The Application of Fourier Transform in Speech Recognition Systems

Changyuan Wang
Department of Physics, University of California, Santa Barbara, CA 93106

(Dated: March 15, 2021)

Speech delivers information that can be transformed into signals by using Fourier transform. After evaluating different forms of signals, speech recognition systems can identify a person, an emotion, or a language. This paper will discuss how we use Fourier transform in such recognition systems. In voice recognition, Fourier transform decodes a human voice and distinguishes different persons by vocal characteristics. It can also convert audio into written words. In Speech Emotion Recognition, Fourier parameter features are useful in extracting various emotional states in a speech. In Spoken Language Recognition (LR), Fourier parameter features help recognize numerous languages from speech signals. In general, Fourier transform is an effective tool to decompose signals and classify vocal information. We can develop various applications from this characteristic of Fourier transform.

I. INTRODUCTION

Fourier analysis has been widely applied in signal processing, digital image processing, and many other types of fields such as Quantum mechanics [1]. The idea is simple: we represent signals in waveforms, and then we use the functions that we are familiar with to approximate the waveforms. Unlike the Fourier series, which simulates a periodic function by a summation of sine and cosine terms, the Fourier transform represents a more general, nonperiodic function by the superposition or integral of complex exponentials [2]. A simple example will be transforming a single image into a two-dimensional waveform and then representing it in a complicated function of exponentials. This is a toy model for digital image processing.

Fourier transform has even more advantages. It serves as an extremely powerful mathematical tool to examine non-periodic signals because it is one of the simplest transforms among the other transformation methods used in mathematics [3]. The time consumption is much less due to this method. In the modern world, it provides easier solutions for the problems [1]. For instance, image processing like iris recognition and signal processing like voice recognition use Fourier transform to identify a person's iris or voice features. In particular, the iris technique is fast and accurate, so it is vastly applied to many security systems like the police force and customs.

In this paper, we will discuss how Fourier Transform is used in voice and speech recognition systems, such as identifying a person, an emotion, or a language. These features serve as an important part of identification systems.

II. MATERIALS AND METHODS

This section will be centered on the scientific and mathematical methods which are applied to the speech recognition system. The content will include what the methods are and how they are used in the systems.

Voice Recognition, Speech Emotion Recognition Systems, and Spoken Language Recognition all use Fourier transform to turn the input audio information into data and figures. Researchers will analyze the figures and functions, divide them into categories based on their characteristics, and then obtain the desired output.

In the following subsections, I will first briefly introduce the general idea of Fourier transforms and discuss the three kinds of Speech Recognition systems one by one.

A. Fourier Transforms

As we have mentioned in the introduction, Fourier transform is a tool that breaks a waveform (a function or signal) into an alternate representation, mainly an integral or summation of complex exponentials [2]. It is an significant mathematical breakthrough devised by Joseph Fourier in 1822. Equation 1 is the definition of Fourier transform.

$$f(x) = \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(k) e^{ikx} \, dk \tag{1}$$

$$g(k) = \tfrac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-ikx} dk \tag{2}$$

We often see Fourier transform in sinusoidal forms because complex exponentials can be written into a linear combination of sine and cosines. The relationship between complex exponentials and trigonometric functions is clearly explained by Euler's formula:

$$e^{i\theta} = \cos\theta + i\sin\theta \tag{3}$$

The main idea of Fourier transform is to transform from the time domain and the frequency domain. The reason is that signals are functions of time, but patterns of signals in the frequency domain are easier to be dealt with [2]. Therefore, we transform functions with respect to time into frequency-related forms and analyze them in the frequency domain. After we understand the data, we can change the desired outputs back into the time domain. Figure 1 shows an example of such transformation. We can see that amplitudes become more apparent in the frequency domain, so information can be processed more efficiently.

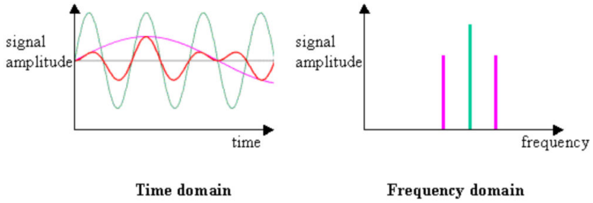


FIG. 1. An example of a typical Fourier transform. In the left subplot, the red curve represents a signal, and it is made up by pink and green sine waves. The red curve is transformed into the curves in right subplot [3].

