TEL AVIV UNIVERSITY
The Iby and Aladar Fleischman Faculty of Engineering
The Zandman-Slaner School of Graduate Studies

PROSODIC FEATURES CRITERION

A thesis submitted toward the degree of
Master of Science in Electrical and Electronic Engineering

by

Ben Fishman

February 2020

TEL AVIV UNIVERSITY
The Iby and Aladar Fleischman Faculty of Engineering
The Zandman-Slaner School of Graduate Studies

PROSODIC FEATURES CRITERION

A thesis submitted toward the degree of
Master of Science in Electrical and Electronic Engineering

by

Ben Fishman

This research was carried out at Tel Aviv University
in the School of Electrical Engineering
Faculty of Engineering
under the supervision of Prof. Hagit Messer-Yaron
and Dr. Irit Opher

February 2020

*In memory of my grandfather **Reuven Berger**,*
who taught me how to think like an engineer

Acknowledgments

First and foremost, I would like to express my gratitude to my advisors:
To **Prof. Hagit Messer-Yaron** for her guidance and for her constant support throughout this work.
To **Dr. Irit Opher** who supported, focused, and pushed me forward during the work on this thesis, sometimes during strange hours of the day. Irit supervised my work with great patience, insisting on the finest nuance, but at the same time let me be the lead of this project and feel it is mine. For this and for your great dedication and engagement, I would like to thank you!

Next, I would like to thank **Dr. Yitzhak Lapidot** for his collaboration, for providing fruitful and productive feedback, and for never giving up on any mathematical formulation or notation.

I would also like to thank the members of the Afeka Center for Language Processing (ACLP): **Ruth Aloni Lavi**, **Ella Erlich**, **Yermiyahu Hauptman**, **Noga Hellman**, and **Noam Lotner** with whom I consulted and who provided me with great advice whenever possible.

I would like to thank all 36 speakers who donated their voice to our self–collected Hebrew dataset which served us for our basis experiment.

Finally, I would like to thank my family, which without them any of this would not be possible:
To **Sivan**, who was my partner in the beginning of this project, and who during its course became my wife. For your love, support and patience during the long weeks and weekends in which I was absorbed in this research; For taking an active part of this project, starting

which being the first person who donated her voice, by assisting in labeling the data, by correcting and polishing my English time and again, and even for having some good engineering insights (even though you claim you are not a mathematical person).
To the **Fishman** family, my parents – **Yossi** & **Nurit**, my siblings – **Shirel** & **Noa** and the entire extended family, for believing in me, and for your unconditional love, support, and encouragement during this long period of time. I love you all!

1 Abstract

Prosody is the non-lexical information conveyed in a speech signal, such as the intonation of the sentence, the loudness, rhythm and tempo, timbre and more.
Prosody provides valuable information and plays an important role in everyday communication. It can reveal the intention and attitude of the speaker and can even be used to assess their emotional state and some medical conditions.

Prosody is a multidisciplinary field that has been researched extensively for many years in a few different domains, both engineering and non-engineering.
Nevertheless, little attention has been given to formulating the speech features that represent prosodic information. This is a significant gap that serves as the basis of this thesis.

In this work, we aim at defining what a prosodic feature is in the sense of quantifying the prosodic information a feature carries. We introduce the Prosodic Feature Criterion (PFC), a criterion for evaluating the prosodic nature of a speech feature. We also show a methodology for calculating the PFC.

We apply the PFC to two feature sets: (1) a collection of standard speech features, (2) a subset of the well known OpenSMILE toolkit, which consists of thousands of features. All experiments are carried out using two datasets: (1) a Hebrew dataset, especially designed for researching prosodic features, that contains two prosodies: neutral and question. (2) An English dataset by the Linguistic Data Consortium (LDC) that contains 15 emotional states.

We use several methods to validate the PFC: comparing the PFC results to common knowledge in the field, and to classification scores based on widely-used methods. We also show visualizations of the PFC scores using dimension reduction of multiple features representations, where we saw good separation between prosodic classes that received high PFC scores.

All our validation tests show positive results, suggesting that the PFC can be used to measure the prosodic quality of the feature, in a quantitative and objective way.

Table of Contents

List of Figures

List of Tables

2 Introduction

In this chapter, we explain the term prosody and introduce our motivation, research question, and how our approach contributes to prosody representation and research.

2.1 What is Prosody?

Human speech can be divided into two types of information: (1) "what was said," which refers to the lexical content, i.e., the words, and (2) "how it was said", which refers to additional, non-lexical information which can be perceived by the listener. The term for this kind of information is called **prosody**.

Prosody can be defined as the study that relates to the non-contextual information conveyed in speech. An alternative and common definition is: "the suprasegmental[1] aspects of the speech stream, that do not represent the verbal content" [107, 3].

Examples of prosodic manifestations are the intonation of the sentence, the loudness, rhythm and tempo, timbre, and more. Prosody provides valuable information and plays an important role in everyday communication between people [153] by helping them maintain dialogue structure [56], separate the speech into chunks of information, and parse discourses into meaningful syntactic and semantic units [167].
Prosody can sometimes reveal the subtle meaning, intention or attitude of the speaker [62], decipher speech-acts [34] and may even sometimes be used to asses a speaker's emotional state. it can assist in identifying characteristics of the speaker such as gender, age [160, 166, 163] and sometimes it can even indicate some medical condition of the speaker. Elaboration on the role of prosody in various research fields and the terms mentioned above is found in chapter 3.

[1] Suprasegmental - the units of speech which are larger than a single phonetic segment (vowel or consonant), i.e the syllables, words or sentences.

2.2 Motivation

During the initial stages of our work, we were interested in generally researching the field of prosody in human speech.
We started by reviewing this multidisciplinary field from a few points of view and were looking for the standard ways to describe the prosodic part of a speech signal. In other words, we were trying to find a regulated way to describe prosodic phenomena.
We called it a prosodic language - a system that is built from building blocks that describe different aspects of prosody and therefore called **descriptors**.

From a linguistics point of view, the descriptors of the language's prosody can be a writing system or another standard labeling method. We reviewed many previous works, as will described in section 4.1, and found that there are a few such methods. The famous one being the ToBI standard [15].
The common ground for all of these methods is that they label prosodic-events only (i.e., pitch accent, pitch rising and falling, phrase boundary, and more) and do not include a full set of all prosodic manifestations. For example, emotional states affect the prosody in a way that in most cases cannot be labeled or described by these methods.

From an engineering point of view, the descriptors can be features that were extracted from a speech signal and carry prosodic information. We can call them **prosodic features**. As will be explained further in section 4.2, by looking at previous works, we were trying to find a definition for the term prosodic-feature. We also tried to find some standard prosodic features that have been used before. We did find three families of features (pitch, duration, and energy), that are considered to be prosodic in nature by many authors. We also found that out of these three families, we can extract a massive amount of features; however, not all of them necessarily carry prosodic information. On the other hand, other features (such as voice quality [64, 158]) also carry prosodic information but are not part of these three families.

To summarize, to the best of our knowledge, there is currently no standard nor definition of what prosodic features are. We think that this is an important gap, and this is the basis and the primary motivation of this work.

2.3 Research Question and Contribution

We believe that the building blocks or descriptors of a prosodic language can consist of well defined **prosodic features**. Since the amount of features is not limited, they can

describe complicated prosodic manifestations. In addition, these features can describe prosody in a quantitative and mathematical way, which is more accurate than a discrete-symbols-labeling-system that describe the prosody qualitatively.
Following these conclusions, we defined our research question as: **how can we define what is a prosodic feature, and to what extent can a feature be considered as conveying prosodic information**.

In this work, we try to address with this research question by suggesting an optional definition for prosodic features. We also introduce the **Prosodic Feature Criterion (PFC)**: a criterion for determining whether a feature represents prosodic information and to what degree. We show a methodology of calculating and evaluating the PFC.

The contribution of this work is by introducing a novel mathematically formulated criterion which can measure in a quantitative, objective and continuous way the prosodic quality of a feature.

We presented the initial formalism of this criterion in [156] and showed a validation of this work using different feature set in [155].

2.4 Thesis Structure

The structure of this thesis is as follows:
In chapter 3, we present the background of the speech field with an emphasis on prosody. We show it from two points of view: (1) non-engineering as a general background, and (2) engineering , showing how classical speech processing tasks can use prosody.

Chapter 4 elaborates on the different descriptors of prosody. We will first review the standard methods of prosody labeling and then explain what we currently know and what we think are the gaps regarding prosodic features formulation.

Chapter 5 introduces our main development, the Prosodic Feature Criterion (PFC). It includes the main idea of the criterion, a mathematical formulation, and a methodology of calculating the criterion.

In chapters 6 and 7 we present the datasets and feature sets we used in this work.

