.7% in error rate. These two error rates are significantly improved compared to 3.2% of the original MFCC and 4.1% of the spectrogram.

The result for the dialect recognition task is similar to the gender recognition task. The two best results come from ES MFCC with 1.8% and LE MFCC with 2.1% in terms of error rate. Compared with the gender recognition task, the difference in this experiment is in the performance of Algorithm 1. It is not better than the baseline features. In the case of the cepstrum, the LE cepstrum shows a higher error rate than the original cepstrum. This means that the coefficients from LE are not always helping the model gain a better prediction.

VI. DISCUSSION

A. WHICH SPEECH APPLICATIONS COULD BE MATCHED WITH PROPOSED FEATURES?

In the section experiments, speech processing tasks were presented and showed the results. In speech recognition, the performance is generally too low, while the results are much better in the speaker gender and dialect recognition. In this section, we will analyze some aspects of these applications and why the proposed methods are good or not.

In speech recognition, much training data is required. This is because both input, a speech signal, and output, a sentence, of the task, are too complicated and diverse. The processing model is a function to map the input to the output, so the complexity of the model is too large. This implies that we need a large dataset to train this model. Due to some limits, this research only runs the experiments with a small dataset to verify the idea. In such a small dataset as TIMIT and VIVOS, the proposed methods, including Algorithm 1 and Algorithm 2, perform so well and gain a much better error rate than conventional methods. Although we do not run our methods in a large dataset such as CommonVoice, they have demonstrated their potential and could be applied to a larger dataset.

With speaker demographic recognition, the improvement of proposed methods is not as straightforward as in speech recognition, but it also shows that Algorithm 1 and Algorithm 2 are applicable for speaker gender and dialect recognition. These recognition tasks are much simpler than speech recognition, so we can conclude which proposed methods can be used for speaker demographic recognition in real applications.

Comparing Algorithm 1 and Algorithm 2, obviously, Algorithm 2 performs better than Algorithm 2. In comparison with the baseline in all experiments, Algorithm 2 significantly improves while Algorithm 1 only reduces the error rate by around 1%. Table 4 briefly summarises the proposed algorithms in the relationship with three speech processing tasks.

B. ANALYSIS IN THE PROCESS OF PROPOSED METHODS

In the two proposed algorithms, two separate manipulations contribute the most to the feature quality. The first is the linear envelope coefficients, and the second is the normalizing step by subtracting the signal with its linear envelope. The second one is also the difference between Algorithm 1 and Algorithm 2.
In this section, we did a small experiment to exploit the contribution of each step to the quality of the feature vector. As in section experiments, MFCC was the best feature for recognition tasks, so we also applied this method as the used extractor in the experiment and ignored the other extractors. We applied the original MFCC and its emphasizing forms for the speech recognition task. In the feature extraction step, we respectively used the methods listed in table 5.

| Feature          | Group      | Size | +LE | -LE | Using a normal extractor |
|------------------|------------|------|-----|-----|--------------------------|
| MFCC             | Baseline   | 39   | No  | No  | Yes                      |
| LE MFCC          | Algorithm 1| 41   | Yes | No  | Yes                      |
| ES MFCC          | Algorithm 2| 41   | Yes | Yes | Yes                      |
| Var-ES MFCC      | Algorithm 2| 39   | No  | Yes | Yes                      |

Var-ES MFCC is a modification of Algorithm 2. From the original form of Algorithm 2, we replaced the step \( Feature[i] \leftarrow (a_i, b_i, F_i) \) with \( Feature[i] \leftarrow F_i \). This means that Var-ES MFCC is the ES MFCC after reducing \((a_i, b_i)\) in the feature descriptor. In other words, Var-ES MFCC is an emphasizing feature with only subtraction only, not including the coefficients from the LE function.

To compare these feature extraction methods in speech recognition, We applied the state-of-the-art models for speech recognition tasks, such as the combination of Transformer with Conformer (T+C) [37], [38] and Transducer [39]. These architectures perform so well in recognition problems, so the difference in results mainly comes from the quality of the feature representation. The model was trained and verified via the VIVOS dataset, with the ratio for training/validating/testing being 80 : 10 : 10. The detailed result is presented in table 6.

