Linguistic Portfolios

Volume 12 Article 2

2023

A COMPREHENSIVE REVIEW OF INTONATION: PSYCHOACOUSTICS MODELING OF PROSODIC PROMINENCE

Ettien Koffi
St. Cloud State University

A COMPREHENSIVE REVIEW OF INTONATION: PSYCHOACOUSTICS MODELING OF PROSODIC PROMINENCE

ETTIEN KOFFI

ABSTRACT

Bolinger (1978:475), one of the foremost authorities on prosody of a generation ago, said that "Intonation is a half-tamed savage. To understand the tamed or linguistically harnessed half of him, one has to make friends with the wild half." This review provides a brief explanation for the tamed and untamed halves of intonation. It is argued here that the pitch-centered approach that has been used for several decades is responsible for why one half of intonation remains untamed. To tame intonation completely, a holistic acoustic approach is required that takes intensity and duration as seriously as it does pitch. Speech is a three-dimensional physical entity in which all three correlates work independently and interdependently. Consequently, a methodology that addresses intonation comprehensively is more likely to yield better results. Psychoacoustics seems to be well positioned for this task. Nearly 100 years of experimentations have led to the discoveries of Just Noticeable Difference (JNDs) thresholds that can be summoned to help tame intonation completely. The framework discussed here expands the analytical resources and facilitates an optimal description of intonation. It calculates and ranks the relative functional load (RFL) of pitch, intensity, and duration, and uses the results to compute the melodicity score of utterances. The findings replicate, based on JNDs, how the naked ear perceives intonation on a four-point Likert melodicity scale.

Keywords: Psychoacoustics, Text Normalization, Linguistic Tokenization, Acoustic Tokenization, Relative Functional Load of Pitch, Relative Functional Load of Intensity, Relative Functional Load of Duration, Just Noticeable Difference Thresholds, Melodicity Score, Melodicity Scale

1.0 Introduction

Over the past few years, I have written "Comprehensive Reviews" of F0 (Koffi 2019a), intensity (Koffi 2020), and duration (2021b). The present review is the epitome of the three previous ones because the cumulative insights derived from them are brought to bear on the study of intonation. Comprehensive reviews are an academic genre that aim at summarizing and highlighting salient points of an issue of scholarly interest in order to open new avenues for research. A good example of this is Klatt (1987) in which he summarized the state of the art in text-to-speech (TTS) synthesis in 56 pages.

The launching pad of the current review is Bolinger's (1978:475) summary of the state of intonation studies, which he described as follows, "Intonation is a half-tamed savage. To understand the tamed or linguistically harnessed half of him, one has to make friends with the wild half.*"* The coarse subdivision between the tamed and the untamed halves of intonation serves as the two main foci of this review. In the first, the tamed half, is reviewed succinctly. The second, which is the longest, describes the psychoacoustic methodology that promises to help tame the wild side. New concepts are introduced, discussed, and illustrated with Utterances 1 and 2, taken from Ladefoged (2001:17).

Utterance 1: <*When danger threatens, your children call the police*>
Utterance 2: <*When danger threatens your children, call the police*>

Before delving into prosodic patterns per se, we must first attend to ancillary issues such as the validity of read speech, the definition of intonation, and text normalization.

1.1 Fallacies about Read Speech

Some linguists have misgivings about read speech. Their misgivings are encapsulated in the following statement by Ladefoged (2003: 23):

> In a country in which the speakers are literate, it may be possible to ask speakers to read the list, but this is seldom a good idea. Even educated people are apt to read with a different pronunciation from their normal speech.

Gronnum (1998:132) is acutely aware of these misgivings. For this reason, she felt the need to apologize for basing her findings of Danish intonation patterns on read speech:

> The material is to a major extent composed of severely limited and manipulated utterance types, i.e., typical "laboratory speech." Such a procedure may seem inappropriate in view of the fact that the final goal is a description of the intonation of spontaneous speech. However, the method may be defended on at least two grounds. Firstly, it is convenient to investigate the course of F0 in syntactically and pragmatically simple structures which have been produced under controlled circumstances, because this allows you to single out the parameter under scrutiny without interference from other factors which may influence F0. Secondly, you may reasonably expect that natural, spontaneous speech can be described, at least to a certain extent, with the same categories and prosodic structures which have been discovered in edited, read speech.

In my considered opinion, these qualifications are unwarranted because it is a fallacy to believe that one can record spontaneous speech that is "authentic." There are several important constraints and legal restrictions that make authentic spontaneous speech unobtainable. First, the mere fact that people know that their speech is being recorded for linguistic analyses automatically disqualifies such recordings from being authentic because participants are inclined to be on their best linguistic behavior. Secondly, in the USA, and probably in other countries, it is against state and/or federal laws to record people's speech without their consent. Thirdly, even if the participants consent to be recorded "spontaneously" anytime and anywhere, ambient noises can cause such recordings to be unsuitable for acoustic phonetic analysis. Given these considerations, prosody researchers find themselves between a rock and a hard place, having to trade "authentic" speech for "laboratory" speech. Fortunately, the harshest critics of read speech have softened their objections because, as it happens, some of the acoustic phonetic issues being investigated are beyond the conscious control of speakers. When people are reading written texts, they are unaware of what the researcher is looking for. So, even if they monitor their speech, their behavior does not invalidate the findings. Boberg (2021:1-57) provides many compelling reasons for why the vowels produced by American actors and actresses are valid data for sociophonetics research. In fact, when we watch movies, television, listen to the radio, or hear an influencer speak, we forget that their speech has been rehearsed at least once. Unless something is out of the ordinary, their intonation does not strike us as unnatural or unauthentic. It follows from these observations that read texts can be used to legitimately investigate intonation.

1.2 Clarification of Concepts

Definitions are important because they provide a common starting point. In theory, **intonation** should be a very simple concept to define because every time humans open their mouths to speak, intonation gushes out. Because of its pervasiveness in human verbal interactions, we intuitively know what it is. Yet, experts continue to grapple with defining it. They are stumped by it partly because intonation means different things to different people. Hirst (1998:3-4) explains the ambiguity as follows:

> The first ambiguity depends on whether or not intonation is defined in a *broad sense*, that is, as including factors such as word-stress, tone and quantity which can be an essential part of the lexical identity of words, or in a *narrow sense*, as excluding such factors. …. The second ambiguity depends on a distinction between levels of analysis and description. In phonetics, as in all sciences, a distinction may be made between the *physical level*, that of observable and measurable physical parameters, and *the formal level*, which is a rather more abstract level of representation set up as a model in an attempt to describe and explain the observed data [italics added for emphasis].

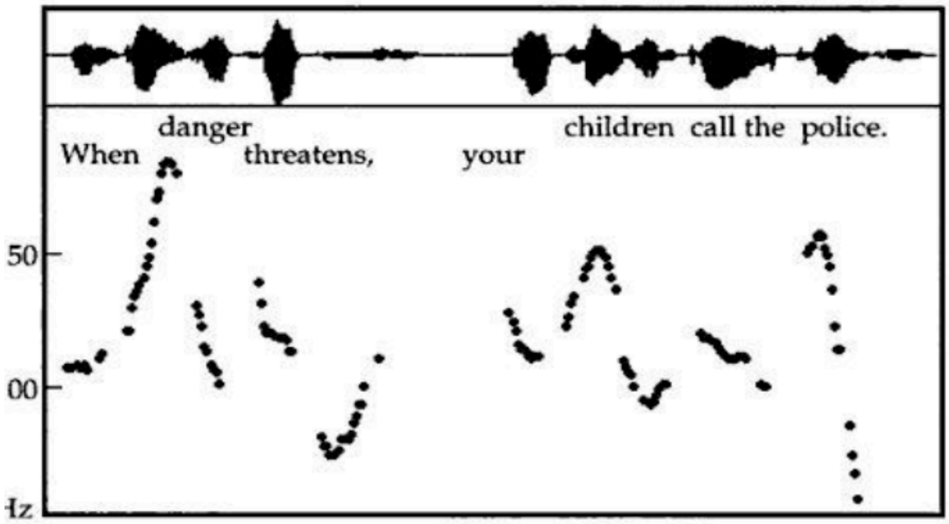
I agree with Svetozarova (1998:264-5), Botinis (1998:300), and Fonagy (1998:331), among others, that the failure to distinguish clearly between various types of intonation has obscured what it is and how to go about studying it. For this reason, let me state my position clearly. I subscribe to the **narrow** sense of intonation. I also study intonation at the **physical** level. Furthermore, I agree with Ladefoged (2001:12) and Benkirane (1998:353) who narrow the scope of intonation further *to prosodic patterns of* ***words*** *in an utterance*. This means that phenomena such as pitch-accent and syllable stress are outside the province of my approach to intonation.

Now that intonation has been defined and my position clearly stated, let's turn our attention to Utterances 1 and 2 that will be used in the remainder of the paper for various demonstrations. These utterances are taken from Ladefoged (2001:12-24) in which he describes the intonation patterns of eight utterances:

1. *<When danger threatens, your children call the police>*
2. *<When danger threatens your children, call the police>*
3. *<I'm going away>*
4. *<Where are you going?>*
5. *<Are you going home>*
6. *<Are you going away?>*
7. *<Jenny gave Peter instructions to follow.>* (i.e., Peter is to follow the instructions)[1]
8. *<Jenny gave Peter instructions to follow.>* (i.e., Peter is to follow Jenny)

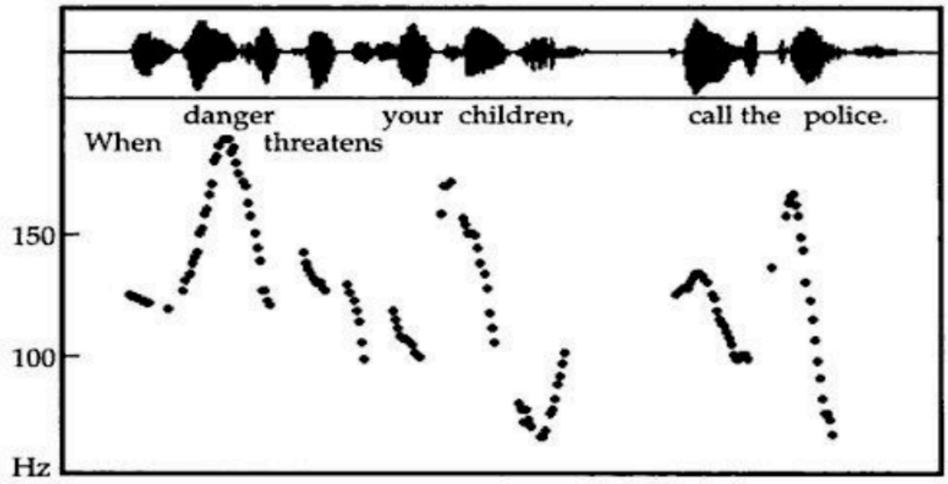
We limit ourselves to Utterances 1 and 2. Yet, the methodology developed here can be used to describe the six other sentences. Furthermore, it can be used to describe the prosodic patterns of utterances in any language. The two utterances under investigation are represented by waveform and pitch tracking as in Figures 1 and 2:

[1] Presumably, intonation patterns can help disambiguate utterances 7 and 8. However, Shattuck-Hufnagel and Turk (1996:196-7) indicate that this is not always the case, because in some cases, there is no one-to-one correspondence between syntax and intonation. We do not confirm nor deny the presence or absence of such isomorphisms in this paper.



(Ladefoged, 2001, p. 17)

Figure 1: Utterance 1



(Ladefoged, 2001, p. 17)

Figure 2: Utterance 2

2.0 Pitch-Centric Approach to Intonation

Intonation is half tamed because the vast majority of studies have focused only on pitch at the exclusion of intensity and duration. Yet, we know since Fry (1958), if not earlier, that speech is a three-dimensional physical entity that includes pitch, intensity, and duration. All three are independent of each other, yet they are interdependent. When humans hear speech sounds, they cannot consciously turn off intensity and duration, and hear only pitch. Similarly, they cannot hear only intensity, and deactivate duration and pitch. Lastly, they cannot focus only on duration and ignore pitch and intensity. All three acoustic correlates are experienced by the mammalian ear at once. Given this auditory reality, it is no wonder that intonation is only half tamed. The inability to tame intonation completely can be blamed on the fact that most of the intellectual efforts have gone into describing pitch alone, as though it were the only correlate that matters when we hear utterances. The analytical priority given to pitch has hampered investigations in the intensity and duration domains. For example, in Ladd's (2008) 349-page book, intensity is mentioned only on six pages. Not a single measurement of intensity is used in any of the descriptions of intonation. Hirst and Di Cristo (1998) edited an important volume on the intonation systems in 20 languages. The invited list of contributors represented the Who's Who of intonation studies before the turn of the 21st century. Yet, not a single author provided an intensity measurement in their analyses. Intensity is sometimes mentioned casually as a descriptive adjective, not as an acoustic correlate. The same can be said for duration, except that it was given slightly more attention. Even so, it was not given the attention it deserves.

Chapter 2 of Ladefoged (2001) is entitled *Pitch and Loudness*. Naturally, one would think that intensity will be discussed on an equal basis with pitch. However, the 13 illustrative figures such as Figures 1 and 2 (above) displayed only pitch tracings. Not a single tracing was provided intensity or duration. In fact, the definition that Ladefoged (2001:12) gave of intonation reinforces the pitch-centric bias. He wrote that "a difference in pitch that changes the meaning of a group of words is called a difference in intonation." The third edition of the same book (2012), now co-written with Disner after Ladefoged's passing in 2006, contains the same illustrations and the same definition. Astruc (2015:126-139) contributed an informative article to the edited volume, *The Bloomsbury Companion to Phonetics*, on the general topic of prosody. The author devoted separate sections to duration, intensity, and pitch. However, when it came to discussing intonation, intensity and duration were left out because she defined it as "the use of prosodic features, especially pitch, to convey differences in meaning at the sentence level and above." The pitch-centric approach permeates the linguistic literature.