In chapter 8 we explain the different experiments that we performed. We also show initial results, which helped us in developing the PFC.

Chapter 9 shows results who validate the criterion by comparing PFC scores of different features and by using several validation tests.

Finally, chapter 10 will conclude and summarize this work and discuss future works.

3 Background and Previous Work

The field of prosody has been researched for many years. This field is multidisciplinary and is related to a few different domains such as linguistics, psychology, neurology, speech therapy, computer science, engineering, and more [107].
In this chapter, we review the current research and prosodic applications. Section 3.1 will shortly review some of the non-engineering fields that deal with prosody. Section 3.2 will review engineering applications, in the general field of speech processing and will show how they make use of prosody.

3.1 General Prosody Research

3.1.1 Linguistics

In this domain, we can find works that relate to prosody, such as "what are the relationship between prosody, discourse, and syntactic structure?" or "what is the pragmatic role of prosody in a conversation?" [107]. One of the first works in the field was [3], which is an important experimental study of the linguistic function of suprasegmentals[1], and on the production and perception of suprasegmental features.
Other works test the differences between different languages in a few aspects, for example [170] investigates the similarities in form and function of prosody, between several languages. [78] deals with the role of prosody to identify "foreign accent." Another interesting work is [39], that deals with learning a new language and offers a method based on prosody to improve a second language (L2) accent.

Many authors examine the differences between tonal and non-tonal languages. In tonal languages, prosody has a lexical role, i.e., words with a different meaning, have the same vowels and consonants sequence, but are pronounced using different prosody [157]. These languages are common mainly in the east Asia and in the southern parts of Africa. [10]

[1] The term suprasegmental is defined in section 2.1

tests how emotions are expressed in tonal vs. non-tonal languages and what are the differences. Finally, the work in [12] tests the ability of a speaker who is a tonal language native speaker to learn the prosody of a non-tonal language. A disclaimer: since tonal languages are a unique field of research, they will be excluded from this thesis.

In our work, we are trying to rank and measure the amount of prosodic information in a feature. Influenced by the above examples, finding better prosodic features can help in many automatic tasks such as analyzing language structure, language recognition, etc. In addition, it can also help to improve linguistic analysis.

3.1.2 Psychology and Cognition

These domains ask questions about the perception of prosody by human beings [94] and how people understand and recognize different types of prosodies.
A few examples of research questions about the relation between prosody and psychology and cognition are:

- Testing how do human beings understand speech-acts using prosody.
  A speech-act is defined at the phrase-level. It represents the state or function of an utterance. For example, the function of a question is to request information, while an answer's function is to provide this information - they are both considered speech-act.
  Other examples are statements, greetings, requests, warnings, promises, and more.

- In psychology, research can test the ability to express and understand the emotional state of the speaker through their speech prosody. A few examples: (1) In [66] a few experiments are performed that revealed that music lessons promote sensitivity to emotions conveyed by speech prosody. (2) Tests in [80] shows that people understand emotional prosody when listening to a foreign language. They show that prosody carries both universal and culture-specific cues. One of the case studies in our work that is presented in section 6.1.2, shows how emotional recognition tasks can use features which convey prosodic information.

- Developmental psychology is a sub-field in psychology. Some works in that field deals with the development of a language and prosody. For example, according to [167], infants are sensitive to the prosodic patterns present in the human speech from around two months after birth. They develop some basic knowledge about the

way their native tongue sounds. Therefore they can distinguish between words with different prosodic patterns.

- Other works try to understand and assess language disabilities in children. For example, the authors in [106] developed an automatic intonation recognition, for the prosodic assessment for children with language impairment.

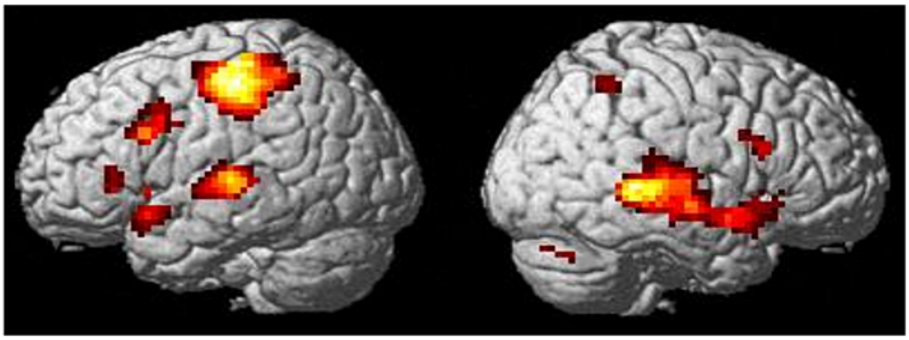


Figure 3.1: fMRI image which shows stronger activity (yellow) of brain regions which react to prosody of simple and complex emotions [121]

3.1.3 Speech Therapy and Neurological Conditions

Many studies in the fields of neurology and speech therapy examine the relations between different medical conditions, prosody, and speech disorders. The domain of speech disorders is important as the production of improper prosody may hurt communicative ability [153]. Examples of research questions are:

- "How is the brain's structure related to prosody and which part of it understands prosody [101]?"
  An interesting example is presented in [67], where an fMRI study of English and Chinese is performed. They examined brain activity and tested which hemisphere (left or right) is more active in different language processing tasks. Figure 3.1 shows fMRI image of the reaction to different types of prosody from a research [121] that compared neural activity during emotions perception in comparison to neutral prosody.

- "How does the human speech-production-system work and how it expresses different prosodies?"

The main organs involved in speech production are: (1) respiratory system which produces the airflow. (2) The laryngeal system which modulates the airflow to be a simple periodic acoustic wave. (3) The supralaryngeal system which provides resonance and articulations, which makes the acoustic wave carry complex speech sounds. A chart of these organs can be seen is figure 3.2.

- "How do different pathologies affect speech?"
  Disorders that hurt the ability to produce standard prosody are usually related to either: (1) pathologies of muscles of phonation and articulation, e.g., vocal fatigue, aberrant speech and voice patterns such as vocal fry [153], or to (2) speech disorders with a primary neurological factor.

  Two well known neurological conditions that affect prosody are: (1) Parkinson's disease (PD), which about one-half of all its patients exhibit speech disorder which is related to prosodic parameters [7, 8]. (2) People with Asperger's Syndrome (AS), that many times show poorly developed skills in understanding emotional messages which are conveyed using prosody [92].

Many works try to find and evaluate new treatment methods for speech disorders. In general, we can say that "prosodic exercises" can achieve significant clinical and psychological improvement for prosodic speech disorders and for a variety of communication disorders [102, 7, 8]. An example of the importance of speech therapy can be seen in [44], that compares speech therapy, surgical and pharmacological solutions to treat speech disorders in PD patients. It shows that speech therapy is the most efficient therapeutic method for improving voice and speech function. Another example is [152], that developed prosodic exercises (an imitation task), that can help with the development and enrichment of prosodic abilities in children with Autism Spectrum Disorders (ASD). The work in [101] discusses the requirements of a clinical effective assessment instrument and treatment programs, for prosodic problems.

During the work on this thesis, we considered developing two applications that can make use of our criterion (the PFC). They are also inspired by some needs raised by previous works: (1) an automatic tool for objective assessment and monitoring of some speech disorder. (2) a treatment tool that is inspired by bio-feedback methods. The application we thought of would visualize the current prosody produced by the user and "how far" it is from the targeted prosody.

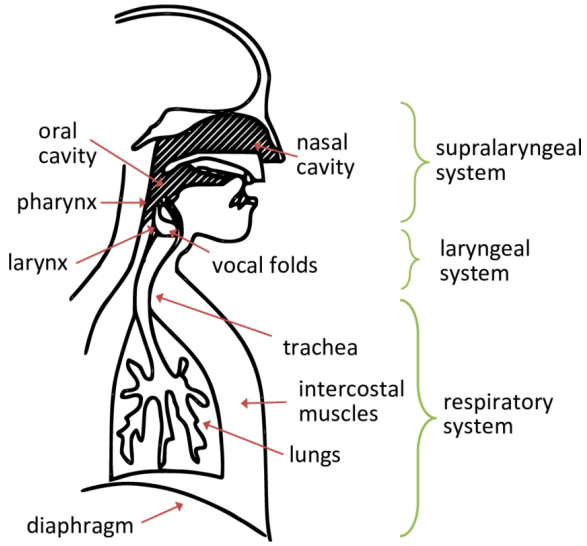


Figure 3.2: The main organs that are involved in speech production divided into three systems: respiratory, laryngeal and supra-laryngeal systems (image taken from [161])

3.2 Engineering Research of Prosody

As we have shown in previous sections, prosody carries important information and is affected by many different factors, such as the speaker's intention, medical condition, emotional state, and more. Therefore it is not surprising that prosody is widely used in many speech-related applications.