Table 6 shows the results in four cases corresponding with four feature extraction algorithms. The result is similar to the first experiment, especially in the four aspects. Firstly, ES MFCC, belonging to Algorithm 2, continues performing well. It achieves much better than the others in both T+C and Transducer, with 23.7% and 28.0% in WER. Secondly, the difference between ES MFCC and Var-ES MFCC is so tiny. This observation concludes that the contribution of the coefficients for linear envelopes in the speech signal is not trivial. On the other hand, ES MFCC and Var-ES MFCC are significantly better than conventional MFCC. This group includes one more phase in the running process, and that phase is the key to pushing the extracted feature’s quality. The phase here is the envelope subtraction step. This step helps to normalize the signal and remove the negative impact of the linear trend on the stationary property of the signal. We can conclude that the LE subtracting process contributes most to improving extracted feature quality. Finally, LE MFCC only improves around 2% compared to the MFCC. It means that the contribution of the LE coefficients to the feature quality is insignificant. This observation is similar to the case of ES MFCC and Var-ES MFCC above.

In a similar experiment with English speech recognition via TIMIT, the result is not different from the result above. We can conclude that there is no appreciable difference between the four features in the two languages (English and Vietnamese). This implies that the two proposed algorithms are language-independent.

**VII. SUMMARY**

This research has shown the existence of linear envelope (LE) and its negative impact on speech feature representation. We have proposed two new algorithms to exploit the LE in feature design to enhance the feature quality. The first algorithm is the extension of the traditional spectrum with LE coefficients. We apply Fourier transform with an extended basis vector set to capture the distribution of the linear trend. The feature is the combination of a conventional feature and LE coefficients. On the other hand, the second algorithm is the combination of exploiting the signal LE and the subtraction step. The extracting process begins by identifying the LE function via linear regression. Then the signal is normalized with LE subtraction to remove its unexpected impacts on the other coefficients. In the end, a standard algorithm for feature extraction is used to extract the stationary properties in the speech. The feature descriptor includes LE coefficients and a feature vector of a normalized signal via the subtraction step.

Experimental results show that ES MFCC, which inherits the strength of MFCC, combines with envelope parameters and the normalization method, performs so well with the best results. The recognizing outcomes are stable in different tasks and different languages. The results are also stable in both Vietnamese and English. This fact also means that the proposed method can be applied to many language speech-processing tasks.

Inside the proposed methods, there are two main steps: identifying LE parameters and removing their impact via a subtraction operator. Among these two phases, envelope coefficients contribute trivially to the final quality of the feature vector. The main improvement in feature quality comes from the envelope subtraction step. Via many experiments, We can conclude that removing the linear trend in the speech signal is the most significant contribution to the correctness of vector representation.

**REFERENCES**

[1] M. Afify and O. Siohan, “Comments on vocal tract length normalization equals linear transformation in cepstral space,” IEEE Trans. Audio, Speech, Language Process., vol. 15, no. 5, pp. 1731–1732, Jul. 2007, doi: 10.1109/TASL.2007.896653.
J. Laver, “The Phonetic Description of Voice Quality,” *IEEE Signal Process. Lett.*, vol. 17, no. 11, pp. 957–960, Nov. 2010, doi: 10.1109/LSP.2010.2079930.

Z. Huang, J. Epps, D. Joachim, and V. Sethu, “Natural language processing methods for acoustic and landmark event-based features in speech-based depression detection,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 435–448, Feb. 2020, doi: 10.1109/JSTSP.2019.2949419.

M. Bhattacharjee, S. R. M. Prasanna, and P. Guha, “Speech/music classification using features from spectral peaks,” *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 1549–1559, 2020, doi: 10.1109/TASLP.2020.9332026.

D. L. Moore, Z. Hamilton, and R. J. Haddad, “Pico-grid smart home energy management system using mel frequency cepstral coefficients,” in *Proceedings*, SoutheastCom, Huntsville, AL, USA, 2019, pp. 1–6, doi: 10.1109/SouthEastCon43211.2019.9020412.

D. S. Widyowaty and A. Suyanto, “Accent recognition by native language using mel-frequency cepstral coefficient and K-Nearest neighbor,” in *Proc. 3rd Int. Conf. Inf. Commun. Technol. (ICOACT)*, Nov. 2020, pp. 314–318, doi: 10.1109/ICOACT50329.2020.9332026.

M. Sahidullah and G. Saha, “A novel windowing technique for efficient computation of MFCC for speaker recognition,” *IEEE Signal Process. Lett.*, vol. 20, no. 2, pp. 149–152, Feb. 2013, doi: 10.1109/LSP.2012.2235067.

