

Introduction

1.1 Motivation

For decades, researchers have been exploring ways in which human speech would be recognized by computers. After numerous advancements, some of which are outlined in this book, robust automatic speech recognition which works in most environments and under a variety of noise conditions is still years away. Today's state-of-the-art speech recognition systems still are to a large extent unable to cope with noisy conditions such as those in parties, subways, and crowded meeting rooms.

Yet, speech recognition is a field where we know that an eventual solution should exist since the human speech recognition can cope with most practical situations (with some obvious exceptions, of course). The main question that researchers in speech recognition have asked in the previous decades, and continue to ask to this day, is what are the processes and mechanisms employed by humans that allow robust speech recognition even in the presence of severe noise.

So, on the one hand, we have the human hearing and speech recognition system which works seamlessly and in most cases adequately. On the other hand we have speech recognition systems which lack the robustness, efficiency, and accuracy that are provided by their biological counterparts. What are the differences, and how can these differences be duplicated in artificial speech recognition systems?

There are aspects to the human speech processing and recognition system that we do not know yet. However, what we do know or can measure experimentally illustrates a clear difference between the artificial and the

biological systems. Humans have two ears, allowing for directional hearing, localization, and tracking. The equivalent of such abilities in artificial systems would be an array of microphones, which also could provide for directional hearing, localization, and tracking.

Second, the human ear is not just the equivalent of a simple microphone. Some of the aspects of the design of the human ear allow for greater capability than the simple microphone model.

Another, and perhaps most important, difference is in the way speech and sound in general is processed by humans. Current artificial speech recognition systems, for the most part, utilize only the amplitude of the incoming sound waves at different frequencies in order to recognize speech. First of all, the use of spectral (or cepstral) transformations for speech processing may not be the ideal representation, although they are almost universally used by all speech recognition systems. Second, the phases of the different frequencies are often tossed out without much attention or processing. An open question in speech processing has been whether phase should play a more dominant role in the speech recognition process.

1.2 The Meaning of Phase

At this point, it is useful for us to define what exactly phase is. As will be illustrated in the next chapter, every recorded sound signal can be expressed as a summation of sinusoids. These sinusoids would be of the form $A \cos(2\pi ft - \phi)$, where A is the magnitude of the sinusoid, f is the frequency in Hz, and finally, ϕ is the phase or the starting point of the sinusoid.

Now, why would the starting point of this sinusoid be important? As it turns out, there are some things about phase that make it very unique for speech processing applications. First of all, just like FM radio receivers are less noisy than traditional AM radio receivers, the estimation of phase can often be done with more reliability than the estimation of amplitude. Another important aspect of phase is the timing information that it provides, which is essential in multi-microphone applications.

In order to better illustrate the importance of phase even with single microphone applications, consider the following example. In this example, we assume that a sound signal consisting of three sinusoids (one with frequency 100 Hz, another with frequency 105 Hz, and a third with frequency 110Hz) is recorded by a microphone for a total duration of 50ms. The

recorded signal $x(t)$ can be expressed as:

$$x(t) = \sin(2\pi 100t) + \sin(2\pi 105t) + \sin(2\pi 110t) \quad (1.1)$$

and is illustrated in Figure 1.1.

Now, let us assume that after this signal is recorded and is transformed into the frequency domain, its phase is tossed away and only its spectral magnitude is kept. As a result of this operation, any of the time domain signals that are shown in Figure 1.2 are possible.

Clearly, these signals look different than the original signal. Yet, since the phase information is tossed away, there is no simple way to distinguish between the different possibilities. This ambiguity, of course, could be a significant problem when it comes time for speech recognition since most recognition systems do not utilize phase directly.

1.3 Dual Microphone Speech Processing, or Why Two Ears Are Better Than One

The title of this section is the title of a talk that one of the authors has often given to introduce audiences to microphone arrays. As mentioned before, the binaural hearing capability of humans allows for directional hearing, localization, and tracking. In this section, a brief overview of microphone arrays will be given.