B. Voice Recognition

In a voice recognition system, Fourier transform is widely applied in processing acoustic information and transforming them into numerous desired outputs [4]. However, we can not use Equation 1 directly, because the data stored in computers are not continuous. We need a transformation that works with discrete data. Therefore, we will use discrete Fourier transform (DFT) instead:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N}kn} \tag{4}$$

$$x_n = \frac{1}{N}\sum_{k=0}^{N-1} X_k \cdot e^{i\frac{2\pi}{N}kn} \tag{5}$$

In this way, Fourier transformation can be applied to a discrete sequence of data. However, DFT is not the optimal approach for analyzing a large database. Typically, we will use fast Fourier transform (FFT), which is a common type of DFT that will accelerate the transformation on computers [4]. FFT has specific algorithms that help to simplify the converting procedure in matrix forms, such as Cooley–Tukey algorithm. By utilizing FFT, acoustic information can be processed efficiently [3].

After a person's voice features have been stored and analyzed, we can vary the outputs based on our needs. For example, we can match up the voice features with stored speech information to convert audio into texts. We can also distinguish a specific speaker from other persons by vocal characteristics. These features are common in electronic devices [4].

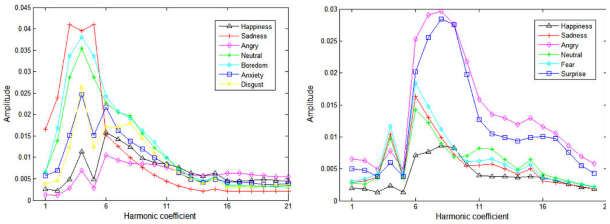
C. Speech Emotion Recognition

Speech emotion recognition is defined as identifying the emotional states of a speaker from his or her speech. Recently, this recognition system catches more and more attention, because people believe it may be helpful in various fields such as criminal investigation and psychological health systems [5].

During the past few decades, studies have shown that harmonic features have the potential of developing speech emotion recognition systems because voice signals are harmonically related [6]. To analyze the harmonics, we first need to transform them into convenient mathematical forms. Here is how Fourier transform comes into play. Since we need frequency, amplitude, and phase to fully specify a harmonic, we can use sines and cosines to estimate it. Also, we know that Fourier transform works well with sines and cosines. Therefore, Fourier analysis becomes the main tool in decomposing the harmonic sequences. For convenience, researchers give the new name ”Fourier parameter (FP)” to a given set of harmonics [5]. As its name suggests, the FP model contains the implication of using Fourier transform to decode harmonic features.

Experimental results show that FP features are effective in classifying various emotional states in speech signals [5]. Take a specific analysis from a research group in IEEE as an example. IEEE researchers aimed to compare and contrast Fourier parameters for different emotions. They proposed an FP model which describes the unique characteristics of a signal by using a set of harmonic coefficients, H_1 to H_{20}. These harmonic coefficients are actually different Fourier parameters that represent features of harmonics. Researchers found that twenty is the optimal number of coefficients to pinpoint an emotion from any voice signal [5].

Figure 2 clearly illustrates what happens. To extract FP features, IEEE researchers selected two large speech emotion databases to build their model: a German emotional corpus (EMODB) and a Chinese emotional database (CASIA). These databases are provided by prestigious institutes in Germany and China, and each database contains thousands of wave files. The data in Figure 2 are the average behaviors of the entire databases [5]. From the figure, we can see that the researchers chose six to seven most common emotions in daily life. The harmonic coefficients, H_1 to H_{20}, have different amplitudes and trends for different emotions. For example, in German, the pick for sadness is H_3 but the pick for angry is H_6. Sadness has the lowest average values from H_9 to H_{20}, but it also has the highest average values from H_1 and H_5. Distinct emotions have different combinations and values of harmonic coefficients, so we can use the twenty coefficients to specify an emotion from any speech signal. We can also notice that the same emotions have different coefficients in different languages, and we will discuss more it in Section III.



(a) German

(b) Chinese

FIG. 2. The amplitudes of the harmonic coefficients for different emotions in German and Chinese. Data points are the mean values of harmonic coefficients, H_1 to H_{20} [5].

To be noted, for this FP model to be widely applied, we need to make sure that the recognition system is speaker-independent. This is one of the latest and most challenging concerns in the field of speech emotion recognition. IEEE researchers specifically addressed this concern. They extracted more than two thousand of features that are related to pitch, spectrum, voice quality, and others. They used massive statistics to conclude a general pattern, and it turned out that harmonic coefficients can successfully remove the dependence on speakers. This is a breakthrough because it has better generalization than the speaker-dependent approach that we used before [5].