2.1 Speech Synthesis and Intonation

Some people refer to Fry (1958) as the reason for excluding intensity and duration from their study of intonation. However, Fry's study was based on the auditory perception of lexical stress on homographic words. He stated many times throughout his article that he was not researching English intonation per se. His correlate rankings applied directly to lexical stress, not intonation. More will be said about this in 5.0. The contemporary understanding of the limitations of the pitch-centric approach comes from Klatt (1987). His desire to produce synthetic speech that is intelligible and natural forced him to approach intonation comprehensively, as explained in this quote:

> A pure tone can be characterized in physical terms by its intensity, duration, and fundamental frequency. These induce the sensations of loudness, length, and pitch, respectively. In speech, it is the change over time in these prosodic parameters of intensity, duration, and F0 that carry linguistically significant prosodic information, (Klatt 1987:759-760).

Notice that, unlike most linguistic definitions of intonation, Klatt's definition includes all three correlates, not just pitch. His subsequent publications underscore the importance of all three correlates. The desire to produce naturally sounding TTS systems has raised awareness that prosody consists of three inextricably bound and inseparable correlates. Consequently, if we desire to tame the half savage portion of intonation, we must also pay attention to intensity and duration, not only to pitch. Bunnell (2022:14) notes that the desire to account for expressiveness in synthetic speech is forcing researchers to explore other dimensions of intonation. In a way, speech synthesis is compelling linguists to re-evaluate old practices and assumptions about prosody.

2.2 Psychoacoustics and Intonation

Advances in psychoacoustics regarding how humans process speech signals have made the pitch-centric approach to intonation untenable. Yost (2007:223) reports, for example, that auditory neurons are hyper-specialized. Some process pitch information, some intensity information, and some duration information. All this takes place in the Central Auditory Nervous System (CANS) where variegated data is integrated. The three important phases in the auditory processing of speech signals, namely, **discrimination**, **integration**, and **resolution** involve all three correlates. It follows from advances in neuroacoustics that intonation is half tamed because intensity and duration have been left unaccounted for. There are hints in Bolinger (1978:476-477) that all three correlates are involved, even if he was dismissive of intensity. He stated that "Intensity has been the most overrated of the three major correlates of prominence, of which the two others are pitch and duration." As for duration, he off-handedly acknowledged that it could potentially play a role, saying, "Rhythm is of course the system of repeating or alternating durations and their rate of succession. It appears to play a secondary role, supporting (and sometimes replacing) other parts of the prosody." After mentioning these two correlates as side comments, Bolinger went on to devote 53 pages of his article to pitch, highlighting its role in intonation. Shattuck-Hufnagel and Turk (1996) also mention all three correlates, but they foreground F0 while paying lip service to the other two, especially intensity. If the wild side of intonation is ever to be tamed, a holistic methodology will be required. This is what I seek do in the remainder of the paper.

3.0 A Holistic[2] Acoustic Approach to Intonation

The best way to deal comprehensively with intonation is to appeal to psychoacoustics because it embraces all three correlates equally and uses them equally to account for speech events. As far as academic disciplines go, psychoacoustics is relatively new because it is less than a century old. Yost (2015) traces its birth to the 1960s and to Harvey Fletcher for the groundbreaking research that he and his team conducted at the Research Laboratories of the Bell Telephone System. He summarizes the advances from that era as follows:

> Fletcher oversaw a litany of psychoacoustic research achievements unmatched in the history of the field, which included measurements of the auditory thresholds (leading to the modern-day audiogram, the gold standard for evaluating hearing loss), intensity discrimination, frequency discrimination, tone-on-tone masking, tone-in-noise masking, the critical band… Two of the more important psychoacoustic contributions of the Bell Laboratories years are the critical-band and equal-loudness contours (Yost 2015: 49).

Physicist Harvey Fletcher published a seminal paper in 1940 in which he posited, based on mathematical calculations, the frequency response of the basilar membrane to auditory signals. His analyses led to the advent of the **Critical Band Theory**. Fletcher became a touring figure in psychoacoustics. His findings led to breakthroughs in many areas, including audio engineering. Physicist Georg von Bekesy is also a touring figure in psychoacoustics. His 1947 article describes some of the ingenious experiments that he conducted to verify if Fletcher's mathematical model of the basilar membrane was grounded in anatomical reality. He was awarded a Nobel Prize in 1961 in Physiology/Medicine for his breakthroughs on how the human ear perceives and processes auditory signals.

[2] The simpler term "holistic" is used here instead of the more technical phrase "Equal Energy Hypothesis" that appears in the acoustics literature on noise and vibration. EEH assumes that the acoustic correlates of frequency, intensity, and duration are co-equal. If so, then any one of them, or in a combination, can be used to encode intonation. This also means that none of them should be, a priori, excluded.

These major accomplishments and subsequent ones make it difficult to provide a succinct definition of psychoacoustics. Fortunately, Fastl and Zwicker (2007:VII) have given us a working definition of it as an attempt to find "the correlation between acoustical stimuli and hearing sensations [that] is investigated by acquiring sets of experimental data and by models which simulate the measured facts in an understandable way." For our purpose, the key phrase in this definition is "understandable way." The way in which psychoacousticians have sought to make their findings understandable is by means of Just Noticeable Difference (JND) thresholds. Their experiments have involved hundreds, if not thousands of people, for nearly 100 years. The results have led to the discovery of auditory thresholds at which properties of speech are optimally perceived. Many of the JNDs have been endorsed by reputable national and international regulatory bodies such as the American National Standards Institute (ANSI), the International Standardization Organization (ISO), and the International Electrotechnical Commission (IEC) for the manufacturing of audio devices. The three JNDs that are directly relevant to the study of prosody are listed below:

Auditory Discrimination in Pitch/F0[3]

Of two speech signals **A** and **B**, **A** is perceived auditorily as having a higher pitch than **B** if and only if there is a difference of 5 Hz or more between them.

Auditory Discrimination in Intensity

Of two speech signals **A** and **B**, **A** is perceived auditorily as being louder than **B** if and only if there is a difference of 3 dB or more between them.

Auditory Discrimination in Duration

Of two speech signals **A** and **B** lasting less than 200 msec, **A** is perceived auditorily as being longer than **B** if and only if there is a difference of 10 msec or more between them.[4]

More will be said about these JNDs later in the paper. Suffice it to say for now that when JNDs such as these are used, they are automatically statistically significant because, for a JND to be considered valid, it must clear at least 75% of correct responses (Stevens 2000:225, Houtsma 1995:271). It is also worth mentioning that these JND thresholds are universal because they work for all human beings, irrespective of their native languages. This means that, though they are used to account for intonation patterns of English sentences, the same methodology and insights can be used to account for prosody in any human language. However, before applying them to describe Utterances 1 and 2, we must attend first to a few procedural matters having to do with **text normalization**.

4.0 Text Normalization

The phrase "text normalization" is used to describe the preprocessing steps that take place before a text is made ready for TTS synthesis (Bunnel 2022:15). The phrase is applied to intonation analysis to describe the kind of preparatory work that is necessary before a full-blown analysis is carried out. Two preprocessing steps are required. The first deals with **"linguistic tokenization,"** and the second with **"acoustic tokenization."** Both are described succinctly below and are applied to Utterances 1 and 2.

[3] There are two different JNDs for F0. One focuses on lexical stress and the other on intonation. For lexical stress, the nucleus of the stressed syllable must be ≥ 1 Hz higher than all other nuclei in the same word. However, for intonation, the word whose F0 is ≥ 5 Hz higher than all other words in the utterance is the one that receives prosodic prominence. These two JNDs must not be confused.

[4] Additional refinements will be provided in 7.0.

Utterance 1: <*When danger threatens, your children call the police*>
Utterance 2: <*When danger threatens your children, call the police*>

4.1 Linguistic Tokenization

Linguistic tokenization addresses a sundry of issues that have a direct bearing on the utterances under consideration. Among them are issues dealing with the syntactic structure of the utterance. This consideration leads generally to the classification of utterances into three types, depending on the number of main verbs they contain:

1. Simple
2. Compound
3. Complex

The easiest way to determine the type to which an utterance belongs is by counting the number of main verbs in it. If an utterance has one main verb, it qualifies as **simple sentence**. If it has two or more main verbs, a coordinating conjunction, a correlative conjunction, or a conjunctive adverb, it is a **compound sentence**. A **complex sentence** is one that has at least two main verbs and a subordinating conjunction. In this regard, Utterances 1 and 2 are classified as complex sentences because they each contain a subordinate clause <*when danger threatens*> or <*when danger threatens your children*>. Also, they each have a main clause, <*your children call the police*> or <*call the police*>. The information structure of Utterances 1 and 2 is also noteworthy because the subordinate clause precedes the main clause (see 9.6 for additional discussions).

Tokenization also helps to identify the grammatical mood of the utterance because modality has a direct impact on intonation (Hirst and Di Cristo 1998:24). Utterances of all types can be classified according to five **grammatical moods**:

1. Declarative/indicative
2. Interrogative
3. Imperative
4. Conditional
5. Subjunctive

Utterances 1 and 2 belong to two different grammatical moods. The verbs <threatens> and <call> in Utterance 1 are both in the declarative mood. However, the two verbs in Utterance 2 belong to two different moods. The verb <threaten> is in the declarative mood, while the verb <call> is in the imperative mood.

Linguistic tokenization also calls for identifying the **parts of speech** of the words in an utterance. This is important because the parts of speech of the words in utterances can dictate the shape of their intonation pattern. Eight parts of speech are found in English and most languages. They are:

1. Nouns
2. Verbs
3. Adjectives
4. Adverbs
5. Pronouns
6. Prepositions

7. Conjunctions
8. Articles

These parts of speech are divided into **content words** (major parts of speech) and **function words** (minor parts of speech). Utterances 1 and 2 contain five content words <*danger, threatens, children, call, police*> and three function words <*when, your, the*>. We will see in 7.1 that differentiating between content and function words plays an important role in calculation of the rhythmicity of utterances. Another aspect of linguistic tokenization is identifying the grammatical function of the words in utterances, as displayed in Table 1:

No.	Words	Utterance 1	Utterance 2
1.	when	subordinating conjunction	subordinating conjunction
2.	danger	subject noun	subject noun
3.	threatens	transitive verb used absolutely	transitive verb
4.	your	possessive adjective	possessive adjective
5.	children	subject noun	direct object noun
6.	call	transitive verb, indicative mood	transitive verb, imperative mood
7.	the	definite article	definite article
8.	police	direct object noun	direct object noun

Table 1: Grammatical Analysis of Utterances

We see that, even though the two utterances contain the same words, the shaded grammatical functions differ. These differences can potentially be significant for the intonation patterns of both utterances. If the analysis warrants it, tokenization can involve paying attention to the phrasal categories that an utterance contains. In fact, di Cristo (1998:200) is of the opinion that phrasal boundaries correspond to rhythmic units. Five **phrasal categories** are universally recognized, which are listed as follows:

1. Noun Phrase (NP)
2. Verb Phrase (VP)
3. Adjective Phrase (AdjP)
4. Adverbial Phrase (AdvP)
5. Prepositional Phrase (PP)

In contemporary syntax, the best way to display the constituents of utterances is by means of tree diagrams such as Figures 3 and 4. The symbol "**S**" for sentence is increasingly being replaced by "**IP**" which stands for "Intonation Phrase." Normally, the highest IP dominates lower IPs. However, in the diagrams below, "IP" is used only once as a synonym of "S." Lower constituents are labelled in accordance with conventional syntactic analyses.

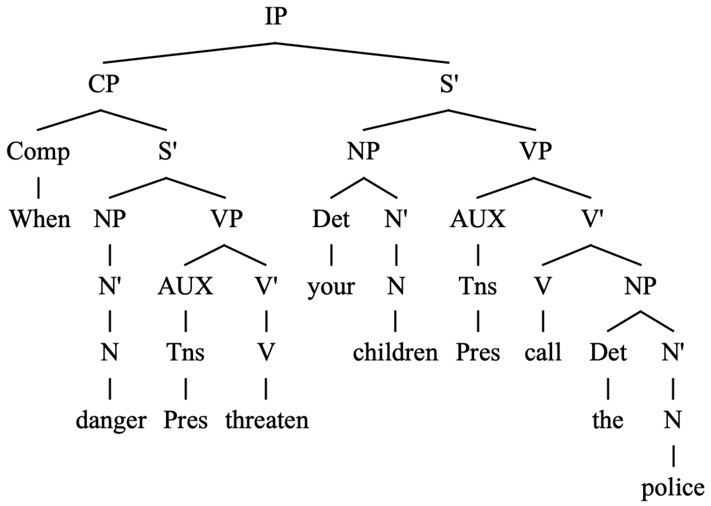


Figure 3: Tree Diagram of Utterance 1

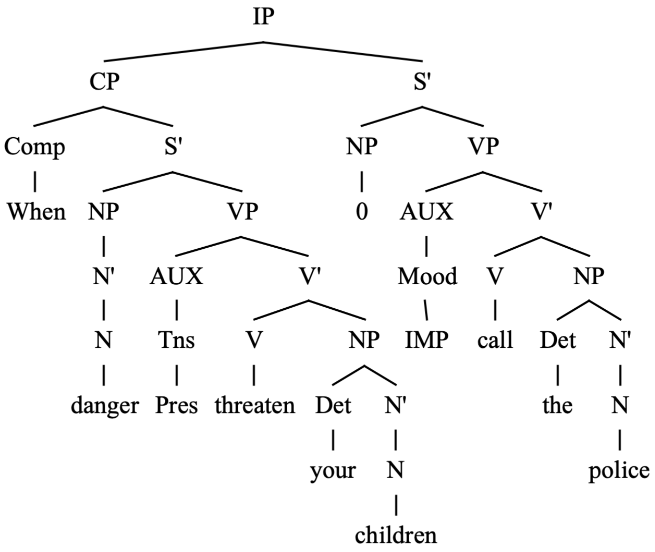


Figure 4: Tree Diagram of Utterance 2

Care should be taken not to turn tokenization analysis into a syntax paper. It is meant to simply highlight the information structure and key syntactic elements that are relevant for the intonation pattern of the utterance. I covered many topics in this demonstration in order to provide a good overview of topics that are relevant during the text normalization portion of intonation analysis.