This section will cover some of the works under the domains of computer science and engineering that deal with prosody or may use prosody. The review is divided into three parts: *Prosody Based Analysis (3.2.1)*, *General Speech Processing Tasks (3.2.2)* and *Speech Generation (3.2.3)*.

3.2.1 Prosody Based Analysis

This part includes systems whose primary purpose is "prosody classification," i.e. determine which prosody was used in the speech utterance.

Speech and Dialogue-Act Detection

As defined in 3.1.2, dialogue-acts and speech-acts represent the function of an utterance in a dialogue or the speaker's intention in an utterance. Common examples are questions,

statements, greetings, hesitations, and more. Modeling and automatically identifying the structure of spontaneous dialogues is important to better interpret and understand dialogues. Many applications can use this dialogue structure, for example: human-computer dialogue applications, machine translation, Automatic Speech Recognition (ASR) systems, topic identification [27], natural speech understanding [32] and more.

Most of the works agree that prosodic information is essential for tasks like speech and dialogue-acts detection and also for the speaker's intention detection. Prosody can be used for recognition tasks or as a performance assurance for detection tasks [88]. A few examples of common tasks in this field are: (1) using prosodic features for the task of boundaries-detection of dialogue-acts [48, 47, 38]. (2) using prosodic features for recognition of speaker intention. A few examples are irony detection [60, 111], sarcasm detection [129] or interrogative-intonation detection, like the works in [79, 88], that extended previous works on French, and found that using prosodic features is a useful way for these tasks.

Following these kind of works, many engineering systems use prosody. Some of them, as a single type of information, e.g., [32], that shows that prosody made a significant contribution to the dialogue-acts recognition and classification tasks. Other works use a combination of prosodic and lexical information, e.g., [32, 58, 31].

In this work, we will use a question detection task to validate our criterion. This task is a common dialogue-act detection task. The dataset and the experiments are presented in section 6.1.1.

Prosodic Events Recognition

In this task, the goal is to automatically detect and classify events such as pitch-accents, stress, emphasis, brakes, and phrase boundary in a speech signal [151]. It includes two sub-tasks [114]: (1) detection and localization of the presence or absence of a prosodic event in a sentence, and (2) classification - what is the type of a prosodic event [22]. This task can be helpful for automatic data labeling.

3.2.2 General Speech Processing Tasks

In this part, we review two of the most traditional tasks in speech processing: (1) Speech To Text (STT) and (2) speaker and language-related tasks. Most of these automated tasks do not use prosody.

Speech To Text (STT)

Speech To Text systems automatically transcribe human speech. The input is the speech signal and the output is the automatically produced transcript [135]. This research area is very important and can assist in many fields, for example:

1. Human-Computer Communication: can be more heavily based on speech as speaking is a natural form of communication for humans[112, 174]. Some examples: mobile devices, voice search [116] and personal assistant such as Siri or Cortana [147].

2. Human-Human Communication: can also be improved using STT, for example, by using speech to speech translation systems for communication between people who speak different languages.

STT has been an active research field for over five decades [174, 112], but many difficulties have arisen during these years. The first basic systems arrived just around the 1980s, and they were very limited [112]. Even today, after many successfully deployed commercial products, lots of times, the performance of STT in real-life is not always accurate enough. This is especially notable in difficult conditions such as overlap speech or noise [115, 132]. In recent years, performance has been improving, so that many applications have started to use speech as a significant part of the system [112, 174].
There are many reasons why STT is so difficult. Other than the technical challenges, the variability in human speech is a very complicated issue: people speak in different styles, and there are many differences between ages, dialects, and accents. Even if we analyze the speech of one specific person, the way he/she talks can change dramatically in different scenarios. The environment also has a significant influence: background noises, side talks, distortion, and additional factors have a substantial effect [135, 174]. We can summarize by using a basic rule of thumb stating that as we move from constrained tasks to real-world, STT becomes more complex [174].

As explained in section 2.1, speech contains two types of components: lexical and prosodical information. A naive approach can suppose that STT systems (which are intended to extract the lexical content), do not necessarily have to use prosody. Indeed, most of STT applications these days do not use prosody [90]. Nevertheless, prosody actually carries essential information that can be helpful for STT [151]. The main reason is that prosody carries suprasegmental information (syllable, word or utterance level), in contrast to the traditional acoustic features which are usually extracted over a very narrow window

and therefore miss some useful information [90]. A few examples where prosody can help are: (1) on a **word level**: different words are written the same way but have a different meaning depending on the word stress. These differences can only be distinguished according to stress (which is part of prosody), e.g., the word record which has two meanings: (a) **re**cord: as in "medical records" or the term "for the record," and (b) re**cord**: as in "tape-recording" or "recording sounds." (2) On a **sentence level**: the same sequence of words can have different meanings depending, in part, on prosody [32]. For example, the sentence, "did you go to school?". By stressing the word "you," we are asking **who** went to school, and by stressing the word "school," we are asking **where** did you go.

The use of prosody can help make STT systems more robust [165] and there are several (but not many) previous works which have already begun to research in this direction e.g. [172, 113].

Speaker and Language Recognition

These fields have been researched for many years [24]. We can divide them into a few types of tasks (1) *Speaker Recognition* that includes many applications such as (1.a) *Speaker Verification* ("Is it really X who is talking?"): these systems receive a speech utterance and a claimed person identity as an input and need to verify that the utterance's speaker is indeed the claimed person [61, 55]. (1.b) *Speaker Identification* ("Who is talking?"): these systems receive a speech utterance as input and need to decipher who is the speaker out of a list of known speakers or to classify the speaker as an unknown [30, 85]. (2) *Language Recognition* ("Which language is it?"): these systems receive a speech utterance as input and need to recognize the language of the utterance.

There are several possible usages for speaker and language recognition, including security systems that can confirm a person's identity in phone calls or as an entrance control system which is based on voice [109]. Language recognition systems can be used for routing incoming calls in call centers or for emergency services where the person does not know how to speak the local language [24] or even as a primary system before the ASR system which recognizes the language and decides the relevant ASR model.

Humans can distinguish between languages and voices of different people, relying also on prosody. Two main factors that help this ability are: (1) physiological parameters such as vocal tract shapes, larynx sizes, and other parts of the voice production organs, which are different between people and affect the voice production. (2) Every person has their own style of speaking, including rhythm and pronunciation pattern, intonation, and more

[109]. This is also related to language distinction because every language has a different and unique style that expressed by its melody, rhythmic pattern, stress locations, and accent.
These two factors are mainly considered to be prosodic parameters, and therefore, understanding the prosodic features can be very useful for recognition systems.

3.2.3 Automatic Speech Generation

In this section, we explain the process of generating speech signals using multiple types of inputs (either text or speech). Two common tasks are (1) text-to-speech synthesis, which is generating speech from text, and (2) voice transformation, which is changing the way a speech utterance sounds. As these tasks produce speech, prosody is a crucial component of the process.

Text To Speech (TTS)

This is the task of synthesizing artificial human speech using text [149, 140]. This field has been researched for many years, and the first TTS systems were developed during the 1960s.

TTS can serve multiple applications such as: (1) systems that read-out-loud stories, news, reports, etc. for example to children or visually-impaired people. (2) automatic responses in call centers [168], (3) navigation systems (e.g., Waze) which read driving instructions while the driver continues to look at the road, (4) artificial personal assistants which communicate over voice [147] and more.

In order to asses the quality of a TTS system, it is common to use two measurements: (1) *intelligibility*, which measures how well the listener can understand the output message and (2) *naturalness* which measures how closely the output sounds like human speech. Most systems try to maximize both of these parameters.
Even though intelligibility is a necessary condition to start using the system, it was found that naturalness is also very important. It affects the level of comfort of the listener therefore influences the amount of use of these systems. Naturalness tests several aspects, including the quality of the voice, language level nuances, consistency of prosody, and more [148, 168]. Prosody plays an essential role in the naturalness experience of the user in these systems.

The first TTS systems were of low quality and sounded very mechanical, and indeed few people used them [168]. As the intelligibility became better, the need to improve the

naturalness became more important [54]. In recent years, we can see that there is a significant advancement in quality, which comes together with the adoption of this technology [168]. An example of one aspect of naturalness that is still missing in some TTS output is emotional expressions, which is still in its very primary stages of development [54].

In recent years the use of Deep Neural Networks (DNN) became popular, and there are many works which replaced part or all of the traditional TTS pipelines with DNNs and achieved a very impressive and human-like output [138], e.g., DeepVoice [149], WaveNet [146], Char2Wav [148].