D. Spoken Language Recognition

Spoken language identification (LID) or spoken language recognition (LR) is the system of identifying an language from speeches. Experimental results show that FP features are also useful in this recognition system because they can effectively recognize different languages from speech signals [6].

We will use research from the National Institute of Technology Puducherry as an example. To extract FP features, researchers compared two large multilingual databases, the Indian Institute of Technology Kharagpur Multilingual Indian Language Speech Corpus (IITKGP-MLILSC) and Oriental Language Recognition Speech Corpus (AP18-OLR) [6]. The mechanisms of transforming the acoustic information into analytical FP features are similar to the one used on Speech Emotion Recognition.

Figure 3 presents the results of six representative trails, a-f, where data are from the database AP18-OLR. The magnitude of data points is means of harmonic coefficients. These coefficients have the same meanings and functions as we discussed in Section II C. We can find that different languages correspond to different peaks, and 120 peaks are enough for distinguishing one language from the other [6].

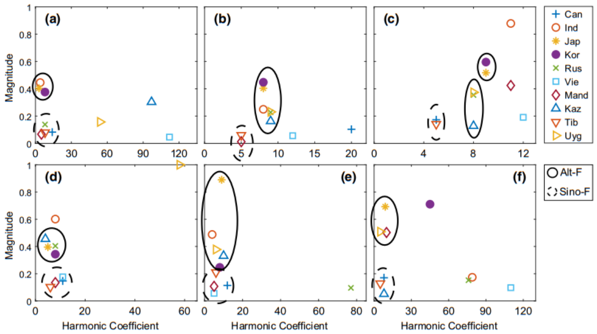


FIG. 3. The amplitudes of the harmonic coefficients in 10 oriental languages. Highest peaks are the mean values of harmonic coefficients, H_1 to H_{120}. This plots shows six random trails a-f, where data are from six sets of random speech signals from AP18-OLR. Note that x-axis are different for each subplot [6].

As we expected, similar languages will have closer peaks. To clarify ”similar,” we will use a linguistic definition, which is classifying languages into families. In Figure 3, different line styles denote different Asian linguistics families. The solid line is for the Altaic family (Alt-F in Fig. 3), while the dashed line is for the Sino-Tibetan family (Sino-F in Fig. 3) [6]. We can see that harmonic amplitudes from the same linguistic family tend to form clusters. In Figure 3a, Altaic languages have mean values around 0.4, while Sino-Tibetan languages have means around 0.1. In figure 3a, d, e, Cantonese (Can.), Mandarin (Mand.), and Tibetan (Tib.) form clusters around amplitudes of 0 to 0.2. The result makes sense because these three languages not only all belong to the Sino-Tibetan family, but also are all languages from China. The fact that they are used in regions that are geographically close to each other makes them even more similar. Moreover, Altaic languages tend to have higher peaks than Sino-Tibetan languages in general. We can see that the Alt-F circle is usually above the Sino-F circle in Figure 3a-f.

Like speech emotion recognition, to make the spoken language recognition system widely applicable, we need to make it speaker-independent. According to the researchers, IITKGP-MLILSC consists of recorded speech files in 27 major Indian languages from 3 major Indian linguistic families. Due to the massive data, the conclusion drawn from the database is successfully speaker-independent and accurate. The recognition system made from this database has a 96% accuracy for identifying languages from 30-45s duration speech utterance, regardless of the ages, genders, and etc. of the speakers. AP18-OLR database, on the other hand, contains speech data in 10 oriental languages from 5 Asian linguistic families. This database is not Indian language-specific, so the conclusions drawn from it are more general. For example, the system may be able to only tell if a language is Indian, but it can not specify which Indian linguistic family the language belongs to [6].

III. RESULTS

We have seen that both FFT and FP come from Fourier transform, and they are important processing tools in Speech Recognition systems. In fact, Fourier transform remains to be the most time-saving method in the field of Speech Recognition systems so far [5].

To be noted, if we want to make the Speech Emotion Recognition system more accurate, we need to take languages into consideration. Figure 4 tells us that the same harmonic coefficient has different means and patterns for different languages.

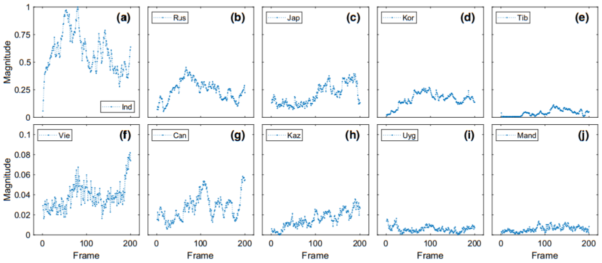
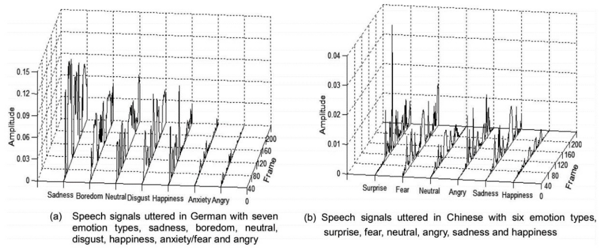


FIG. 4. Means of harmonic coefficients H_3 for 10 languages in AP18-OLR database [6].