4.2 Acoustic Tokenization

Acoustic tokenization can be subdivided into two parts: the counting of minor **intonational phrases** within an utterance and, if necessary, **spectrographic editing**. Any given utterance may contain one, two, three, or more intonational phrases, as illustrated by sentences 6a through 6d in Shattuck-Hufnagel and Turk (1996:197). The number of intonational phrases in an utterance is determined simply by counting the number of demarcative pauses. What is a **demarcative pause**? For written utterances, it usually corresponds to the number of punctuation marks (commas, semi-colons, colons, periods, etc.). For spoken utterances, one simply counts the number of places where the talker pauses. Klatt (1976:1210) found that "pauses make up about 20% of the time during fluent reading, and a good deal more, about 50% of the time, in a conversation."

Another concept that is closely associated with intonational phrase is that of **interlexical link calculations**. It has to do with a basic count of the words in an utterance. An important aspect of intonation is to account for the pitch, sonority, and rhythmicity modulations and variations between **two** consecutive words. Shattuck-Hufnagel and Turk (1996:229) write that the expanded ToBI transcription system is now paying attention to adjacent pairs of words. However, as will be explained in 5.0 and subsequent sections, the approach discussed in this paper takes the analysis to a whole new level of rigor. One counts the number of words in an utterance in order to arrive at the correct number of possible interlexical links. The calculation is based on the following formula:

Interlexical links = Number of Words – (minus) Number of Demarcative Pauses

If an utterance contains 10 words, and only one period at the end, then there are 9 possible interlexical links. This corresponds also to the number of possible auditorily perceptible pitch or sonority movements in the utterance. If the utterance has 10 words, a comma, and a period, then there are 8 possible interlexical links, and so on and so forth. The number intonational phrases and their position in utterances matter a lot in prosody analysis. For example, even though Utterances 1 and 2 have 8 words and 6 interlexical links each, they differ prosodically because their intonational phrases contain different number of interlexical links. In Utterance 1, the first intonation phrase contains three words, <*when, danger, threatens*> and two interlexical links, while the second has five words, <*your, children, call, the, police*> and four interlexical links. In Utterance 2, the first intonation phrase has five words, <*when, danger, threatens, your, children*> and four interlexical links, while the second has three words, <*call, the, police*> and two interlexical links. These differences are displayed and illustrated as follows:

Utterance 1: <*When —danger —threatens// your —children —call —the— police#*>
Utterance 2: <*When —danger— threatens —your —children// call —the —police#*>

By convention, slashes "//" stand for minor pauses, the pound symbol "#" is used for major pauses, and the dashes "—" represent interlexical links. The latter is an adaptation of the use of dashes in historical reconstruction, as explained by Fromkin et al. (2017:352-353). This analysis makes it possible to determine the possible number of pitch and sonority movements in an utterance. In both Utterances 1 and 2, there are six possible pitch movements and six possible sonority movements. More will be said about this in sections 5.1 and 6.0. Suffice it to say that this process is very important for calculating the relative functional load (RFL) of pitch and sonority in utterances. This process is also essential for calculating the overall melodicity score of the utterance, as discussed 8.0. Once the intonational phrases and interlexical links of an utterance have been accounted for, acoustic tokenization is deemed complete, and another phase of the intonation analysis can proceed, unless there are issues such as the one in Figure 5 that one must attend to.

The spectrographs of Utterances 1 and 2 in Figures 6 and 7 do not present any problem upon visual inspection. However, for the purpose of illustration, I am using another utterance by the same female speaker that has an issue that can derail the accuracy of an intonation analysis. The issue of concern, and to which analysts must pay close attention, is what is commonly referred to as **octave error** or **spurious pitch**. Pitch values that are to be used in intonation analyses are those that originate from the glottis, as a result of the vibration of vocal folds. However, occasionally but frequently enough to deserve mention, the pitch tracking algorithm finds pitch where none should exist. Himmelmann and Ladd (2008:270) offer the following explanation for why a spurious pitch may occur:

> Finally, it is important to remember that automatic F0 extraction is based on mathematical algorithms applied to the digitized acoustic signal, not on human pattern recognition. These algorithms can occasionally be fooled and give *spurious F0 values*. The most important case is that of "octave errors," in which the reported F0 value is exactly twice or exactly half what it should be … Octave errors can sometimes happen for no apparent reason, but they are often associated with *slightly irregular phonation* [italics added for emphasis].

In the utterance <*Jenny gave Peter instructions to follow*> produced by the same female speaker, a spurious pitch is found in the area within the red circle (color online), in Figure 5.

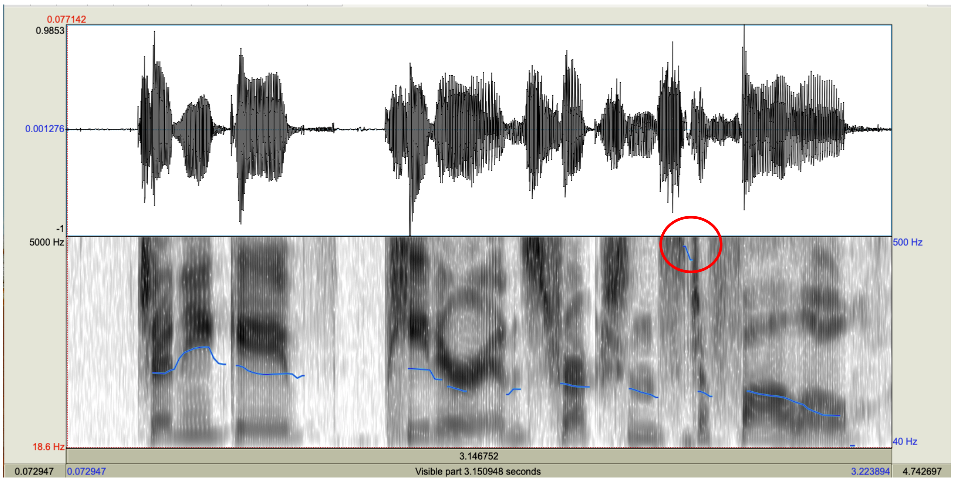


Figure 5: Example of an Octave Error [Red circle online]

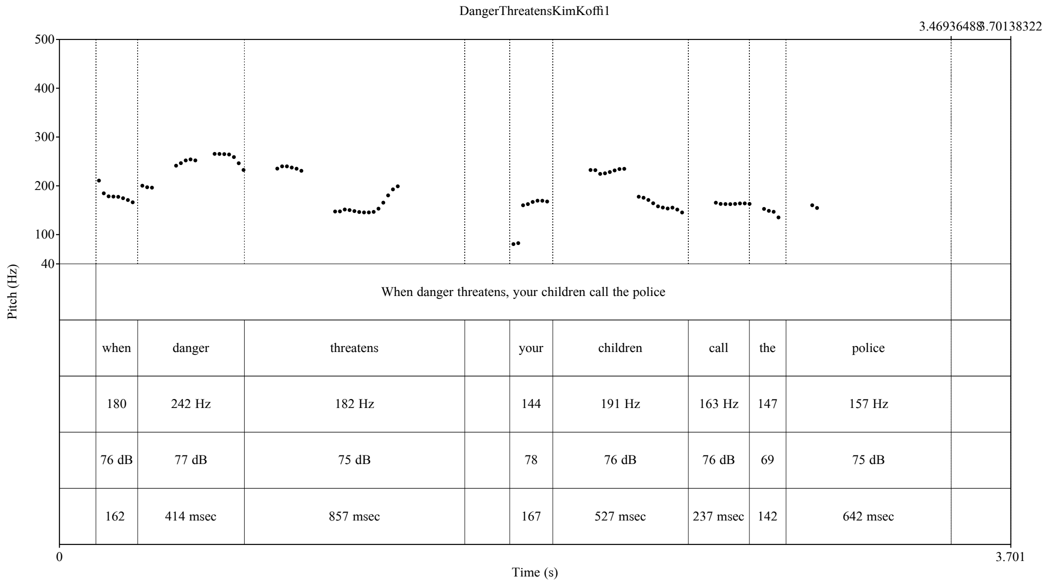
The speaker began <Jenny> with a mean F0 value of 258 Hz. However, in the spectrograph itself, inside of the red circle, we see another pitch whose mean value is 478 Hz. This is an octave error that must be removed, otherwise it will cause the analysis to be erroneous. See Appendix 1 for how to remove octave errors in Praat.[5] Octave errors are one of the reasons why some experts are leery about bulk extraction of acoustic correlates:

> Most forensic-voice-comparison researchers and practitioners employing formant measurements would use human-supervised measurement. This is generally considered to provide *greater validity and reliability* than a fully automatic measurement but requires a *greater investment in human labor* and is not itself free of difficulties, (Zhang et al. 2013:797) [italics added for emphasis].

[5] A quicker solution for removing octave errors consists simply in changing the pitch setting in Praat. Usually setting the pitch range from 75-200 Hz will do for adult males and setting it from 75-250 will do for adult females.

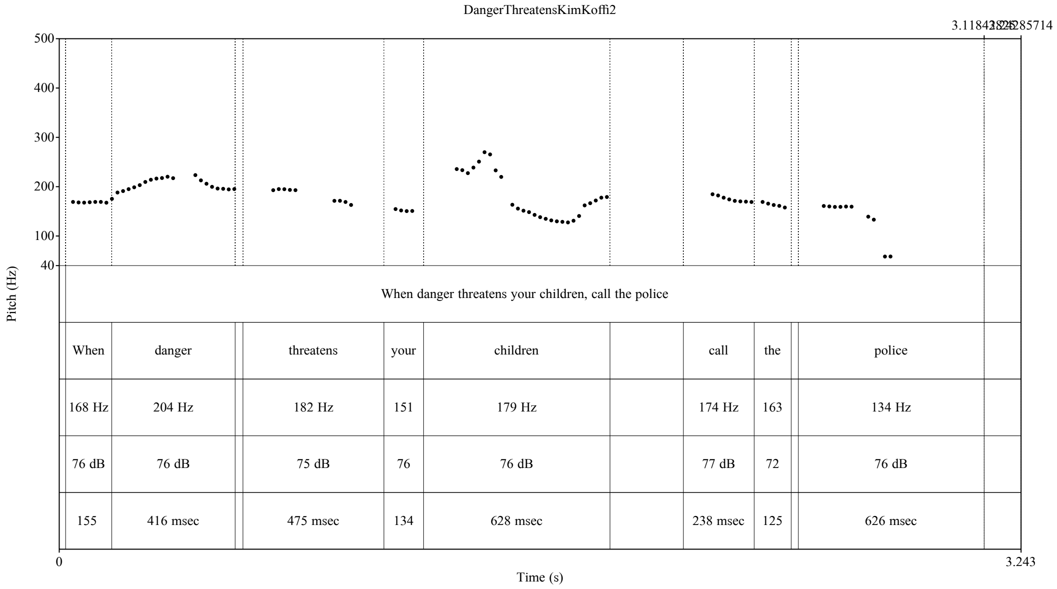
Baken and Orlikoff (2000:3) have issued six principles to guide acoustic phonetic analyses. Principle 5 is stated as follows: “Never trust a computer completely.” This is a fair warning that should not go unheeded in prosody studies.

Another aspect of text normalization consists of removing reading dysfunctionalities. Many times, the reading goes smoothly. But there are times when some readers stumble over their words, skip words, etc. In some situations, it is not feasible to have the reader re-record himself/herself. If words are skipped from a standardized elicitation paragraph, this must be noted. When one is done with editing the spectrograph, and everything is as clean as possible, the next step in the acoustic tokenization process is to **segment** the utterance into words and **annotate** the spectrographs for correlate extraction. This step is illustrated by Figures 6 and 7:



	When danger threatens, your children call the police									
	when	danger	threatens		your	children	call	the	police	
	180	242 Hz	182 Hz		144	191 Hz	163 Hz	147	157 Hz	
	76 dB	77 dB	75 dB		78	76 dB	76 dB	69	75 dB	
	162	414 msec	857 msec		167	527 msec	237 msec	142	642 msec	

Figure 6: Annotated Spectrograph of Utterance 1



When danger threatens your children, call the police										
When	danger		threatens	your	children		call	the		police
168 Hz	204 Hz		182 Hz	151	179 Hz		174 Hz	163		134 Hz
76 dB	76 dB		75 dB	76	76 dB		77 dB	72		76 dB
155	416 msec		475 msec	134	628 msec		238 msec	125		626 msec

Figure 7: Annotated Spectrograph of Utterance 2

Now that Utterances 1 and 2 have been fully normalized and the needed measurements have been extracted, we can embark on a holistic approach of intonation, rather than the pitch-centered one that researchers usually focus on. The JNDs listed in 3.0 are used to interpret measurements, calculate RFLs, and determine the melodicity score of utterances.