Voice Transformation (VT)

Voice transformation (VT) is another field that deals with speech generation. Its main goal is to change or modify one or more speech's parameters [104] while keeping the same lexical content, i.e., modifying the non-lexical information of the input speech.We can divide VT into two main categories:

1. Changing the speaking style (i.e., "how does it sound?"), while keeping the utterance sounds like the same speaker is speaking [13]. In the context of this work, changing the prosody of an utterance can be considered as VT. Examples for applications can be changing the emotional state [89, 71], and other parameters of the speech that are related to style [127, 130].

2. Changing speaker-related characteristics, so that the output will be perceived like a different speaker is speaking [150, 139]. It can be done either by changing the source speaker into a specific target speaker [11] or by blending a few voices into a new artificial voice [49].
   Many applications can use it, for example, for entertainment purposes such as movie dubbing while maintaining the voice of the original actor, or as a singing voice conversion [100]. It can be also be used in security applications, either to protect a person by changing their voice or for fraud purposes against speaker verification systems [133, 134]. Speech-to-speech translation systems can use it to maintain the voice of the original speaker [171, 86]. TTS systems can also use VT to personalize the output and make it sound like a specific person [139], e.g., for people who lost the ability to speak and use a speech synthesizer [36, 87].

Even after many years of research VT systems still suffer from quality problems, especially on the naturalness level. Because these systems deal with prosodic parameters,

understanding the prosodic features better can lead to better performance.
One of the main reasons for VT quality issues is that lots of systems perform the transformation on the frame-level. We recall that prosody is suprasegmental by definition, i.e., related to more than one segment (longer than a frame). Therefore, it is obvious that systems that working on a frame-level will suffer from issues on the prosodic level.
The work in [150], explicitly stated that "the main challenge is the absence of certain high-level features during conversion, which hugely affect human prosody." In regards to future works they mentioned that "developing more complex prosody models... is an important research direction," and it "would enable the capture of complex prosodic patterns and thus enable more effective transformations."

3.3 Summary

To summarize, in this chapter we tried to explain the importance of prosody by reviewing previous works that deals with prosody.
We started by showing the massive research on prosody that has been done from different perspectives in several non-engineering domains such as linguistics, psychology, neurology and speech therapy.
We then reviewed the engineering point of view and showed three different types of tasks: (1) "prosody based analysis" that includes tasks whose main goal is to classify what kind of prosody has been used, (2) the added value prosody can contribute to "general speech processing" tasks that are not necessarily using prosody, and (3) speech generation tasks that have to use prosody to reach their goals.

4 Prosodic Descriptors

The motivation of our work is that we would like to find a way to describe the prosody of a speech utterance (see section 2.3). In this chapter, we will elaborate about prosody's "descriptors" from both linguistic and engineering points of view.
We will first review the importance of prosody labeling and will show a few initial attempts that have been made towards prosody formalism. Then we will describe the ideas that led us to develop the Prosodic Feature Criterion (PFC) (presented in chapter 5).

4.1 Prosody Labeling

A writing system and standard notations are common for most languages. They serve as a documentation tool and as a way to express the language using symbols. Two examples are: (1) the language of music, whose standard notations are composed of notes and rests that represent the pitch and the rhythm of the music. (2) Human verbal languages, that in most cases have many types of writing systems. A few examples are (a) the alphabet method, e.g. the Greek, Latin and Arabic alphabet , and (b) the symbolic methods who can represent either words or phonemes (e.g. Hieroglyph).
But what about prosody, and how can we transcribe or label it?

A standard annotation system for prosody is crucial for researching this field, and can be used to learn the relations between the (labeled) prosodic events to the speech signal, or the relations to other linguistics events such as lexical, syntactic and semantic structure [77]. In the last decades, there has been extensive work towards standardization of prosody transcription [77], although the nuances of prosody are often hard to express on paper.

4.1.1 Types of Labeling Systems

There are two types of labeling systems:

Encoding Systems

This type of systems encode linguistic events by discrete categories or by translating auditory information into symbols. These systems are quite popular. The best known labeling system is the *Tones and Break Indices (ToBI) standard* [15].
ToBI's transcription includes two classes of symbols:

- **Tonal events** which in general track pitch dynamics (i.e rising or falling of pitch contour). Tonal events contain sub-classes such as: (a) *Pitch accents*, which describe prosody on a word-level. These usually indicate the most informative words in the utterance. (b) *Boundary tones*, which mark the edges of a phrase. (c) *Phrase accents*, which are the tones between a pitch accent and a boundary tone.

- **Brake indices** which indicate how strong is the break between words.

The original ToBI standard was developed for American English. Over the years, other ToBI systems were developed for different languages. A few examples are J-ToBI for Japanese, K-ToBI for Korean, GR-ToBI for Greek, G-ToBI for German [57] and more.

There are also other and less-known prosody labeling systems. For examples, [45] whose authors noted that manual labeling using ToBI is less reliable and therefore created a simpler and more robust annotation system which is called ToBI-lite. [65] is an additional example, which shows a development of a system for automatic prosody labeling. The authors of this paper created their own labeling scheme.
In general, the ToBI system is very comprehensive but difficult to label manually. It was found that simpler prosodic representations are good enough for certain speech applications such as disfluency, sentence boundary, and dialog act detection [29, 81, 32].

Parametric Systems

In contrast to "Encoding Systems" which try to give some linguistic interpretation to acoustic events in the speech signal, the "parametric systems" aim only at describing some objective measure of the signal, mostly the pitch contour [95].

There are some works in this field, but most of them are not new. One example is the TILT intonation model [37]. It analyzes the F0 contour and represents the intonation as a sequence of parameterised events. The events are pitch-accents and boundary-tones, and they are both characterised using the same parameter called Tilt. The authors claim that

the Tilt model is a more appropriate and powerful than ToBI.
Other examples of parametric models are the Fujisaki model [6], and International Transcription System for Intonation (INTSINT) [18].

We were interested in this type of models, even though they are not so new, as they all claim that there are some limitations to ToBI, and try to find alternatives. They are all showing the need for a more objective and numerical way to describe and model prosody. This type of models, in addition to the attempts for automatic labeling (see in section 4.1.3), led us to use representations of prosodic descriptors, namely prosodic features. We will describe the prosodic features in section 4.2. These features are one of our main motivations for developing the PFC.

4.1.2 Limitations of Labeling Systems

As we can see, there have been many attempts to create different annotation systems. Still, there is no one universal methodology that has been accepted [83]. There are several common issues with the current existing labeling methods, which among others, lead to a limited amount of prosody-labeled datasets [77]:

1. **Complexity**: the main issue, which has already been mentioned, is the complexity of manual labeling which makes the labeling process expensive and time-consuming [114]. . Indeed, the majority of available training data for most speech applications, miss manual labels of prosodic information [143].

2. **Discrete symbols**: most of the labeling methods represent prosody by discrete events that are less accurate than continuous methods. Representation of prosody using this way leads to a loss of a large amount of information. Examples:
   (a) Different types of prosodic events can be mapped to a single symbol. E.g., all of the events where low pitch rising to high pitch could have one symbol.
   (b) Discrete symbols can be considered as a signal that was sampled using a very rough "sampling rate". Therefore they cannot represent high-frequency events. These types of events can carry prosodic information, e.g., the numerical features Jitter and Shimmer (see section 7.1.3).

3. **Subjective methods**: manual tagging performed by humans is subjective. Two different labelers can label the same prosodic event in different ways, or different

prosodic events can be labeled as the same type of event. We should note the subjectivity is a general issue with most types of tagging and not unique to prosody. Phonetic transcription, for example, also suffers from the same issue.

4. **Multiple methods**: there is no one agreed-upon system that fits all of the languages. This is because most of the labeling systems encode linguistic events that differ between languages. The ToBI authors themselves mentioned this issue [82].

5. **Partial symbols set**: most of the labeling methods have symbols for **some** prosodic events only, such as breaks, tones-direction (going up or down), phrase and boundaries of a dialogue act. These kinds of symbols do not cover all the prosodic information. For example, speaker's gender, age, or emotional state, can definitely be understood from the speech prosody but are not represented by these common prosody labeling systems.

4.1.3 Automation of Labeling Systems

In order to solve a few of these issues, many works have tried to develop systems for automatic labeling. These systems should be faster, objective, and hopefully more accurate, and can improve the performance of many speech processing applications [77, 65, 45], for example: (1) Text To Speech (TTS) systems, which could more rapidly adapt to new speakers and new domains [40] by using more prosody annotations. (2) Speech-to-speech machine translation, that can use annotations for correct word emphasis transfer [143].
Creating automatic systems require the use of measurable properties which will be described in the next section as prosodic features.
Initial attempts using acoustic information alone have been reported in [17]. Many other works tried to add lexical and syntactic cues, like in [77, 65, 68], which uses supervised learning (and therefore, was restricted to small labeled training datasets). The work in [74] presents an unsupervised word-prominence labeling algorithm. Another unsupervised algorithm of annotates-accent and boundary-events in speech can be found in [77].
Some works are only based on text, like [16] for an accent-labeling task using syntax, or [19], which predicts prosodic stress from parts-of-speech (POS). The work in [51] shows a method that sped up manual labeling by using an automatic system based only on text and then manually correcting its results using humans.