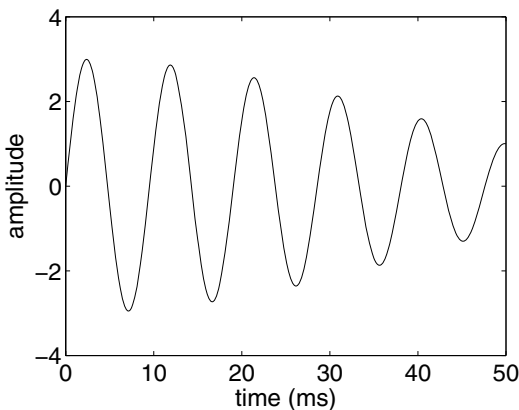


Figure 1.1: A simulated signal consisting of three sinusoids.

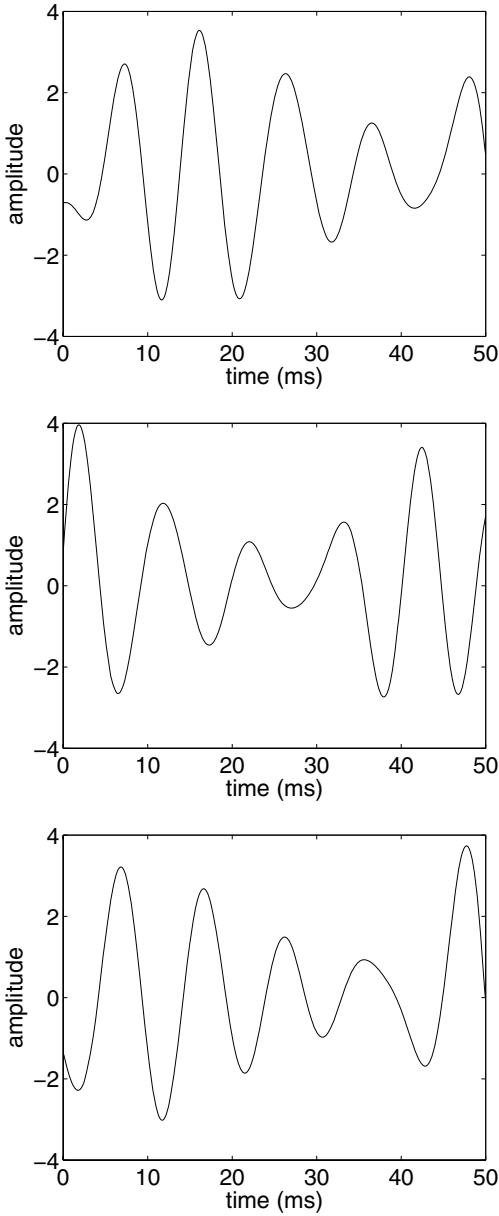


Figure 1.2: A set of signals with the same amplitude spectrum as Figure 1.1 but with random phases.

When a sound is produced by a source, it will arrive at possibly different times and intensities at different microphones. The farther a microphone, the lower the intensity of the arriving sound waves and the greater the time delay of arrival. As a result, when looking at the specific recordings made by, say, two microphones, their intensity will be different and there will be a time shift as long as one of the microphones is closer to the sound source.

Now, in practice we do not know where the source is, or what the sounds from the source look like. All that we have are recordings made from microphones at different spatial locations. From these recordings, we must determine the location or direction of the sound source as well as the specific sounds generated by the source (apart from other sounds or noises at other locations).

Often, the intensity difference across microphones will be so small (especially if the microphones are placed close together) that it will not provide a significant amount of useful information. This is true especially in the presence of noise, when minute differences in intensity cannot be reliably estimated. Instead of intensities, the time difference of arrival at different microphones can be estimated with more reliability. In fact, it is possible to localize and enhance a sound source based on the time difference (or, equivalently, based on frequency dependent phase differences). As a result, for multi-microphone applications, phase plays a very important role.

1.4 The Microphone From the 22nd Century - The Human Ear

For all speech processing applications, we must record the frequency, phase, and amplitude of the sound waves at specific points in space. This measurement will also disturb the sound waves at the measurement points depending on the size and shape of the recording device.