5.0 The Role of Pitch in a Holistic Study of Intonation

The JND for the auditory perception of pitch is ≥ 5 Hz. According to t'Hart (1981:812), this JND has a long history that goes as far back as 1931, if not further. However, since the experiments that may have led to its discovery are unknown, it is better to limit ourselves to Fry (1958). Pages 141 to 144 describe his experiments and findings. Here is a quick summary. Fry wanted to know the acoustic correlates that underly peoples' perception of homographic pairs such as <**per**mit> (noun) and <per**mit**> (verb). He recruited 43 participants and conducted several auditory perception experiments. He controlled pitch levels by various increments. He found that when frequencies on two different words differed only by 3 Hz, the participants did not reach a consensus as to which word had a high pitch and which had a low pitch. However, when he increased the pitch levels by 5 Hz, a clear consensus emerged among the participants. He did more testing by varying F0 levels by increments of 5 Hz, that is, 10 Hz, 15 Hz, 20 Hz, and so on and so forth up to 60 Hz. He discovered that, no matter the levels of subsequent increases in F0, the consensus about pitch detection remained the same as when he increased it only by 5 Hz. He, therefore, concluded on page 141 that "Increase in the size of the frequency step appears to produce no mark trend in the results." The significance of the ≥ 5 Hz threshold is also discussed by Houtsma (1995:277). More recently, it has been verified by Liu's (2013) experiments. Jongman et al. (2017) have appealed to it in their analysis of pitch detection by Mandarin and English listeners.

Let's apply the JND of ≥ 5 Hz and the insights from Fry's experiment to Utterances 1 and 2, as displayed in Figure 8:

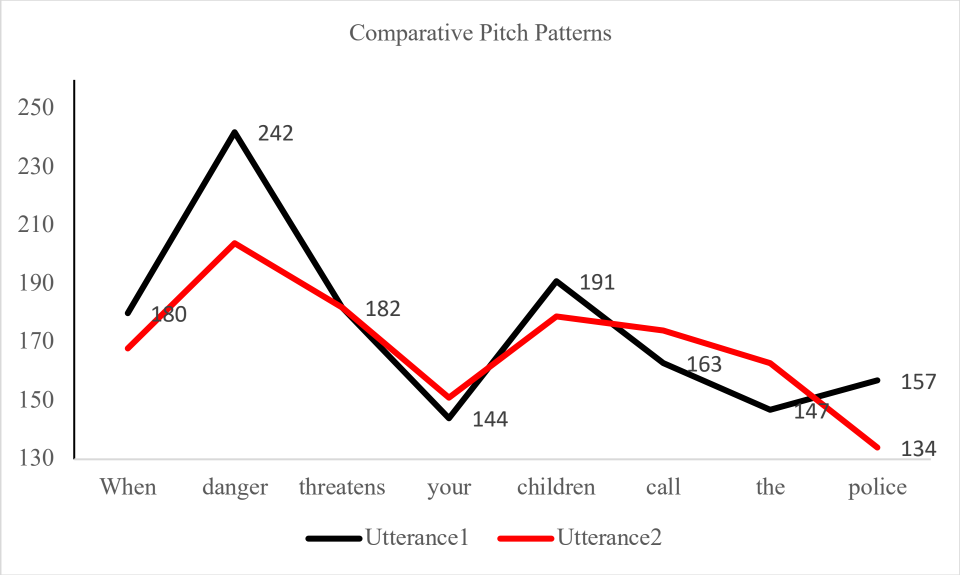


Figure 8: Pitch Movements

When the two utterances are placed side by side in the same graph, we see how they resemble each other in pitch. The beginning of both utterances is similar in that both begin with a **cone-shaped,** delta **Δ pattern**. The word <*danger*> is the peak of the rise. We also see a **V-shaped** pattern in the middle of the utterance where <*your*> is the trough of the fall. The only notable difference between the two utterances is that Utterance 1 concludes with a terminal rise, while Utterance 2 has a terminal pitch fall. The latter is indicative of the so-called “Valley Girl Speech” or “uptalk” (Curzan and Adams 2009:129). The claim that the intonation patterns in Figure 8 are similar, except for their endings, may come as a surprise to many because they assume that visual displays of pitch correlate with auditory reality. However, when interpreting pitch, “seeing is **not** believing.” The magnitude of pitch movements is not to be interpreted literally. Fry (1958:151) explains why:

> Change in fundamental frequency differs from change of duration and intensity in that it tends to produce an *all-or-none effect*, that is to say, the magnitude of the frequency change seems to be relatively unimportant while the fact that *a frequency change has taken place is all-important* [italics added for emphasis].

In other words, the difference of 62 Hz between <*when*> (180 Hz) and <*danger*> (242 Hz) in Utterances 1 and the difference of 36 Hz between <*when*> (168 Hz) and <*danger*> (204 Hz) in Utterances 2 is not perceived by the naked ear as significant. What matters most is that a difference of 5 Hz has been detected. Houtsma (1995:288) provides an additional insight into why the magnitude of pitch movements is not well perceived:

> Among people appearing to possess absolute pitch, he distinguishes between genuine and acquired pitch absolute pitch skills. Possessors of *genuine* absolute pitch typically make quick absolute identifications, accurate within a semitone, with octave confusion being the principal source of errors. *Acquired* skills are behaviorally characterized by slow judgments, as if subjects are trying to recall some learned reference like A_4 for orchestra musicians or extreme vocal range singers. Given enough time, these subjects can make fairly accurate absolute pitch judgments, but if forced to respond quickly they will typically make large errors.

This means that, ordinary human beings who do not have the skills of orchestra singers are not good at detecting pitch movements. Even the great Ladefoged (2003:75) confessed his inability to perceive pitch movements accurately, “I’ve never been very good at transcribing tones. … I’ve always found it hard to give good descriptions of subtle changes in pitch.” Numerous experiments, including t’Hart (1981:818), have shown that people are for the most part not very good at detecting pitch movements. Furthermore, some people are better at perceiving one type of pitch movement than the other. For example, most Dutch participants in t’Hart (1981) found it more difficult to perceive pitch falls than pitch rises. Yet, she mentioned a study that found that British speakers were better at perceiving pitch falls than pitch rises. Liu (2013:3017-3018) did an experiment involving 10 American English and 10 Mandarin. It was found that the listeners scored similarly on the perception of level tone, but they scored differently on the perception of pitch movements.

There are two main types of pitch movements: a **fall** and a **rise**. Given the JND of ≥ 5 Hz, of two words **A** and **B**, a pitch fall occurs if and only if the pitch of **A** is higher by 5 Hz or more than the pitch of **B**. On the other hand, if the pitch of **A** is lower than the pitch of **B** by 5 Hz or more, a pitch rise takes place. The actual pitch distance between **A** and **B** may be much greater than 5 Hz. However, since the naked ear perceives pitch distances on an *all-or-none* basis, what matters the most is that a pitch difference of 5 Hz has been detected. A decision tree such as the one below helps us visualize the pitch detection process:

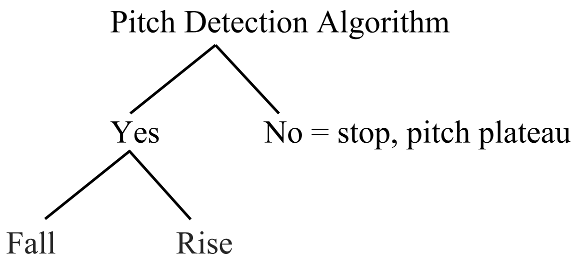


Figure 9: Pitch Detection Algorithm

This algorithm is to be interpreted as follows. Given any two adjacent words, if the pitch distance between them is less than 5 Hz, the naked ear does not perceive any pitch difference between them. This results in a **pitch plateau**. Therefore, pitch processing stops. However, if the pitch distance between two successive words is at, or exceeds 5 Hz, pitch is perceived auditorily.

5.1 Calculating the RFL of Pitch

The RFL of pitch is calculated on the basis of auditorily perceptible movements. Because of the *all-or-none effect*, we do not concern ourselves with falls or rises when calculating the pitch RFL of utterances. Rather, what matters the most is that pitch has been detected or not detected. The formula for calculating the RFL of pitch is as follows:

$$\text{RFL of Pitch} = \frac{\text{Number of Pitch Plateaus}}{\text{Number of Auditorily Expected Perceptible Pitch}} \text{ x } 100$$

The equation is formulated this way because in most utterances, there will be fewer pitch plateaus than auditorily perceptible pitch movements. Let's illustrate how the formula works by applying it to Utterances 1 and 2. The acoustic tokenization process discussed in 4.2 yielded six possible auditorily perceptible pitch movements. The pitch measurements in Figures 6 and 7 show that the F0 distances between successive pairs of contiguous words are all higher than the JND of ≥ 5 Hz. Therefore, there are no plateaus in Utterances 1 and 2. So, the RFL of pitch is 100% for each utterance.

Now, let's pretend for the sake of demonstration that the acoustic tokenization revealed that the utterance contained two pitch plateaus out of six possible auditorily perceptible pitch movements in Utterance 1. We would calculate the RFL of pitch by dividing 2 by 6, which yields 0.3333. Then, we will multiply the product by 100 because RFLs are calculated in percentages. The result will be 33.33%. We will then subtract the percentage of plateaus from 100. This will yield an RFL score of 66.66% for auditorily perceptible pitch movements.[6]

6.0 Role of Intensity in a Holistic Study of Intonation

There are a few synonyms that swirl around the intensity correlate. For this reason, a brief terminological clarification is needed. For an in-depth analysis, readers are encouraged to refer to Koffi (2020), a paper devoted entirely to a comprehensive review of intensity. **Intensity** is a theoretical acoustic concept that has various sensory correlates. The first is **amplitude**. It can be understood as intensity in reference to some distance. The second is

[6] Mathematically, it does not matter if the number of auditorily perceptible pitch is divided by the number of expected pitch movements and then multiplied by 100. The problem is stated this way because fewer pitch plateaus are expected.

loudness. It is intensity that elicits a psychological response. The third is **sonority**. It is used in reference to how the ear perceives the intensity distance between two signals. One or more of these terms are used depending on the academic discipline for which a reference to intensity is necessary. Since we are dealing with acoustic phonetics and linguistics, sonority is the synonym that matters the most. Sonority has been appealed to by phoneticians in the study of syllables. Here, we use it to account for the intensity distance between two contiguous words.

The JND of 3 dB is found literally everywhere. This is not a hyperbole. One cannot read anything about the auditory perception of intensity without seeing it. Nearly all the authors of *Hearing*, a 1995 edited volume, make mention of it. It is found in several dozen pages in Yost (2007). It is even listed on the packages and instructions leaflets of many audio products. The information displayed in Table 2 is found in the sound settings in iPhones. The JND listed below explain how the naked ear responds to variations in intensity levels.

N0	**Perception of Increases**	**Intensity Levels in dB(A)**
1.	Imperceptible change	1
2.	Barely perceptible change	3
3.	Clearly perceptible change	5
4.	Twice as loud	10
5.	Four times as loud	20

Table 2: Relative Intensity Thresholds

When the **JND** of **≥ 3 dB** is used in intonation analyses, it means that if the sonority distance between two consecutive words is less than 3 dB, the naked ear does not perceive any difference between them. The decision tree for the auditory perception of sonority can be represented as follows:

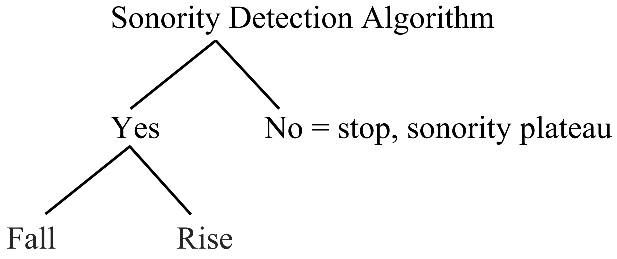


Figure 10: Sonority Detection Algorithm

Fry (1958:151) notes that the auditory response to intensity is different from its response to pitch. This was alluded to in 2.2 on how neurons specialize in conveying frequency, intensity, and duration information. Intensity is **not** perceived on an *all-or-none* basis like pitch because a sonority difference of ≥ 5 dB is perceived clearly and unambiguously. Schnitta (2016:55, Table 1) writes that a sonority difference of 7 dB amounts to an increase in sound power of 87%, while a difference of 10 dB is equal to the doubling of the intensity of the previous sound. It is worth noting that a sonority distance of 10 dB or more between two adjacent words is unlikely unless the speaker starts yelling in the middle of an utterance (see 9.5 for details). Figures 11 allows us to contrast the sonority profiles of Utterances 1 and 2.