In conclusion, many prosody labeling standards were developed; some are simpler to tag than others. Because of a few drawbacks, many works were tried to develop automatic

labeling systems that use acoustic features. These features can describe and represent the prosodic building blocks which are perceived by the listener. Nevertheless, little attention has been given to generalizing and formulating these features, a step which is crucial in the way to exploit their full potential.

4.2 Prosodic Features

There are a few facts about prosodic features that are widely agreed upon: (1) they are a subgroup of the speech features. (2) They are related to some prosodic manifestations. (3) They are usually quantified to their mathematical form influenced by a parameter that was researched by linguistics. For example, F0 is the mathematical form of Pitch, which is a qualitative parameter used by linguistics.

Three of the best known features families which are considered to be prosodic according to many previous works such as [77, 65] are (1) **F0**: measures the pitch of the voice, (2) **duration**: measures timing aspects such as tempo and rhythm of segments such as phonemes or words, and (3) **energy**: related to the loudness of the speech.
we elaborate on these feature families in section 7.1.3.

It is widely accepted that these three feature families are indeed not related directly to the lexical content of the sentence [1] (i.e., the words) but to the manner of the speech (i.e., the prosody). Based on that, many authors used some variation of these feature families and called them - prosodic features.

These features families indeed well describe prosodic manifestations of some datasets and some prosodic classes. Nevertheless, there are some issues that have to be addressed. The primary issue is that these three are an incomplete set, i.e., they are not the only features that can be considered prosodic [158].
One example of a feature that may also be considered prosodic is *voice-quality*. This feature has been shown to contain prosodic related information, such as significant correlations with the manner of speech and with speech-act [64, 158]. It even has been suggested to call it the fourth prosodic parameter [65]. Voice-quality can be hard to estimate reliably, and therefore many works do not use it in practice [65].

There has been extensive work on extracting various acoustic and spectral features for both specific prosodic research (e.g., classification of prosodic events, see 3.2.1) and

[1] We recall that we excluded Tonal Languages from this work. Therefore we can claim that features such as pitch are not related to the lexical content.

general use of prosody for other tasks (e.g., better emotion detection, see 3.2.2). Different works used different prosodic features but in general there are two common grounds:

1. The vast majority of the works used features that fitted their own classification or detection task and did not look for a descriptive nature of the feature. Meaning, they looked at the features selection or extraction process from the "task point of view" and not from the "features point of view." These works probably chose a feature set that maximized the performance of their trained classifier (i.e. separated between the task's classes). However, these features are not necessarily the best ones that would describe the classes.

2. Most of the papers go directly to one or more of the above three feature families (pitch, duration, and energy) and do not search for new ones. A few examples are: the LLD[2] that has been used in [83] was F0, and the functionals[3] were related to F0 direction and peak positions. [77] uses intensity, F0, and timing as LLD. [45] uses common and standard functionals such as min() and max(), while [65] uses self-developed methods such as functionals that perform a non-linear transformation. A different representation scheme for modeling prosodic features has been proposed in [95]. They use an n-gram model that creates a kind of prosodic contour but still uses the same three feature families as their LLDs.

An interesting project for extracting speech features is called OpenSMILE [108]. This project is an open-source toolkit for extracting thousands of types of acoustic and spectral features; it is widely used and was cited by thousands of papers. It includes many features that are considered to be prosodic. Lots of them are derived from the above three families, but many others are not. In this thesis, we use OpenSMILE to validate our criterion results. We elaborate on our use of OpenSMILE in section 7.2.2.

To summarize, we can say that numerous papers mentioned the term prosodic features. Nevertheless, to the best of our knowledge, there is still no proper and complete definition for what prosodic features are.When someone suggests a new prosodic feature, there are no objective measures to test this feature while taking prosody into consideration. Therefore the features' selection process becomes complicated and not trivial.
We believe that this is a significant gap, and there is a need to better define which features can be considered prosodic. This gap is the basis of our work and the primary motivation to develop the Prosodic Feature Criterion - PFC.

[2] LLD - Lower Level Descriptors. Defined in section 7.1.1
[3] Functionals. Defined in section 7.1.1

5 Prosodic Feature Criterion (PFC)

In previous chapters, we discussed the importance of prosody in many research fields. We showed it by reviewing works in a few different disciplines which are using prosodic information (chapter 3). Then we discussed several ways to describe prosodic manifestation (also called "prosodic descriptors") from both linguistics and engineering points of view. That led us to present the gap we believe we found in the current definition of **prosodic features** (chapter 4).

In this chapter, we are going to present our suggested methodology to overcome some of this gap. We introduce the Prosodic Feature Criterion (PFC) - a criterion that can be used to determine whether a single feature carries prosodic information, and to what degree.

We will first explain our motivation and the criterion requirements in a qualitative way (section 5.1.1). Then, we will mathematically formulate these requirements (section 5.1.2). Finally, we will present a general methodology for calculating the PFC score, including a few specific examples (section 5.2). We will also discuss the suitable functions that should be used in this methodology (section 5.3).

5.1 The Criterion

5.1.1 Motivation

We wish to define a criterion for measuring how well does a feature represent prosodic information conveyed in speech utterances. We aim that the criterion will be simple and suggest that a prosodic feature should satisfy the following requirements:

1. A prosodic feature, should be **dependent** on some prosodic manifestations.

   Meaning that: if the prosody of a speech utterance is changed, the values of the relevant prosodic features should be changed as well.

Clarification: we do no suggest that **every** prosodic feature is dependent on **every** prosodic manifestation. On the contrary - most of the prosodic features are affected just from **a few** prosodic manifestations.
Examples: (1) the sentence "That is just what I needed today!" can be said happily or sarcastically. (2) The sentence "this is what you need" can be said as a statement or as a question.
We expect that changing between these prosodies (for each example), should affect and change the values of a prosodic feature.

2. A prosodic feature should be **independent** of changes in non-prosodic speech parameters.

   Meaning that: if some parameter of the speech utterance is changed (e.g., the content of the utterance), while the same prosody is used, the values of a prosodic feature should not be changed.
   An example: both sentences "What is your name?" and "Where are you?" have different content but can be uttered in the same question prosody. This change in content should not affect the values of a prosodic feature.

Throughout this work the second requirement relates most of the time only to the lexical content of the utterance. This is because we use an oversimplification where only two attributes characterize an utterance - prosody and content.
The term "lexical content" refers to the sequence of the words in the utterance. This means that when we say that there is "change in content," we refer to any change, either acoustic or semantic. For example, the utterances "The kid likes cats" and "The child love kitties" are semantically similar but still considered as utterances with different content.

We also assume that prosody and content are independent of each other. Naturally, this assumption does not hold for some languages such as tonal languages, where different tonal patterns convey content [118]. Therefore these languages will have to be considered separately.

5.1.2 Mathematical Formulation

Following the above requirements, we mathematically formulate the criterion. We will start by defining a more general (and weaker) criterion by referring to prosody only, and not to the content. Then we will show how we can also refer to the content by simply changing the range of two indices.

Notations

Suppose we have a set of utterances U_{pc}^{k}, where $p = 1,2...N_p$ is an index representing the different prosodies in our dataset and $c = 1,2,...N_c$ is an index representing the different content types. N_p and N_c are the number of prosodic and content classes respectively. The index $k = 1,2,...N_{pc}$ runs through all utterances of type pc, i.e., there are N_{pc} utterance with prosody p and content type c.
From that point, when using the notation $\forall p,c,k$, we mean that p, c or k stand for all values between 1 and N_p, N_c or N_{pc}.
F_{pc}^{k} is an instance of the examined feature F extracted from the utterance U_{pc}^{k}.

Conditions For Prosodic Feature

We suggest that feature F would be denoted prosodic if the following conditions hold:

Condition A:

The dissimilarity between the extracted values of F is sufficiently small, for most utterance pairs with the **same** prosody p:

$$Pr\left(d_{same}^{p} < T_1\right) > x_1 \tag{5.1}$$

$$d_{same}^{p} \in D_{same}^{p} \tag{5.2}$$

where $Pr\left(\cdot\right)$ is the probability function, T_1 is a threshold we use to define "sufficiently small" and x_1 is a threshold we use to define "most of the utterances".
The set D_{same}^{p} (eq. 5.3) contains random variables over the probability space that represents the dissimilarity between the values of the feature F that were extracted from two different utterances k and l with the same prosody p:

$$D_{same}^{p} = \left\{ d_F\left(F_{pc}^{k}, F_{pr}^{l}\right) \right\}, \forall c,r,k,l \tag{5.3}$$

where $d_F\left(\cdot,\cdot\right)$ is a function that measures the dissimilarity between features of two utterances.
Note that this condition deals with prosody only, so D_{same}^{p} can include dissimilarities between pairs of utterances with same content or different content.