Figure 5 explains how languages affect emotions more clearly. For sadness, the harmonic amplitudes of German range from 0 to 0.9, but amplitudes for Chinese are mostly under 0.1. Also, German has higher amplitudes in general. Same emotion results in different patterns of the harmonic coefficient for different languages.



(a) Speech signals uttered in German with seven emotion types, sadness, boredom, neutral, disgust, happiness, anxiety/fear and angry

(b) Speech signals uttered in Chinese with six emotion types, surprise, fear, neutral, angry, sadness and happiness

FIG. 5. Amplitudes of harmonic coefficients H_1 to H_{200} for German and Chinese. Same emotions in German and Chinese are harmonically different [5].

Since figure 4 and 5 reveal obvious differences of amplitudes between languages, we can safely generalize the conclusion that language is a necessary factor in Speech Emotion Recognition. Since languages in the same linguistic family have similar harmonic amplitudes, exploring the emotional patterns of a linguistic family may be an effective approach.

IV. DISCUSSION

In conclusion, a person's unique voice features can be represented by functions, and then we can describe voices mathematically and systematically. In voice recognition, FFT is an efficient approach to convert audio into texts. In Speech Emotion Recognition and Spoken Language Recognition, Fourier parameter features help recognize emotional states or various languages from different shapes or functions of speech signals.

There are also other methods which have been used for this problem. For Spoken Language Recognition and Speech Emotion Recognition, both of the researchers tried MFCC features, which is also a common technique other than the Fourier parameter. MFCC stands for mel-frequency cepstral coefficients, which are coefficients that consist of Mel-frequency cepstrum. Mel-frequency cepstrum also takes the Fourier transform of signals, but the difference is it will map the power spectrum to Mel-scale. Since it takes logs of power during processing, its outputs are closer to humans auditory system than linearly-spaced scale because human's hearing is not linear [6]. Experimental results have shown that when FP and MFCC features and combined, the accuracy is improved [5].

The recognition systems that we mentioned in Section II are useful in an even broader context. For instance, Voice recognition is also commonly used for vocal commands where we do not need to press any buttons [4]. Speech emotion recognition is very helpful for intelligent assistance and man-machine interaction [5]. The spoken language recognition system can help linguistic programs because it can analyze thousands of speech in seconds, and give researchers their desired result without having them waste time listening to the speeches one by one [6]. When the system is mature enough, maybe someday it may even replace the occupations of translators.

ACKNOWLEDGMENTS

I would like to thank our professor, Dr. Jean Carlson, for teaching us and providing the necessary complex analysis knowledge. Her classes serve as a guide for my work and help me understand other online research papers. I also owe gratitude to our teaching assistant Zihang Wang, our learning assistants Davis Thuillier, and Aryan Zaveri. I want to thank them for helping me with the materials of Fourier transforms and enhancing my understanding. I am also grateful for the insightful comments offered by my peer reviewers Zhaoyi He and Haopu yang. They generously pointed out my errors and I improved some of the drawbacks in this paper.

[1] Deepika Babel Nisha Chittora, "A brief study on fourier transform and its applications," IRJET **05**, 1127–1131 (2018), 2395-0056.

[2] The Introduction to the Fourier Transform, www.thefouriertransform.com.

[3] Danielle Collins, "How are fast fourier transforms used in vibration analysis?" (2019), www.motioncontroltips.com/how-are-fast-fourier-transforms-used-in-vibration-analysis/.

[4] Qicong Liao Weiqiang Liu, "Approximate designs for fast fourier transform (fft) with application to speech recognition," IEEE Transactions on Circuits and Systems **66**, 4727–4739 (2019).

[5] Ning An Kunxia Wang, "Speech emotion recognition using fourier parameters," IEEE Transactions on Affective Computing **06**, 69–75 (2015), 2395-0056.

[6] N.Sugan N.S.Sai Srinivas, "Recognition of spoken languages from acoustic speech signals using fourier parameters," Circuits Syst Signal Process **38**, 15018–5067 (2019).