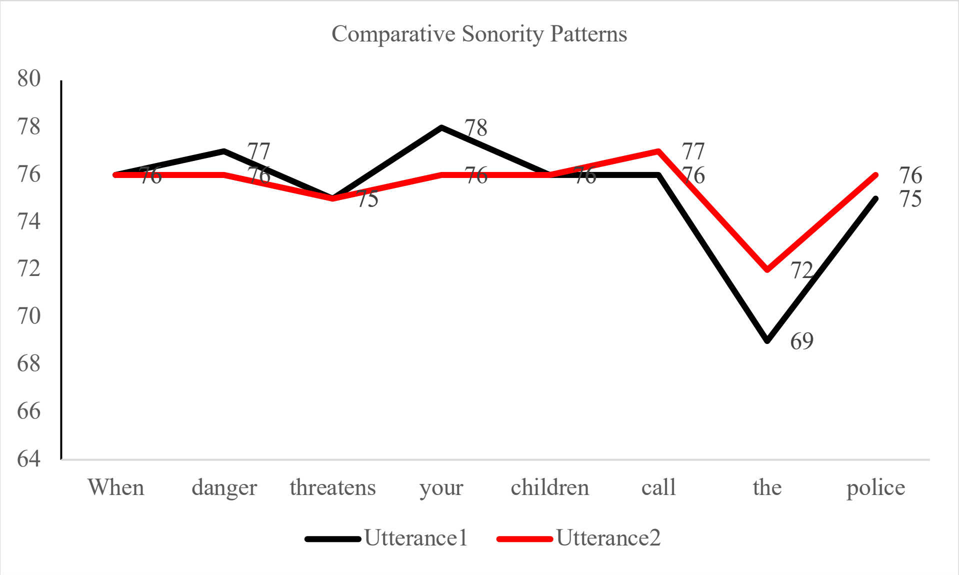


Figure 11: Sonority Movements

The shapes of the two utterances are similar. The sonority movements are similar. More importantly, both utterances end with auditorily perceptible terminal sonority rises. The initial and medial portions of the utterances are similar because the rises and falls of sonority are less than the JND of ≥ 3 dB required for audibility.

6.1 Calculating the RFL of Sonority

The formula for calculating the RFL of intensity is as follows:

$$\text{RFL of Sonority} = \frac{\text{Number of Sonority Plateaus}}{\text{Number of Auditorily Expected Perceptible Sonority}} \text{ x } 100$$

The acoustic tokenization analyses in 4.2 yielded a total of six possible auditorily perceptible sonority movements for Utterances 1 and 2. In each utterance, we see that sonority plateaus abound. Four of the six possible sonority movements are plateaus. In Utterance 1, we see only two auditorily perceptible sonority movements. The first is a fall from <call> (76 dB) to <the> (69 dB), and the second is a rise from <the> (69 dB) to <police> (75 dB). Therefore, the RFL of sonority is 4 divided by 6, which yields 0.6666. When it is multiplied by 100, the result 66.66%. We calculate the RFL of intensity by subtracting 66.66% from 100%, which yields 33.33%. The same goes for Utterance 2 in which we find only two auditorily perceptible sonority movements. We see a sonority fall from <call> (77 dB) to <the> (72 dB), and a sonority rise from <the> (72 dB) to <police> (76 dB). The RFL of Utterance 2 is also 33.33%.

7.0 Role of Duration in a Holistic Study of Intonation

Duration is an important physical correlate of speech. The way the naked ear perceives and processes variation in duration between two consecutive speech sounds is referred to as **rhythmicity**. Bolinger (1978:477) defines it as "the system of repeating or alternating durations and their rate of succession." Koffi (2021b) notes that duration measurements are to be interpreted linearly, but not literally because the naked ear perceives rhythmicity partly on a linear scale and partly on a logarithmic scale (Eddins and Green 1995:223 and Hollien 1990:29). This explains why Fry (1958:151) states that duration is **not** perceived on an *all-or-none* basis like pitch. The description of neural activity and specialization discussed in 2.2. is seen once again here. The JNDs required for the auditory perception of rhythmicity are as follows:

1. Of two speech signals **A** and **B** lasting < 200 ms, **A** is perceived as being longer than **B** if and only if the temporal distance between them is ≥ 10 ms.
2. Of two speech signals **A** and **B** lasting ≥ 200-299 ms, **A** is perceived as being longer than **B** if and only if the temporal distance between them is ≥ 20 ms.
3. Of two speech signals **A** and **B** lasting ≥ 300-399ms, **A** is perceived as being longer than **B** if and only if the temporal distance between them is ≥ 30 ms.
4. Of two speech signals **A** and **B** lasting ≥ 400-499 ms, **A** is perceived as being longer than **B** if and only if, the temporal distance between them is ≥ 40 ms.
5. Of two speech signals **A** and **B** lasting ≥ 500-599 ms, **A** is perceived as being longer than **B** if and only if the temporal distance between them is ≥ 50 ms.
6. Rhythmic **plateaus** occur when durational differences are below the JND threshold for audibility.

The JND of 10 msec is the most well-known of all. Its discovery goes as far back as 1920s when Fletcher found that the naked ear could not perceive temporal differences of less than 10 msec for signals lasting 200 msec or less. Subsequent experiments have confirmed this all-important JND. Yost (2007:148) explains the ear's inability to detect differences below 10 msec as follows, "It appears that the auditory system is not a constant-energy detector below 10 msec." It has also been found that 300 msec is an important durational landmark. Again, we turn to Yost (2007:149) to learn why:

> The duration of the signal used to establish tonal threshold is important. If it is longer than approximately 300 msec, the thresholds represent intensity in units of power; if the signal is between 10 and 300 msec, the thresholds reflect approximately constant energy; for signals of duration less than 10 msec, the spread of energy makes the determination of thresholds dependent on frequency and is more difficult to determine.

In studying intonation, we are bound to come across words of various lengths, many of which surpass 300 msec. In such cases, we turn to Quéne (2004) who proposes a rule of thumb of 10% as a JND for assessing rhythmicity for all words, no matter how long they last. In gauging rhythmicity, Klatt (1976:1211) notes that the presence or absence of demarcative pauses influences duration. He reports that syllables at or before pauses increase in duration by about 60-200 msec. This phenomenon is called **prepausal lengthening**. Let's apply the aforementioned insights to Utterances 1 and 2 in Figure 12 to assess how they compare in rhythmicity:

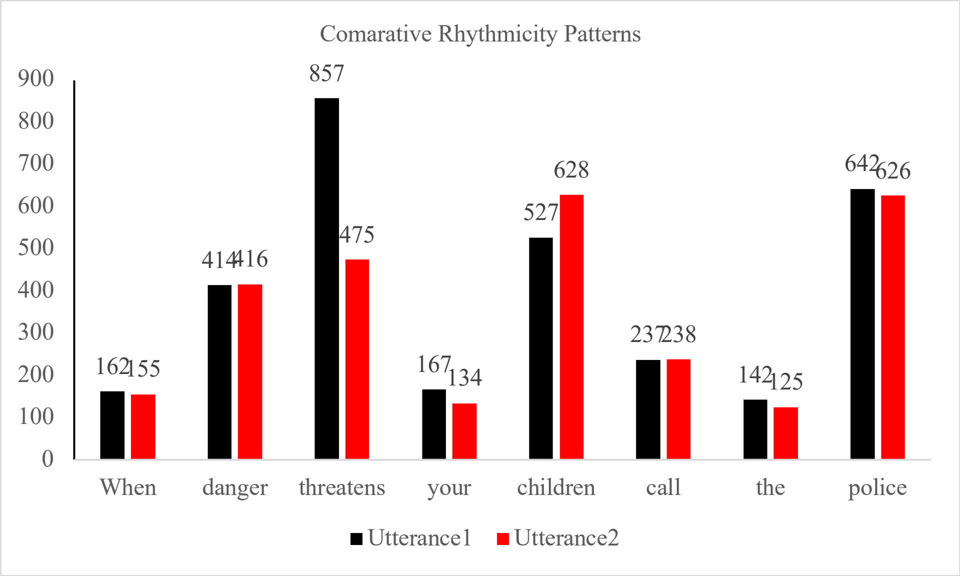


Figure 12: Rhythmicity Patterns

The claim that demarcative pauses affect duration is borne out. The word <threatens> (857 msec) in Utterance 1 is 382 msec longer than <threatens> (475 msec) in Utterance 2 just because the former appears before a demarcative pause and the latter does not. Similarly, <children> (628 msec) in Utterance 2 is 101 msec longer than <children> (527 msec) in Utterance 1. Both durational differences are greater than their respective JNDs. Since <threatens> in Utterance 1 is 857 msec, its JND threshold for audibility is 85 msec. The difference between <threatens> in Utterance 1 and <threatens> in Utterance 2 is 382 msec, which is four times greater than the JND. Similarly, the JND for <children> of 62 msec is 1.62 times greater than the audibility threshold. We see clearly that demarcative pauses affect rhythmicity.

When we examine the differences between the other words in both utterances that do not occur before demarcative pauses, we see that their differences are not auditorily salient. For example, the difference between <when> (162 msec) in Utterance 1 and <when> (155 msec) is not auditorily perceptible because it is less than 10 msec. The difference between the two <danger>s is only 2 msec. The difference between the two <call>s is only 1 msec, and the difference between the two <police>s is also not auditorily perceptible because 16 msec is below the JND 64 msec. Yet, the difference between <your> (167 msec) in Utterance 1 and <your> (134 msec) in Utterance 2 is 33 msec, which is auditorily robust.

Why is there a difference between the two <your>s? Koffi (2022) explains that when lexical items occur inside the orbit of the verb, their duration is reduced. He refers to this as the **Proximity Principle**. A difference exists because <your> in Utterance 1 occurs outside of the orbit of the verb <threatens>, whereas the one in Utterance 2 occurs inside the orbit of <threatens>, functioning as its the direct object. Since the Proximity Principle applies to <your> in Utterance 2, this explains why its duration (134 msec) is shorter than <your> (167 msec) in Utterance 1 by 33 msec. Clearly, the analysis of grammatical function undertaken in the tokenization phase of the analysis has proven relevant. If overlooked, the explanatory power of the findings is greatly diminished. Finally, we have no syntactic explanation for why <the> (142 msec) in Utterance 1 is 17 msec longer than <the> (125 msec) in Utterance 2. Klatt (1987:762) attributes such unknowable cases to "performance variables." Paralinguistic reasons unknown to researcher, and maybe even to the talkers, may lead them to adopt certain intonation patterns.

7.1 Calculating the RFL of Rhythmicity

The algorithm for detecting rhythmic differences between consecutive words can be represented as follows:

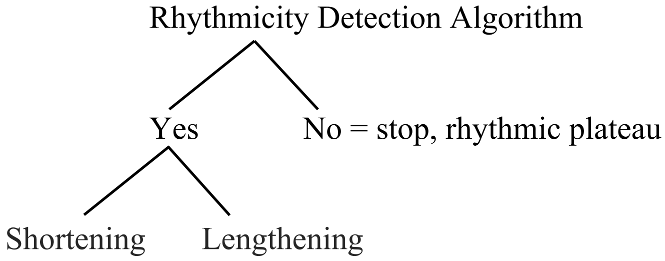


Figure 13: Rhythmicity Detection Algorithm

Once a durational difference has been detected auditorily, the next step in the analysis is to see if the rhythmic difference is one of **shortening** or **lengthening**. The determination is mostly based on JNDs. Instances of shortening can be attributed to the Proximity Principle, while lengthening is brought about by demarcative pauses. In some cases, it seems right to talk of **extra-lengthening**, especially when duration greatly exceeds expected norms. Such is the duration of <threatens> (857 msec) in Utterance 1 versus <threatens> (475 msec) in Utterance 2. The former is 1.8 times longer than the latter.

In the study of prosodic patterns, a deal of great attention has been paid to durational differences between **content** and **function** words. These two types of lexical items contribute greatly to the overall rhythmicity of utterances. This explains why the RFL of rhythmicity is the ratio of the duration of function words over the total duration of content words in utterances, as in the following formula:

$$\text{RFL of Duration} = \frac{\text{Duration of function words[7]}}{\text{Total Utterance Duration}} \times 100$$

When we apply the formula to Utterances 1 and 2, we see that the tokenization analysis showed that both utterances had each five content words, <*danger, threatens, children, call, police*> and three function words, <*when, your, the*>. All the words in Utterance 1 lasted 3,148 msec. Its five content words lasted 2677 msec, while its three function words took 471 msec to produce. So, the RFL of rhythmicity of Utterance 1 is 85.03%. The total duration of Utterance 2 is 2,797 msec. Its five content words lasted 2,383 msec, while its three function words took 414 msec. Its RFL of rhythmicity is 85.19%. The two utterances differ only by 0.16%.

This RFL difference could be taken to mean that the differences in rhythmicity between the two utterances is not important. However, this is not true. We must remember that duration calculations should not be interpreted literally. If we turn to JND thresholds, we see that the naked ear perceives rhythmic differences between Utterances 1 and 2. First, with regard to total duration, the difference of 351 msec between Utterance 1 (3,148 msec) and Utterance 2 (2,797 msec) is greater than the JND of 314 msec. Secondly, the duration of the content words in Utterance 1 (2,677 msec) is perceived auditorily as being longer than that of Utterance 2

[7] Some sentences lack a function word, such as "*Bob loves acoustic phonetics.*" In such instances, rhythmicity is calculated by dividing the shortest word by the sum of the longest words. Suppose that *Bob* is 271 msec, *loves* is 412 msec, *acoustics* 508 msecs, and *phonetics* is 658 msecs. The RFL of rhythmicity is calculated by dividing 271 by 526 (the average of 412+508+658), which is = 51.52%. Content words matter the most in the calculation of rhythmicity. So, the RFL of rhythmicity for the utterance is 48.47%, which is simply 100%-51.52%.

(2,383 msec). The difference of 294 msec is greater than the JND of 267 msec. Lastly, the temporal difference of 57 msec between the function words in Utterance 1 (471 msec) and those in Utterance 2 (414 msec) is also auditorily robust because it is greater than the JND of 47 msec. All in all, Utterance 1 is perceived as being longer than Utterance 2. The reason for this is the fact that the demarcative pause caused the extra-lengthening of <threatens> in Utterance 1.