Condition B:

The dissimilarity between extracted values of F for utterances with **different** prosodies is high enough for most such utterance pairs.

$$Pr\left(d_{diff}^{p,q} > T_2\right) > x_2, \forall p \neq q \tag{5.4}$$

$$d_{diff}^{p,q} \in D_{diff}^{p,q} \tag{5.5}$$

where T_2 is a threshold for "high enough" and x_2 is a threshold for "most of the utterances". $D_{diff}^{p,q}$ (eq. 5.6) is a set contains random variables over the probability space that represents the dissimilarity between the values of the feature F that were extracted from utterances pairs with different prosody:

$$D_{diff}^{p,q} = \left\{d_F\left(F_{pc}^{k}, F_{qr}^{l}\right)\right\}, \forall c, r, k, l, p \neq q \tag{5.6}$$

The thresholds T_1, T_2, x_1 and x_2 should be tuned for each feature F. The function $d_F(\cdot,\cdot)$ can be also defined for each feature separately, while taking into account feature dimensionality.

As mentioned at the beginning of this section, these conditions can hold for both cases of "same" or "different" content type, and therefore they only satisfy requirement 1 from section 5.1.1 (dependency of prosodic manifestations). They will be naturally stronger if we also add requirement 2 from section 5.1.1 (independence of non-prosodic parameter). In order to do that, we should change the above conditions such that:

- In condition A the content is changed, i.e. $c \neq r$
- In condition B the content is unchanged, i.e. $c = r$

5.2 Methodology for PFC Calculation

In order to calculate the PFC based on conditions A (eq. 5.1) and B (eq. 5.4), we should tune the thresholds T_1, T_2, x_1 and x_2. We should also estimate the conditional distributions of $d_F(\cdot,\cdot)$ given DIFF (using $D_{diff}^{p,q}$ subsets), and $d_F(\cdot,\cdot)$ given SAME (using D_{same}^{p} subsets). Estimating these thresholds and distributions is not trivial and requires a large amount of data. Therefore we propose an alternative way which is based on the principles of the above conditions A and B, and using a simpler and numerical methodology.

The following subsections will describe this methodology, step by step. We will present the general case where we have N_p prosodic classes, and for each step we will also show an example for the binary case where we have only two classes of prosodies, i.e., $N_p = 2$.
Figure 5.1 illustrates the methodology pipeline, where each step of the methodology is represented by a different block. The numbers in the parenthesis are the steps numbers.

As a part of the methodology presentation we use several functions. We elaborate on the possible implementation of these functions in section 5.3. In addition we mention distance and dissimilarity functions, a short review about this topic can be found in Appendix A.

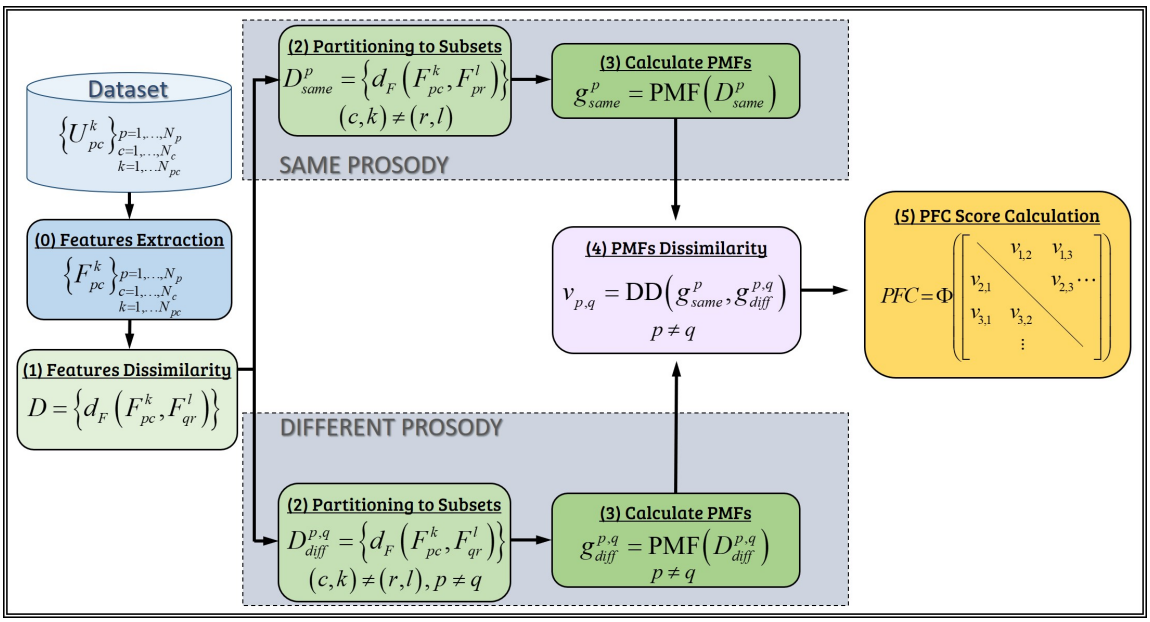


Figure 5.1: A block diagram describing the methodology's steps for calculating the PFC.

5.2.1 STEP 0: Features Extraction

For each utterance U_{pc}^k in the dataset, extract the examined feature F. The extracted values are denoted F_{pc}^k, i.e. this feature's values for the k^{th} utterance with prosody p and content c. To avoid confusion, we assume that F_{pc}^k is a vector. The set of all these extracted vectors is denoted as:

$$\left\{F_{pc}^k\right\} \forall p,c,k \tag{5.7}$$

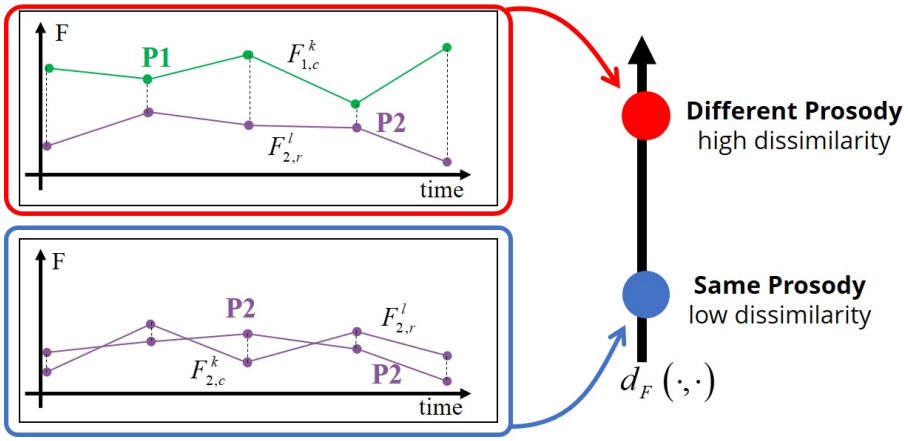


Figure 5.2: STEP 1 - artificial example of $d_F(\cdot,\cdot)$ function outputs. Feature's values of utterance pairs with different-prosody (top) are expected to yield higher dissimilarity-values than utterance pairs with the same-prosody (bottom).

5.2.2 *STEP 1: Features Dissimilarity*

Calculate the dissimilarity-values between the features' values, over all possible utterance pairs in the dataset. The dissimilarity-values are denoted as the set:

$$D \triangleq \left\{ d_F\left(F^k_{pc}, F^l_{qr}\right) \right\} \forall p,q,c,r,k,l \tag{5.8}$$

We recall that $d_F(\cdot,\cdot)$ is a function that measures the dissimilarity between feature's values of two utterances, either scalars or vectors.
For the purpose of this work and in order to simplify the calculations, we require that $d_F(\cdot,\cdot)$ will satisfy the following: (1) non-negativity: $d_F(x,y) \geq 0$, (2) $x = y \Rightarrow d_F(x,y) = 0$, and (3) symmetry: $d_F(x,y) = d_F(y,x)$.

An artificial example for the binary case can be found in figure 5.2. This figure shows 2 pairs of feature's instances, the top with different prosodies (P1 and P2) and the bottom with the same prosody (P2). The top pair received higher dissimilarity-value than the bottom one, as they have different prosody.
Different implementations of $d_F(\cdot,\cdot)$ functions will be discussed later in section 5.3.