Small microphones (such as the common condenser microphones) would result in the least disturbance of the sound field in an environment. However, do we ideally want passive microphones, or more intelligent microphones whose disturbance of the sound wave is done so in a calculated and purposeful manner.

Looking at the biological equivalent of microphones (i.e. the human ear), we see a striking difference between the human ear and the microphones

of today. The human ear, decomposed into three segments known as the outer, middle, and inner ear, performs numerous tasks before the sound waves entering the outer ear are converted to nerve signals coming out of the inner ear. The outer ear acts as a microphone shell, focusing and coalescing sounds towards the ear canal and onto the middle ear. There, the sounds are converted to bone vibrations via the ear drum. The series of bone interconnections carry (and also slightly modify) the sound waves into the cochlea of the inner ear. In the cochlea, thousands of hair cells connected to nerve endings record the incoming vibrations at different spatial points. These nerve signals are then carried to the brain for further processing, and are in simple terms the output signal of the human ear.

Clearly, the human ear is far more sophisticated than any microphone available today. Aside from all the processing performed by the outer and middle parts of the ear, the spatial disparate recording of the inner ear's hair cells allows for the precise measurement of phase of the incoming sound signal.

1.5 Why Smart Computers Are Hard To Find

The human brain consists of approximately 100 Billion neurons with a vast amount of interconnection. Today's most powerful computer processors, on the other hand, have less than 100 Million transistors. The more than thousand-fold difference in basic computing elements is one reason that today's computers are no match for the human brain.

Yet, at this junction, one would be led to ask whether a transistor is really the same as a neuron. A neuron is a non-linear summation of any number of inputs, working in an inherently analog fashion. A transistor, on the other hand, could also have multiple inputs and could also work in an analog fashion. However, in today's computers most transistors have only a few inputs and are, for the most part, designed to operate in a digital form (i.e. they are either fully turned on or fully turned off). Digital design is beneficial and practical for numerous reasons (the most important of which include stability and ease of design), but it is not as efficient as an analog system. Hence, a transistor is not really the equivalent of a neuron, at least not in the way that transistors are utilized today.

The question of form versus numbers is also important in the speech processing today. Speech, because of its small spectral bandwidth, can

experience significant processing, enhancement, and modification in real-time even with the more modest processors of today. However, all the processing in the world may not be enough if the type of processing is inefficient. Today's methods for speech recognition start from standard spectral transformations that have existed for decades. These transformations do not, in a simple way, take the phase of the sound signals and especially the relative phase of the signals into account. While in certain domains these techniques work well, whether they are really the same algorithms and processes that are performed by the human ear AND brain remains to be seen.

1.6 The Bigger Picture

In the end, the robustness and reliability of human speech recognition may be more than just a simple result of either the larger and more optimized processing of the brain, the spatial sampling and directionality afforded by two ears, or the complex sound wave capturing ability of the ear. The human recognition of speech may in fact rely on all of these abilities, as well as on information provided by other senses such as eyes. For example, in very noisy environments, the motions of the lip and their synchronization with the audio provide important clues that could be combined with the noisy audio for improved speech recognition.

1.7 Book Overview

This book focuses mainly on the importance of phase in speech processing applications. We initially focus on the background signal processing theory in Chapter 2. Next, in Chapter 3, we focus on the current state-of-the-art in speech recognition including both the front end processing as well as the back-end probabilistic models. It is then argued that although specific phase information is not utilized in most speech recognition systems, phase restoration algorithms that can infer phase from the magnitude information do exist. Chapter 4 is composed of two parts. In the first part we describe how humans hear, and in the second part we illustrate the importance of phase in human speech through a detailed overview of the concepts as well as through experimental verification.

Chapter 5 provides a detailed overview of a phase-based speech processing algorithms using multiple microphones, including sound localization and phase-based speech enhancement. This chapter also provides an overview of speech enhancement and sound localization in general.

Chapter 6 concludes the book and provides directions for future work.