Three cursory remarks can be made about the calculation of the RFL of rhythmicity. First, it is proportional to the sheer number of content and function words in an utterance. Utterances with more content words will have greater RFLs. Secondly, it is proportional to the number of function words in an utterance. If an utterance has many function words, the RFL of rhythmicity decreases. Thirdly, the RFL of rhythmicity is also proportional to the duration of function words. Of two utterances with the same exact function words, the one in which function words are longer is bound to have a smaller RFL. This explains why the RFL of Utterance 1 is smaller than the RFL of Utterance 2, albeit by only 0.16%. The latter point explains why the RFL of rhythmicity of native speakers of English is usually greater than that of nonnative speakers. They tend not to spend a lot of time on function words. As for non-native speakers, they often do not reduce function words. They tend to hold them longer than native speakers do. This decreases the rhythmicity of their utterances. For example, Pandey (2015:311) reports that English speakers from India produce their function words fully. In general, the RFL of rhythmicity will be smaller than that of speakers of General American English (GAE). Put it differently, if the same utterance contains the same number of words, it takes L2 speakers longer to produce them than native speakers.

8.0 Correlate Ranking

The RFL calculations of pitch, intensity, and duration make it possible to rank the contribution that each correlate makes to the overall auditory perception of intonation. When this is applied to Utterances 1 and 2, it yields the following results:

Utterance 1: Pitch (100%) > Duration (85.03%) > Intensity (33.33%)
Utterance 2: Pitch (100%) > Duration (85.19%) > Intensity (33.33%)

These rankings agree with Bolinger's (1978:477) claim that rhythmicity "appears to play a secondary role, supporting (and sometimes replacing) other parts of prosody." As for pitch, he notes on p. 516 that it "makes a contribution entirely on its own, independently of the rest of prosody." Here, Bolinger may have overstated his case by claiming that "pitch makes a contribution *entirely on its own*, *independently* of the rest of prosody" because as his previous statement indicates, sometimes duration can rank higher than pitch. Bolinger did not give any consideration to intensity, only to say that it is "the most overrated of the three major correlates of prominence, of which the two are pitch and duration," (p. 476). He concluded the brief section on intensity by noting that "[it] is not very reliable." I beg to differ for reasons discussed in Koffi (2020). Moreover, there is a big difference between saying that intensity ranks last versus saying that it is not reliable. When one listens to Utterances 1 and 2, one perceives clearly that the female talker resorted to an uptalk at the end of her utterance. The main strategy she used to accomplish this is intensity, as shown in Figure 11. Figure 8 shows that pitch does not contribute anything to the perception of uptalk in Utterance 2. The merit of an analysis such as this one is that it brings forth the contribution of each correlate to the overall auditory perception of intonation. After all, speech is a three-dimensional physical entity. Acoustic phoneticians should account for all three, not just for one or two. Now that speech synthesis has shown that all three dimensions are important (see 2.1.), it is time that linguists also pay attention to all three correlates.

8.1 Melodicity Scale and Scores

Talkers produce their utterances without being consciously aware that they are using three correlates to encode them. Similarly, hearers are not consciously aware that their ears are performing a kind of Fast Fourier Transform (FFT) analysis on incoming speech signals (Palmer 1995:75, Hartmann 1995:16). It is the job of the researcher to account for how the correlates that are emitted by talkers interact with each other until they hit the ear of hearers. Significant developments in psychoacoustics can help shed some light on all three phases of the auditory perception of intonation (Eddins and Green 1995:208). To summarize, the processes leading to the calculation of RFLs mirror what takes place in the discrimination phase during which the basilar membrane discriminates between frequency, intensity, and duration signals. Thereafter, the signals undergo integration in Heschl's gyrus and/or in the Planum Temporale (Yost 2007:236, 246; Koffi 2021a:56-61). The resolution phase can be equated with what takes place in the thalamus. According to Amerman (2016:434), it discards 99% of the excess information that it does not need. It is quite likely however, nobody knows for sure, that the thalamus computes the melodicity scores of utterances and correlates them with a melodicity scale template such as the one in Table 3. If such a template exists at all, it resides in hearers' auditory memory and helps them interpret the overall melodicity of talkers' utterance. When these insights are applied to Utterances 1 and 2, their melodicity scores of 72.78% and 72.84% correspond to a modal intonation.[8]

N0	Percentages	Melodicity Scale
1.	76-100%	Falsetto
2.	51-75%	Modal/Recitative
3.	26-50%	Monotonous
4.	0-25%	Staccato

Table 3: Melodicity Scale

A short description of the steps on the scale is given, starting from the bottom up. A **staccato intonation** is a fast speech delivered in small chunks. It has been referred to as choppy or singsongy intonation. Reed and Michaud (2015:460) contend that L2-accented English speakers of specific language backgrounds intone their utterances this way. A **monotonous intonation** is one in which there is barely any inflection in the voice because pitch and sonority plateaus abound. This intonation is paralinguistically associated with boredom or disinterestedness. Magdics (1963:146) uses the term **"recitative"** to describe the unmarked intonation pattern that people use when they are not monitoring their diction. It is also known as **modal intonation**. This is an unmarked intonation of every verbal interaction. A **falsetto intonation** is one in which pitch, sonority, and rhythmicity are modulated extravagantly. It has various nicknames. It has been dubbed "service counter" intonation or "nurses' talk." Some extreme versions of it correspond to "foreigner talk," that is, the intonation used when native speakers perceive their L2 interlocutors as lacking proficiency in L1. Another extreme

[8] The melodicity score of sentences that are less than 5 words tends to be high, while that of sentences with 10 words or more tends to be low. This has a lot to do with the distribution of acoustic energy in speech production and speech rate. With shorter utterances, less articulatory energy is spent, whereas with longer utterances, more energy is spent producing the utterance. As a result, more pitch and sonority plateaus are to be expected. Also, in longer utterances, speakers tend to "rush" through words so that they can get to the end of the utterance. Consequently, many words do not reach their articulatory targets and become shorter. This is referred to as "hypospeech," by Thomas (2011:293). Its opposite is "hyperspeech," in which speakers take time to enunciate words clearly and distinctly.

case is "motherese," the exaggerated intonation that parents or caregivers use in addressing preverbal infants (Fernald 1989).

9.0 Other Factors Influencing Intonation

The preceding analyses have shown how the speech intelligibility framework within psychoacoustics can help tame the wild side of intonation. Yet, there are more to intonation than pitch, intensity, and duration. For a comprehensive study, ancillary issues such the biological gender of the speaker[9] and the acceptable number of participants in a study must be investigated. Having an appropriate number of participants is necessary because it allows for interspeaker variability analyses. Then and only then can the results be generalized to a speech community. Additional issues tackled in this section are expressiveness, intelligibility, and pragmatics.

9.1 Biological Gender and Intonation

The consensus in acoustic phonetics is that biological gender is an indexical feature. Just by hearing a person's voice, most people can tell whether the talker is an adult male or female. Table 4 summarizes various measurements found in Stevens (2000:9, 13, 24, 25) that pertain to the biological gender that affect intonation.

Articulatory Characteristics	**Adult female**	**Adult male**
Vocal tract length	14.1 cm	16.9 cm
Pharynx length	6.3 cm	8.9 cm
Oral cavity length	7.8 cm	8.1 cm
Vocal fold length	1.0 cm	1.5 cm
Length of the trachea	2 to 4 cm	2 to 4 cm
Vocal tract volume (closed mouth)	130 cm^3	170 cm^3
Vocal tract volume (mouth open 1cm wide)	150 cm^3	190 cm^3
Volume of the tongue	90 cm^3	110 cm^3
Lip opening (horizontal: from corner to corner)	10 to 45 mm	10 to 45 mm
Lip opening (vertical: from upper to lower lip)	5 to 20 mm	5 to 20 mm

Table 4: Biological Gender Issues in Intonation

Because of these physiological factors, the pitch of females' voices is 1.5 to 1.7 times higher than males' (Kent and Read 2002:191, Klatt 1987:761). The intensity in females' voice tends to be 3 dB lower than males' (Yost 2007:147). However, this was true a generation or so ago, but not so much now. Large data discussed in Koffi (2020:14-15) does not support gender-based intensity differences anymore for most speakers of American English. The reason for the blurring of the intensity line is **vocal fry**. Younger American females who are millennials or younger have adopted this new way of speaking.[10] As a result, the intensity in their voice, especially at prepausal positions is lower now. The duration correlate is also not biologically pertinent in spite of Hillenbrand et al.'s (1995:3103) report that vowels last longer in the speech of American females' than in males'. In fact, Koffi and Krause (2020:66, 81) found no durational differences for vowels in running speech. It follows from this brief review that intonation measurements should be separated by biological gender because of F0/pitch, not so much because of intensity and duration.

[9] I make a distinction between three types of genders: biological gender, grammatical gender, and social gender.
[10] Even though vocal fry is found occasionally among some males, it is by far more prevalent in female speech.

9.2 Optimal Number of Participants

Prosody studies do not usually enroll many participants. Yet, the conclusions drawn from them are often stated so forcefully that one would think that they apply to an entire speech community. The questions that students of acoustic phonetics often ask is the following. "How many participants are needed for the results of a study to be generalizable?" There is no consensus on the optimal numbers of participants, but acoustic phonetic studies involving three participants or less are discouraged. Ladefoged (1968:xi) contends that such studies merely describe speakers' idiosyncrasies and are, therefore, not generalizable. Since Atal (1972), it has been deemed that the results of a study enrolling 10 participants or more are generalizable to an entire speech community. Himmelmann and Ladd (2008:265) contend that for lesser-known languages where it is hard to find participants, eight participants are enough. It goes without saying that in selecting participants, gender parity should be encouraged whenever possible, for reasons indicated in 9.1. To avoid comparing apples and oranges, all the participants in a study must be given the same utterances to produce. If they produce different sentences, the results of the study are simply not comparable. This is one of the reasons why "authentic" speech is not good enough for the study of intonation. The likelihood of two people producing the same exact utterances is very slim. On the other hand, read speech is more amenable to interspeaker variability analysis because all the participants can be given the same elicitation paragraph or the same sentences to read.

9.3 Cross-Linguistic Comparisons

The paucity of acoustic phonetic data makes cross-linguistic comparisons of intonation almost impossible. Even so, it has been surmised, based on impressionistic data, that there is something universal about intonation. Bolinger (1978:511) expresses this view as follows,

> Probably the majority of intonational differences from language to language are instances of doing *the same things* but doing them in *different degrees* – either more often or less often, or to a greater or lesser extreme. If this is the case, then we can think of *an intonation core*, an innate pattern of the sort envisioned by Liberman, from *which speakers and cultures may depart, but to which some force is always pushing back* [italics added for emphasis].

The lack of a coherent theory of intonation has been a major stumbling block for cross-linguistic comparisons. Consequently, statements about intonational universals are couched in generic terms. The psychoacoustic approach proposed here has the potential of resolving this problem because it relies on JNDs that are valid for all human beings, so long as they do not suffer from any hearing abnormality or speech impediment. With such a framework, future researchers will be in a better position to assess claims of intonation universals or lack thereof.

9.4 Intonation, Accentedness, and Intelligibility

The role that intonation plays has been sufficiently well-research in speech synthesis. These insights can be applied to accented speech to assess any putative correlation between it, intelligibility, and intonation. In the absence of acoustic phonetic data, unverified claims have been made that prosodic issues interfere more with the intelligibility of L2-accented English than segmental errors. We know that during the discrimination phase of auditory processing, CANS pays attention to the minutest of acoustic cues. Speech parameters are used for all kinds of indexical discrimination tasks. For example, Purnell et al. (1999) indicate that the <e> in <hello> is enough to discriminate between speakers of some ethnic groups in American English. They write on page 21 that "The acoustic signal carries a variety of information about the individual speaker beyond just the phonemic content of the signal." De Jong (2018:9) also writes that the minutest of acoustic cues is enough to tell whether the speaker is a native or a nonnative speaker of English:

> Research on the acoustics of the speech of nonnative speakers shows how extraordinarily sensitive listeners are to the very subtle details of speech patterns. When we hear a person speaking, we very rapidly not only hear what the speaker is saying but also are cognizant of many other attributes of the speaker: whether male or female, emotional state, whether the speaker has a cold, and whether the speaker is a native member of our own speech community… Speech signals are very rapid but are also defined by a complex and intercollated matrix of information.

Yet, these statements do not speak directly to the role of intonation in intelligibility. There are at least two good reasons for believing that the intonation patterns of L2 English do not affect intelligibility. The first is based on the view that intonation patterns are universals, differing only by degrees, as noted in the quote by Bolinger in 9.3. Secondly, Amerman (2016:383) writes that CANS discards 99% of the acoustic cues that it does not need. So, unless the hearer focuses on indexical cues at the expense of the linguistic information being communicated, the intonation patterns themselves are less likely to affect intelligibility. The claim made by Wright and Souza (2012:183) that accented speech imposes a processing cost and burdens the cognitive load of hearers can be explained by the fact that some hearers cause their thalamus to attend to paralinguistic details instead of letting it abstract linguistically significant information. Even so, this is not prima facie evidence that accentedness infringes on intonation patterns. The real impact of accentedness on intonation on the one hand, and of intonation and intelligibility on the other, is waiting to be adjugated on based acoustic phonetic data, not on impressionistic conjecturing.