5.2.3 *STEP 2: Partitioning to Subsets*

Partitioning the set D into two types of subsets according to the utterances' prosodic classes:

1. Same prosody:
   These sets include the dissimilarity-values of all instances with the **same** specific prosody p. In total we have N_p sets denoted as:

$$D^p_{same} = \left\{ d_F\left(F^k_{pc}, F^l_{pr}\right) \right\}, \forall (c,k) \neq (r,l), p \in \{1,..,N_p\} \tag{5.9}$$

   Figure 5.3a illustrates this step for the binary case where we only have the sets D^1_{same} and D^2_{same}. In the figure we can see two subsets, each with dissimilarity-values (represented by lines) between pairs of utterances (represented by circles) with the same prosodic class. Figure 5.4 (left) illustrates the same for the case of tree prosodies ($N_p = 3$).

2. Different prosody:
   These sets include the dissimilarity-values between all instance pairs where their prosodies are different ($p \neq q$), denoted as:

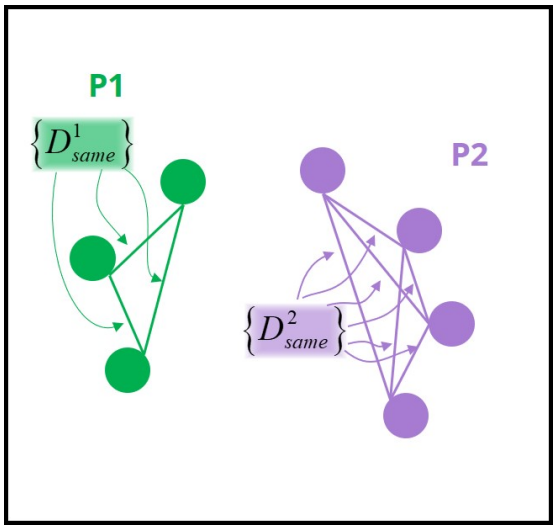
$$D^{p,q}_{diff} = \left\{ d_F\left(F^k_{pc}, F^l_{qr}\right) \right\}_{\substack{1 \leq c,r \leq N_c \\ 1 \leq k \leq N_{1c} \\ 1 \leq l \leq N_{2r}}} \tag{5.10}$$

   The number of sets is all of the possible ways to choose 2 different classes out of N_p classes, which is: $\binom{N_p}{2} = \frac{N_p(N_p-1)}{2}$. We should also note that $D^{p,q}_{diff} = D^{q,p}_{diff}$ and that is because $d_F(\cdot,\cdot)$ is symmetric.
   Illustration of the binary case, where we have only $\binom{2}{2} = 1$ set denoted $D^{1,2}_{diff}$, can be seen in figure 5.3b. This set contains dissimilarity-values (represented by lines) between pairs of utterances (represented by circles), when one is with prosody P1 and the other with prosody P2. An extension of this graph for $N_p = 3$ can be seen in figure 5.4 (right).

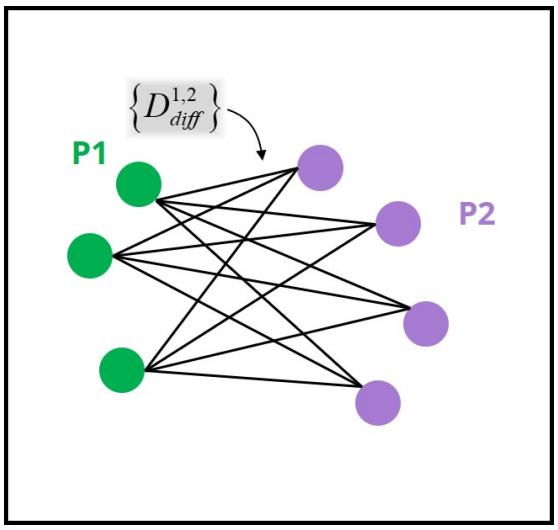
Figure 5.5 illustrates this step over real data. It shows the dissimilarity-values of the F0_mean feature (out of the Initial feature set, see chapter 7), that was extracted out of the Hebrew Q&N dataset that we use in this work (see chapter 6). On top we can see all the dissimilarity-values as a single set, and on the bottom - partitioned into two subsets.

*5.2.4 **STEP 3: Calculate PMFs***

For each of the sets mentioned in STEP 2, estimate its distribution. These sets contain continuous values and therefore, naturally should be represented using continuous random



(a) The sets D^1_{same} and D^2_{same}. The lines connect only between circles with the same prosody. Each set contains the dissimilarity-values between all utterances with the same prosody.



(b) The set $D^{1,2}_{diff}$. The lines connect only between circles with different prosody. This set contains the dissimilarity-values between all utterance pairs with different prosody.

Figure 5.3: STEP 2 - illustration for $N_p = 2$. The lines represent dissimilarity-values, while the circles represent utterances.

variables with Probability Density Functions (PDF). However, due to the relatively small datasets and for computational reasons, we will treat them as discrete random variables and evaluate their Probability Mass Function (PMF), denoted as:

$$g^p_{same} = \text{PMF}\left(D^p_{same}\right) \forall p \tag{5.11}$$

$$g^{p,q}_{diff} = \text{PMF}\left(D^{p,q}_{diff}\right) \forall p \neq q \tag{5.12}$$

where $\text{PMF}(\cdot)$ is a function which estimates the PMF of the set values, for example by using the normalized histograms of the set values. Note that $g^{p,q}_{diff} = g^{q,p}_{diff}$ because the sets $D^{p,q}_{diff} = D^{q,p}_{diff}$.

Figure 5.6 shows example of PMF calculation given subset D^1_{same}. On top - the first stage. Calculating the histogram of the set values. On the bottom - the rest of the stages. Normalizing the histogram by the number of elements in the set, and smoothing the result. This distribution describes real data of F0_mean feature (Initial feature set) that was extracted out of the Hebrew Q&N dataset.

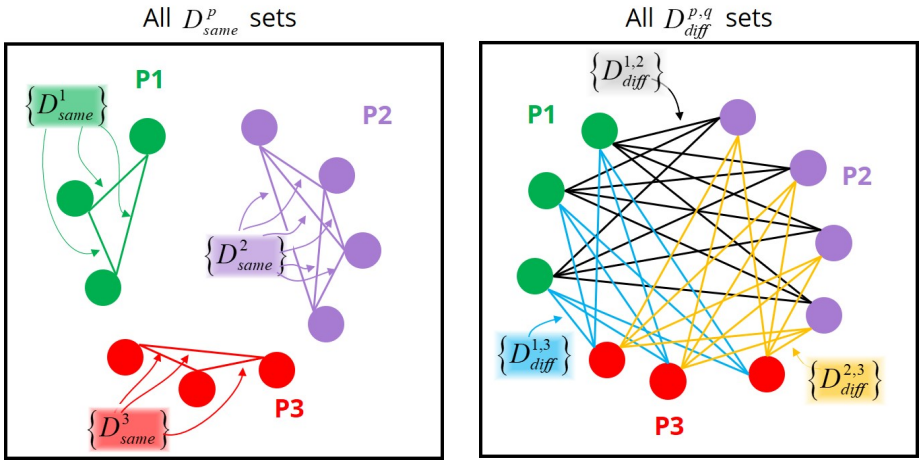


Figure 5.4: STEP 2 - illustration for $N_p = 3$. Dissimilarity-values are represented by lines, utterances by circles. Left: three D^{p}_{same} sets, each containing all dissimilarity-values of pairs with the same prosody. Right: three $D^{p,q}_{diff}$ sets, each containing all dissimilarity-values of pairs with different prosody. e.g., yellow set for the prosodies $P2$, $P3$

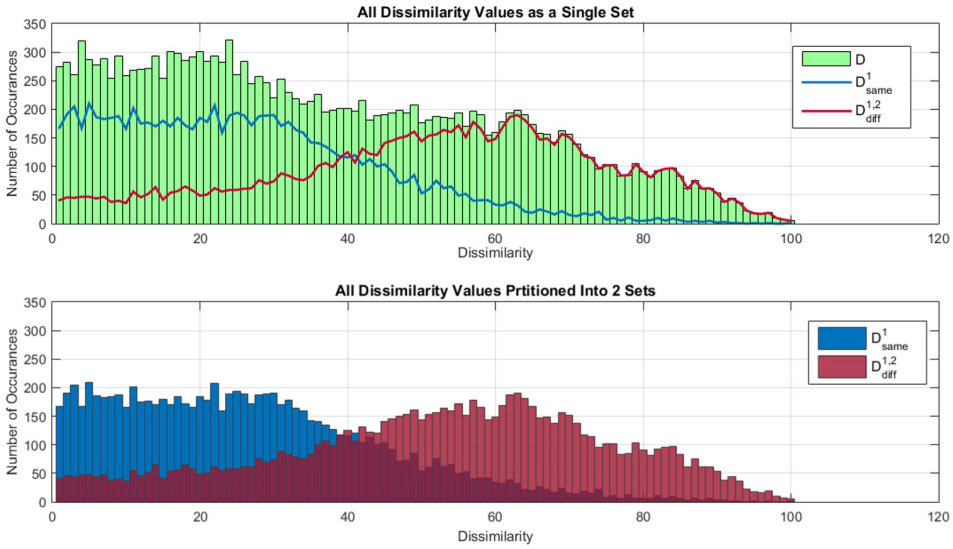


Figure 5.5: STEP 2 - example over real data of partitioning D set into two subsets. On top - histogram of the dissimilarity-values as a single set (D), and the contour of the partitioned two classes. On the bottom - two histograms of dissimilarity-values after partitioning into D^{1}_{same} and $D^{1,2}_{diff}$.