T. Kinnunen, “Low-variance multitaper MFCC features: A case study in robust speaker verification,” *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 7, pp. 1990–2001, May 2012, doi: 10.1109/TASL.2012.2190196.

R. W. Frick, “Communicating emotion: The role of prosodic features,” *Psychol. Bull.*, vol. 97, no. 3, p. 412, 1985.

R. Cowie and E. Douglas-Cowie, “Automatic statistical analysis of the signal and prosodic signs of emotion in speech,” in *Proc. 4th Int. Conf. Spoken Lang. Process. (ICSLP)*, vol. 3, 1996, pp. 1989–1992, [Online]. Available: https://ieeexplore.ieee.org/document/608027, doi: 10.1109/ICSLP.1996.608027.

I. Luengo, E. Navas, I. Hernáez, and J. Sánchez, “Automatic emotion recognition using prosodic parameters,” in *Proc. Interspeech*, 2005, pp. 493–496, [Online]. Available: https://aholah.edu.eus/papers/Interspeech2005.pdf.

C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, “Analysis of emotion recognition using facial expressions, speech and multimodal information,” in *Proc. 6th Int. Conf. Multimodal Interfaces (ICMI)*, 2004, pp. 205–211.

K. S. Rao, S. G. Koolagudi, and R. R. Vempada, “Emotion recognition from speech using global and local prosodic features,” *Int. J. Speech Technol.*, vol. 16, no. 2, pp. 143–160, Jun. 2013.

S. G. Mallat, “A theory for multiresolution signal decomposition: The wavelet representation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 7, pp. 674–693, Jul. 1989, doi: 10.1109/34.192463.

S. Suganya and E. Y. A. Charles, “Speech emotion recognition using deep learning on audio recordings,” in *Proc. 19th Int. Conf. Adv. ICT Emerg. Regions (ICTER)*, Sep. 2019, pp. 1–6, doi: 10.1109/ICTer48817.2019.9023737.

S. Gupta, K. De, D. A. Dinesh, and V. Thakendiyidjooy, “Emotion recognition from varying length patterns of speech using CNN-based segment-level pyramid match kernel based SVMs,” in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2019, pp. 1–6, doi: 10.1109/NCC.2019.8732191.

M. Kawasaki, Y. Yamada, Y. Ushiku, E. Miyashita, and Y. Yamaguchi, “Inter-brain synchronization during coordination of speech rhythm in human-to-human social interaction,” *Sci. Rep.*, vol. 3, no. 1, Dec. 2013, doi: 10.1038/srep01692.

M. Latinus and P. Belin, “Human voice perception,” *Current Biol.*, vol. 21, no. 4, pp. R143–R145, Feb. 2011, doi: 10.1016/j.cub.2010.12.033.

H. D. Do, T. D. Chau, D. D. Nguyen, and S. T. Tran, “Enhancing speech signal features with linear envelope subtraction,” in *Advances in Computational Collective Intelligence (Communications in Computer and Information Science)*, vol. 1463, K. Wojtkiewicz, J. Treur, E. Pimenidis, and M. Malesza, Eds., Cham, Switzerland: Springer, 2021, doi: 10.1007/978-3-030-88113-9_25.

J. S. Garofolo, “TIMIT acoustic-phonetic continuous speech corpus,” in *Linguistic Data Consortium*, 1993. [Online]. Available: https://catalog.ldc.upenn.edu/LDC93S1, doi: 10.35111/17gk-bn40.

H. T. Luong and H.-Q. Vu, “A non-expert Kaldi recipe for Vietnamese speech recognition system,” in *Proc. WLS/IOAF4HL COLING*, 2016, pp. 51–55.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” 2017, arXiv:1706.03762.

G. Auli, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, “Conformer: Convolution-augmented transformer for speech recognition,” 2020, arXiv:2005.08180.

N. Jaitly, D. Sussillo, Q. V. Le, O. Vinyals, J. Sutskever, and S. Bengio, “A neural transducer,” 2015, arXiv:1511.04868.

HAO DUC DO received the B.Sc. and M.Sc. degrees in computer science from the University of Science—Vietnam National University, Ho Chi Minh City, Vietnam, in 2015 and 2018, respectively, where he is currently pursuing the Ph.D. degree in computer science. He is also a full-time Lecturer at FPT University, Vietnam. His research interests include speech signal processing and deep generative learning.