9.5 Intonation and Expressiveness

Bunnell (2022:14) writes that current efforts in speech synthesis have brought to light the need to pay attention to expressiveness in verbal interactions. Simulations are run in order to understand how "**neutral**" speech resembles or differs from "**emotionally charged**" speech. Most of the prosodic patterns encountered in daily verbal interactions qualify as neutral. Yet, we also know that emotionally charged verbal exchanges take place sometimes and have their own unique intonation patterns. Murray and Arnott (1993:1103-1105) list the following, as examples of emotionally charge speech.

1. Anger/aggression
2. Joy/happiness/humor
3. Disgust/hatred/contempt/scorn
4. Sarcasm/irony
5. Grief/sorrow
6. Affection/tenderness
7. Surprise/astonishment

Because it is practically impossible to record emotionally charged utterances live without violating people privacy rights, Murray (1993) has asked participants to simulate them in his lab. Even though simulated emotions are not the real "thing," insightful data have been collected. In their literature review, Gobl and Chasaide (2003:190) indicate that, because emotional outbursts involve greater vocal efforts, frequency and intensity rank high. In their study of screams, Hansen et al. (2017:2961, Table 2) did not measure intensity. But they found that pitch played an important role. In neutral speech, the average F0 is 135 Hz for males and

181 Hz for females. In screams, F0 rises to 233 Hz and 281 Hz for males and females respectively. The differences between neutral speech and screams are 95 Hz for males and 100 Hz for females.

It has also been noted that pitch modulation is an important paralinguistic feature that differentiates between various speech acts. Furthermore, it has been posited that within an utterance, the lexical item that is the most "expressive" gets boosted and is given a higher pitch. To demonstrate this, we return to Figure 8, displayed here as Figure 14, to see which word is the most expressive in the two utterances.

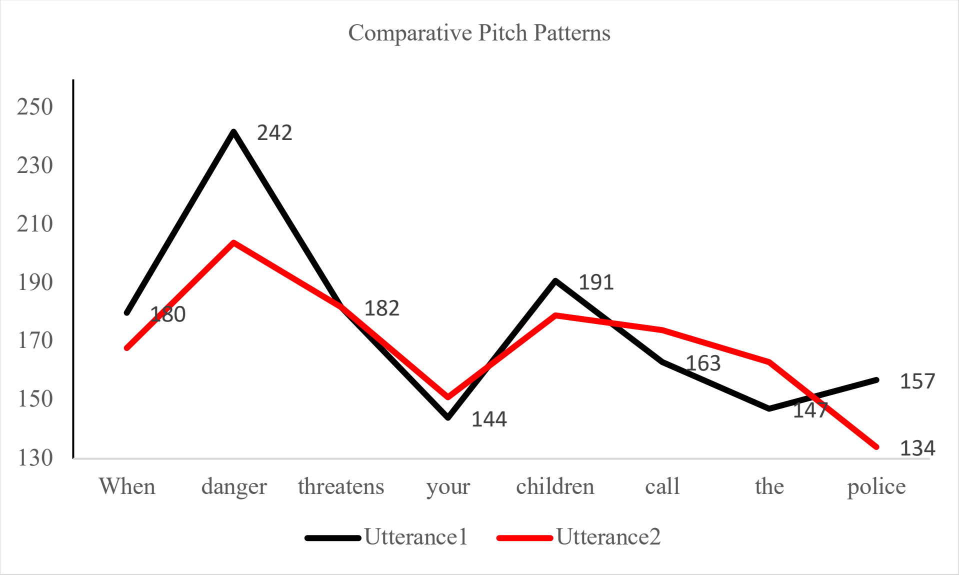


Figure 14: Prosodic Pitch Prominence

In both Utterances 1 and 2, the word <danger> receives **prosodic prominence**. It sits on top of the cone-like pattern. Words that receive prosodic prominence are at the **apex** of the cone. In Utterance 1, pitch rises from <when> (180 Hz) peaks at <danger> <242 Hz> and drops from there to <threatens> (182 Hz). In Utterance 2, the same pattern repeats itself. Pitch rises from <when> (168 Hz) and peaks at <danger> (204 Hz) and drops from there to <threatens> (182 Hz). The fact that <danger> is the peak in both utterances underscores the fact that it is the most expressive word in both utterances.

9.6 Intonation and Information Structure

Syntactic and discursive strategies can be used by speakers to signal which lexical item in their utterance should receive prosodic prominence. One strategy that is often used is **contrastive stress** (Bolinger 1978:488). Talkers can use this strategy to emphasize every single word in an utterance, even a function word, and make it stand out prosodically. This is demonstrated by the boldfaced words in the sentences below:

1. *<**When** danger threatens your children, call the police>*
2. *<When **danger** threatens your children, call the police>*
3. *<When danger **threatens** your children, call the police>*
4. *<When danger threatens **your** children, call the police>*
5. *<When danger threatens your **children**, call the police>*
6. *<When danger threatens your children, **call** the police>*
7. *<When danger threatens your children, call **the** police>*
8. *<When danger threatens your children, call the **police**>*

Alternatively, talkers can make use of **information structure** to convey prosodic prominence. In this case, they simply move constituents out of their default syntactic position to elsewhere in the utterance. In *The Minimalist Program*, Chomsky (1995:228, 250, 261) proposed a "**Move α**" rule according to which a constituent in the deep structure is moved elsewhere in an utterance. When this happens, that constituent receives the "communicative function of focus" (Jackendoff 2002:411-412). To illustrate this point, let's assume that Utterances 1 and 2 have the following basic structure:

Base form: <*You call the police when danger threatens your children*>

The base forms of utterances are those that can be generated directly by Phrase Structure Rules. To obtain the surface structure form, *<When danger threatens your children, call the police>,* the "Move α" operation moved the subordinate clause out from the end of the utterance to the beginning. By doing so, the whole subordinate clause receives prominence. Within the clause, the word <danger> is selected consciously or subconsciously by the talker to receive prosodic prominence. So, syntactic operations can be used to encode prominence. Talkers have paralinguistic tools beyond syntax that they can use. Yet, Klatt (1987:762) notes that formulating rules to account for intonation has remained a challenge for speech synthesis:

> For synthesis by rule, what is needed is a theory that can predict when F0 will rise or fall, and what levels it will reach on individual stressed syllables of a sentence as a function of syntactic structure, stress patterns, and semantic/performance variables (if known) such as the location of the most important word in the sentence, or the speaker's attitude towards what is being said.

It is entirely possible that, because of "performance variables," there may be interspeaker variability among native speakers when they produce the same utterance. However, data collected on Utterances 1 and 2 from 10 native speakers of American English shows that they all encoded prosodic prominence the same way, by foregrounding <danger>.[11]

9.7 Intonation and Pragmatics

Pragmatics can be defined broadly as all the paralinguistic factors that influence speech production and perception. These factors include but are not limited to talkers' mental and emotional dispositions, their assumptions about hearers, and the effects that they expect their utterances to have on listeners. Accordingly, talkers summon the relevant syntactic, semantic, and rhetorical strategies to achieve their communicative goals. Seen from this perspective, pragmatics is an indomitable topic to cover. When prosody is added to the mix, pragmatics become even more intractable. Therefore, it is understandable that Shattuck-Hufnagel and Turk (1996:236) omitted it from their tutorial on prosody. Yet, I contend that the psychoacoustic model proposed here is robust enough to embrace pragmatics so long as the paralinguistic factors that fuel speech acts have traceable manifestations in the speech signals that talkers emit. If so, then F0, duration, and intensity measurements and JNDs can be used to uncover some of the intonational intent embedded in the message. Pragmatics in intonation is likely to remain an unresolvable issue for a long time because talkers choose to send mixed signals with a covert intent to violate the maxims of discourse (Fromkin et al. 2017:165).

[11] The data has been collected but the paper containing the results is yet to be written.

10.0 Summary

The psychoacoustic methodology outlined in this paper provides a sure foundation for the study of prosody. The JND thresholds used to account for intonation are based on nearly 100 years of auditory speech perception experiments. They are reliable because they have been confirmed again and again in a wide variety of experimental settings. National and international regulatory agencies have mandated that audio engineering products conform to their specifications. This means that when these JNDs are used to account for intonation patterns, the results will match as closely as possible how the naked ear perceives an utterance. Additionally, this psychoacoustic-based framework mirrors the three phases of auditory perception. The acoustic tokenization phase of the analysis corresponds to the Fourier-like analyses that the naked ear performs on pitch and sonority movements and on rhythmic patterns. The RFL rankings mirror the integration phase in CANS, while the pegging of the melodicity score on a four-point Likert melodicity scale corresponds to the resolution phase. This holistic approach can help tame the wild side of intonation that has remained untamed for a very long time. In a nutshell, the model discussed in this paper gives researchers new tools to probe deeper into intonational issues such as interspeaker variability, cross-linguistic comparisons, intelligibility, expressiveness, and pragmatics.

ABOUT THE AUTHOR

Ettien Koffi, Ph.D. linguistics (Indiana University, Bloomington, IN) teaches at Saint Cloud State University, MN. He is the author of five books and author/co-author of several dozen articles on acoustic phonetics, phonology, language planning and policy, emergent orthographies, syntax, and translation. His acoustic phonetic research is synergetic, encompassing L2 acoustic phonetics of English (Speech Intelligibility from the perspective of the Critical Band Theory), sociophonetics of Central Minnesota English, general acoustic phonetics of Anyi (a West African language), acoustic phonetic feature extraction for application in Automatic Speech Recognition (ASR), Text-to-Speech (TTS), voice biometrics for speaker verification, and infant cry bioacoustics. Since 2012, his high impact acoustic phonetic publications have been downloaded **54,717** times (**37,140** as per Digital Commons analytics), **17,577** (as per Researchgate.net analytics), and several thousand downloads from Academia.edu, as of **February 2023**. He can be reached at enkoffi@stcloudstate.edu.

Appendix 1: Removing Octave Errors

Praat offers many possibilities for removing octave errors. I have found this one to be the easiest. Here are the steps:

1. Load Praat
2. Load the file containing the error into Praat Object
3. Select that file
4. Find **<Analyze Periodicity>** on the Praat Object Menu, and click on it
5. When the dialog opens, select **<Pitch..>** and click **<Ok>**. A new file is created in Praat Object Menu.
6. Find it, select it and click on **<View and Edit>**. You will see a document that looks like this:

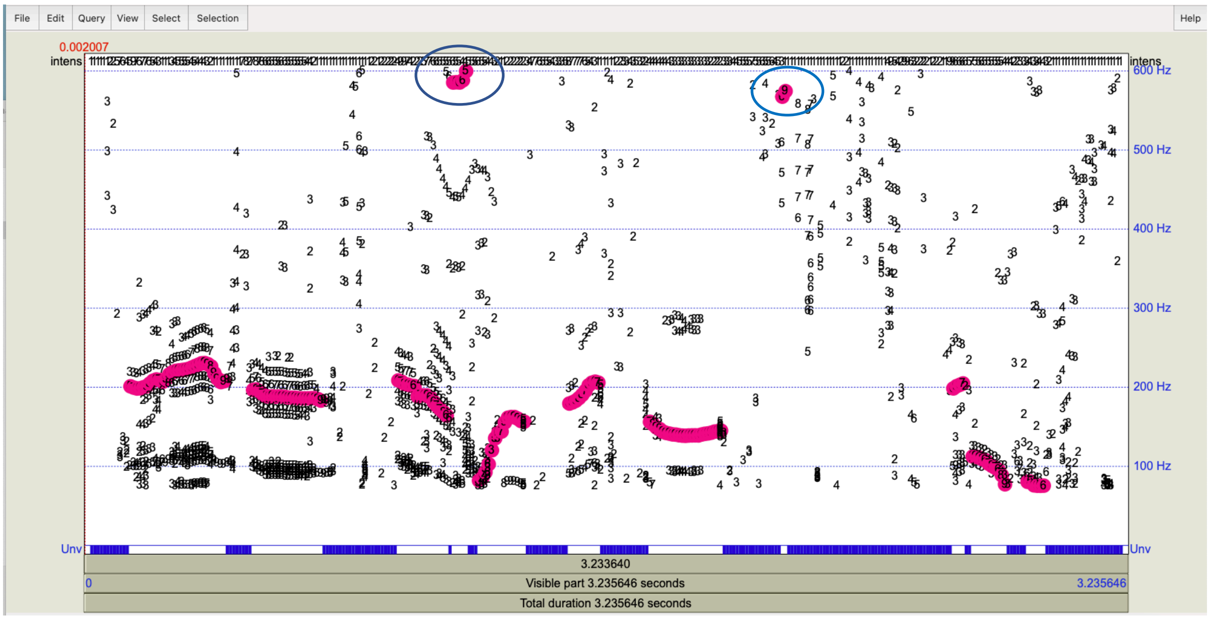


Figure 15: Octave Errors on Top Inside Blue Circles [online]

7. Use the cursor to select the octave errors.
8. Go to tabs on top and find <Selection>. Click on it.
9. Click on <Unvoice>. This will delete the octave errors.
10. Name and save the edited file.

References

Abe, Isamu. 1998. Intonation in Japanese. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo. pp. 363-378. New York: Cambridge University Press.

Amerman, Erin C. 2016. *Human Anatomy & Physiology*. New York: Pearson Education, Inc.

Astruc, Lluisa. 2015. Prosody. *The Bloomsbury Companion to Phonetics.* ed. by Mark J. Jones and Rachael-Anne Knight, pp. 126-139. New York: Bloomsbury.

Atal, B. S. 1972. Automatic Speaker Recognition Based on Pitch Contours. *Journal of the Acoustical Society of America*, 52 (6):1687-1697.

Bachorowski, Jo-Anne and Michael J. Owren. 1999. Acoustic Correlates of Talker Sex and Individual Identity are Present in a Short Vowel Segment Produced in Running Speech. *The Journal of the Acoustical Society of America* 106 (2): 1054-1063.

Baken, R. J. and Robert F. Orlikoff. 2000. *Clinical Measurement of Speech and Voice*. 2nd Edition. San Diego, CA: Singular Publishing Group.

Békésy, Georg Von. 1947. Variations of Phase along the Basilar Membrane with Sinusoidal Vibrations. *The Journal of the Acoustical Society of America* 19 (3): 452-460.

Benkirane, Thami. 1998. Intonation in Western Arabic (Morocco). *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo, pp. 348-362. New York: Cambridge University Press.

Bolinger, Dwight. 1998. Intonation in American English. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo. pp. 45-55. New York: Cambridge University Press.

Bolinger, Dwight. 1978. Intonation Across Languages. *Universals of Human Language*, Volume 2, Phonology, ed by Joseph H. Greenberg, pp. 471-524. Stanford, CA: Stanford University Press.

Botinis, Antonis. 1998. Intonation in Greek. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo, pp. 291-313. New York: Cambridge University Press.

Bunnell, Timothy H. 2022. Speech Synthesis: Towards a "Voice" for All. *Acoustics Today* 18 (1): 14-22.

Childers, D.G. and Ke Wu. 1991. Gender Recognition from Speech. Part II: Fine Analysis. *The Journal of the Acoustical Society of America* 90 (4): 1841-1856.

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: The MIT Press.

Cruz-Ferreira, Madalena. 1998. Intonation in European Portuguese. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo.pp. 219-241. New York: Cambridge University Press.

Curzan, Anne and Michael Adams. 2009. How English Works: *A Linguistic Introduction. Second Edition*. New York, NY: Pearson-Longman.

de Jong, Kenneth J. 2018. Sensitivity to Foreign Accent. *Acoustics Today* 14 (2): 9-16.

Dickerson, Wayne B. Using Orthography to Teach Pronunciation. 2015. *The Handbook of English Pronunciation*, ed. by Marnie Reed and John M. Levis, 488-504. Malden, MA: Wiley Blackwell.

Eddins, David A. and David M. Green. 1995. Temporal Integration and Resolution. *Hearing: A Handbook of Perception and Cognition*. Second Edition, ed. by Brian C.J. Moore pp. 207-242. New York, NY: Academic Press.

Fastl, Hugo and Eberhard Zwicker. 2007. Pschoacoustics: Facts and Models. Third Edition Edition. New York, NY: Springer.

Fernald, Anne. 1989. Intonation and Communicative Intent in Mothers' Speech to Infants: Is the Melody the Message? *Child Development* 60 (6): 1497-1510.

Fletcher, Harvey. 1940. Auditory Patterns. *Reviews of Modern Physics*, Volume 12, pp. 47-65.

Fonagy, Ivan. 1998. Intonation in Hungarian. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo, pp. 331-347. New York: Cambridge University Press.

Fromkin, Victoria, Robert Rodman, and Nina Hyams. 2017. *An Introduction to Language*. Eleventh Edition. Boston, MA: Cengage Learning.

Fry, Dennis. B. 1958. Experiments in the Perception of Stress. *Language and Speech* 1 (2): 126-152.

Fry, Dennis. B. 1979. *The Physics of Speech.* New York: Cambridge University Press.

Garding, Eva. 1998. Intonation in Swedish. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo. pp. 112-130. New York: Cambridge University Press.

Gobl, Christer and Ailbhe Ni Chasaide. 2003. The Role of Voice Quality in Communicating Emotion, Mood and Attitude. *Speech Communication* 40:189-212.

Gronnum, Nina. 1998. Intonation in Danish. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo. pp. 131-151. New York: Cambridge University Press.

Hansen, John H., Mahesh Kurman Nandwana, and Navid Shokouhi. 2017. Analysis of Human Scream and Its Impact on Text-Independent Speaker Verification. *Journal of the Acoustical Society of America*, Volume 93 (2):1097-1108). *Journal of the Acoustical Society of America* 141 (4): 2957-2966.

Hanson, Helen M. 1997. Glottal Characteristics of Female Speakers: Acoustic Correlates. *The Journal of the Acoustical Society of America* 101 (1): 466-481.

Hartmann, William M. 1995. The Physical Description of Signals. *Hearing: Handbook of Perception and Cognition*. Second edition, ed. by Brian C. J. Moore, pp. 1-40. New York: Academic Press.

Hillenbrand, James, Laura A. Getty, Michael J. Clark, and Kimberlea Wheeler. 1995. Acoustic Characteristics of American English Vowels. *The Journal of the Acoustical Society of America* 97 (5): 3099-3111.

Himmelmann, Nikolaus P. and Robert D. Ladd. 2008. Prosodic Description: An Introduction for Fieldworkers. *Language Documentation and Conservation* 2: 244-274.

Hirst, Daniel. 1998. A Survey of Intonation Systems. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo. pp. 1-44. New York: Cambridge University Press.

Hollien, Harry. 1990. The Acoustics of Crime: The New Science of Forensic Phonetics. New York: Plenum Press.

Houtsma, Adrianus J.M. Pitch Perception. 1995. Pitch Perception. *Hearing: A Handbook of Perception and Cognition*. Second Edition, ed. by Brian C.J. Moore pp. 267-295. New York, NY: Academic Press.

Jackendoff, Ray. 2002. *Foundations of Language: Brain, Meaning, Grammar, Evolution*. New York, NY: Oxford University Press.

Jongman, Allard, Ratree Wayland and Serena Wong. 2000. Acoustic characteristics of English fricatives. *Journal of the Acoustical Society of America* 108 (3): 1252–1263.

Jongman, Allard, Zhen Qin, Jie Zhang, and Joan A. Sereno. 2017. Just Noticeable Differences for Pitch Direction, Height, and Slope for Mandarin and English Listeners. *Journal of the Acoustical Society of America* 142 (2): 163-169.

Kent, Ray D. and Charles Read. 2002. *Acoustic Analysis of Speech*, 2nd edition. Clifton Park, NY: Delmar-Cengage Learning.

Klatt, Dennis H. 1976. Linguistic uses of segmental duration in English: Acoustic and Perceptual Evidence. *Journal of the Acoustical Society of America* 59 (5): 1206–1221.

Klatt, Dennis H. 1987. Review of Text-to-Speech Synthesis Conversion for English. *Journal of the Acoustical Society of America*, Volume 82 (3):737-793.

Koffi, Ettien. 2021a. *Relevant Acoustic Phonetic Analysis of L2 English: Focus on Intelligibility*. Boca Raton, FL: CRC Press, Taylor & Francis Group, A Science Publishers Book.

Koffi, Ettien. 2021b. A Comprehensive Review of Intensity and its Linguistic Applications. *Linguistic Portfolios* 9:2-27.

Koffi, Ettien. 2020. A Comprehensive Review of Intensity and its Linguistic Applications. *Linguistic Portfolios* 9:2-28.

Koffi, Ettien and Jessica Krause. 2020. Speech Style Variation of Vowels in Citation Form vs. Running Speech: Intelligibility Implications for AI-Enabled Devices. *Linguistic Portfolios* 9:60-85.

Koffi, Ettien. 2019a. A Comprehensive Review of F0 and its Various Correlations. *Linguistic Portfolios* 8:2-24.

Koffi, Ettien. 2019b. A Template Model Account of Lexical Stress in Arabic-Accented English, ed by J. Levis, C. Nagle, & E. Todey, *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, ISSN 2380-9566, Ames, IA, September 28, pp. 158-167. Ames, IA: Iowa State University.

Koffi, Ettien. 2017. The New Paradigm in Tone Analysis: The Contribution of the Band Theory. *Linguistic Portfolios* 6:147-165.

Koffi, Ettien. 2015. *Applied English Syntax: Foundations for Word, Phrase, and Sentence Analysis*, 2nd edition. Dubuque, IA: KendallHunt Publishing Company.

Kratochvil, Paul. 1998. Intonation in Beijing Chinese. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo. pp. 421-436. New York: Cambridge University Press.

Ladd, Robert D. 2008. *Intonational Phonology*, 2nd edition. New York: Cambridge University Press.

Ladefoged, Peter. 1968. *A Phonetic Study of West African Languages: An Auditory-Instrumental Survey.* Cambridge, UK: Cambridge University Press.

Ladefoged, Peter. 2001. *A Course in Phonetics*. Fourth Edition. New York: Harcourt College Publishing.

Ladefoged, Peter. 2003. *Phonetic Data Analysis: An Introduction to Fieldwork and Instrumental Techniques.* Malden, MA: Blackwell Publishing.

Ladefoged, Peter and Sandra Ferrari Disner. 2012. *Vowels and Consonants.* Third Edition. Malden, MA: Wiley-Blackwell.

Liu, Chang. 2013. Just Noticeable Difference of Tone Pitch Contour Change for English and Chinese Native Listeners. *Journal of the Acoustical Society of America* 134 (4): 3011-3020.

Magdics, Klara. 1963. Research on Intonation during the Past Ten Years. *Acta Linguistica Academiea Scientiarum Hungaricae* 13 (1/2): 133-165.

Murray, Iain R. and John L. Arnott. 1993. Toward the Simulation of Emotion in Synthetic Speech: A Review of the Literature on Human Vocal Emotion. *Journal of the Acoustical Society of America*, Volume 93 (2):1097-1108).

O'Grady, William, John Archibald, Mark Aronoff, and Janie Rees-Miller. 2017. *Contemporary Linguistics: An Introduction*, 7th edition. New York: Bedford/St. Martins.

Palmer, Alan R. 1995. Neural Signal Processing. *Hearing: Handbook on Perception and Cognition*, ed. by Brian C. J. Moore, pp. 75-121. New York: Academic Press

Pandey, Pramod. 2015. Indian English Pronunciation. *The Handbook of English Pronunciation*, ed. by Marnie Reed and John M. Levis, pp. 301-319. Malden, MA: Wiley Blackwell.

Peterson, Gordon E, and Harold L. Barney. 1952. Control Methods in a Study of the Vowels. *The Journal of the Acoustical Society of America* 24 (2): 176-84.

Purnell, Thomas, William Idsardi, and John Baugh. 1999. Perceptual and Phonetic Experiments on American Dialect Identification. *Journal of Language and Social Psychology* 18 (1): 10-30.

Quéne, Hugo. 2004. What is the Just Noticeable Difference of Tempo in Speech? *On Speech and Language. Studies for Sieb G. Nooteboom*, ed. by Hugo Quené and Vincent van Heuven 149-158. Utrecht, The Netherlands: Netherlands Graduate School of Linguistics (LOT).

Reed, Marnie and Christina Michaud. 2015. Intonation in Research and Practice: The Importance of Metacognition. The Handbook of English *Pronunciation*, ed. by Marnie Reed and John M. Levis, 454-487. Malden, MA: Wiley Blackwell.

Rossi, Mario. 1998. Intonation in Italian. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo, pp. 219-241. New York: Cambridge University Press.

Sachs, M.B., I.C. Bruce, R.L. Miller and E.D. Young. 2002. Biological Basis of Hearing Aid Design. *Annals of Biomedical Engineering* 30: 157–168.

Schnitta Bonnie. 2016. Residential Quietude, the Top Luxury Requirement. *Acoustics Today* 12 (3):49-56.

Shattuck-Hufnagel, Stefanie and Alice E. Turk. 1996. A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research* 25 (2): 193-247.

Searle, John R. 1969. *Speech Acts: An Essay in the Philosophy of Language*. New York: Cambridge University Press.

Stevens, Kenneth. 2000. *Acoustic Phonetics*. Cambridge, MA:MIT Press.

Stevens, Kenneth N. 1998. Models of Speech Production. *Handbook of Acoustics*, ed. by by Malcolm J. Crocker, pp. 1231-1244. New York: A Wiley-Interscience Publication, John Wiley and Sons, Inc.

Svetozarova, Natalia. 1998. Intonation in Russian. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo, 264-277. New York: Cambridge University Press.

Thomas, Erik R. 2011. *Sociophonetics: An Introduction*. New York: Palgrave Macmillan.

t' Hart, Johan. 1998. Intonation in Dutch. *Intonation Systems: A Survey of 20 Languages*, ed. by Daniel Hirst and Albert di Cristo. pp. 96-111. New York: Cambridge University Press.

t' Hart, Johan. 1981. Differential Sensitivity to Pitch Distance, Particularly in Speech. *Journal of the Acoustical Society of America*, 69 (3): 811-821.

Wichmann, Anne. 2015. Functions of Intonation in Discourse. *The Handbook of English Pronunciation*, ed. Marnie Reed and John M. Levis, pp.175-189. Malden, MA: Wiley Blackwell.

Wright, Richard and Pamela Souza. 2012. Comparing Identification of Standardized and Regionally Valid Vowels. *Journal of Speech, Language, and Hearing* 55:182-193.

Yost, William A. 2007. *Fundamentals of Hearing: An Introduction*. New York: Elsevier.

Yost, William A. 2015. Psychoacoustics: A Brief Historical Overview. *Acoustics Today* 11 (3):46-53.

Zhang Cuiling, Geoffrey Stewart M, Ewald Enzinger, Felipe Ochoa. 2013. Effects of Telephone Transmission on the Performance on Formant Trajectory-Based Forensic Voice Comparison – Female Voices. *Speech Communication* 55:796-813.