5.2.5 STEP 4: PMFs Dissimilarity

Calculate the dissimilarities between the PMFs of "same" sets (eq. 5.11) to the PMFs of "different" sets (eq. 5.12), denoted as:

$$v_{p,q} \triangleq DD\left(g^{p}_{same}, g^{p,q}_{diff}\right), p \neq q \tag{5.13}$$

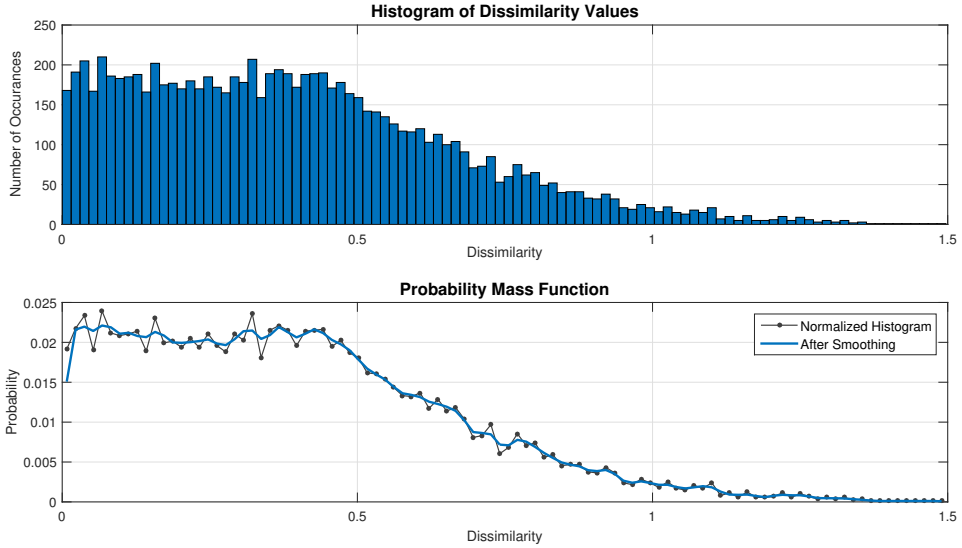


Figure 5.6: STEP 3 - PMF calculation over real data. Top: histogram of the values in set D^1_{same}. Bottom: the normalized histogram (i.e. PMF) in black, and the smoothed PMF in blue.

where the function $DD(\cdot,\cdot)$ measures the dissimilarity between distributions and therefore is called *Distribution Dissimilarity (DD)*. We also require it to satisfy the following: (1) non-negativity: $DD(x,y) \geq 0$, (2) $x = y \Rightarrow DD(x,y) = 0$, and (3) symmetry: $DD(x,y) = DD(y,x)$ for the same reason as in STEP 1.

Note that $v_{p,q} \neq v_{q,p}$ as they use different inputs. In total we have $N_p(N_p - 1)$ different elements. For example, in the binary case, we will have only two elements: $v_{1,2}$ and $v_{2,1}$. Figure 5.7 shows a synthetic example of calculating them. On left - the different PMFs we estimated in STEP 3. On right - the calculation of these two element. Each element has different inputs, and therefore they have different values.

Figure 5.8 shows another example of calculating $v_{1,2}$ and $v_{2,1}$ values for the binary case, over real data (F0_mean from the Initial feature set that was extracted out of the Hebrew Q&N dataset), using Helinger distance.

Motivation: the $v_{p,q}$ elements measure how much the examined feature F depends on prosody changes:

- As long as the distributions ("same" and "different") **get closer**, the relevant $v_{p,q}$ value decreases. Meaning that statistically, the feature's values do not change when

switching between prosodic classes p and q, i.e. the feature is **less dependent** on prosody.

- As long as the distributions **grow farther apart**, the relevant $v_{p,q}$ value increases. Meaning that statistically the feature's values change when switching between prosodic classes p and q, i.e. the feature's values **depend** on prosody.

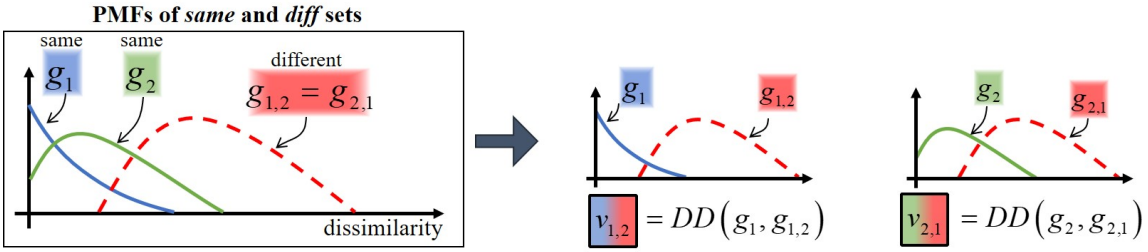


Figure 5.7: Illustration of STEP 4, for the binary case. Left- the PMFs of same prosody sets (g_1 and g_2) and different prosodies set ($g_{1,2}$). Right- the two outputs of $DD(\cdot,\cdot)$ function for that case: $v_{1,2}$ and $v_{2,1}$ - each has different inputs and different values.

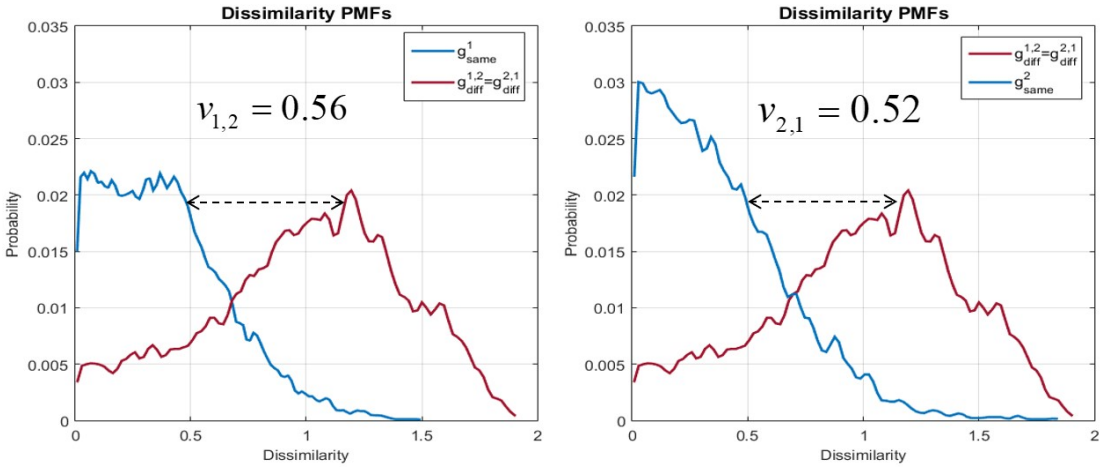


Figure 5.8: Illustrating of STEP 4 using real data. Left - the PMFs g^1_{same} and $g^{1,2}_{diff}$. Dissimilarity between these distributions is $v_{1,2}$. Right - same as left but for the value $v_{2,1}$.

5.2.6 STEP 5: PFC Score Calculation

We can arrange the outputs of the *DD* functions from STEP 4 in the following non-symmetric table:

$$T \triangleq \begin{bmatrix} / & v_{1,2} & v_{1,3} & \cdots \\ v_{2,1} & / & v_{2,3} & \\ v_{3,1} & v_{3,2} & / & \\ \vdots & & & \end{bmatrix} \tag{5.14}$$

The diagonal of the table do not contain any element. Therefore we have $N_p\,(N_p-1)$ different elements that represent the full relationship map between the different prosodic classes.
Each row of the table represents different prosody class, e.g., the first row represent prosody 1, as it measures how far is the distribution of the dissimilarity-values of prosody 1 (g^1_{same}) from the distribution of dissimilarity-values between prosodies 1 and q ($g^{1,q}_{diff}$ $2 \leq q \leq N_p$).

In order to have a single value which represents the PFC, we will now apply a function to all the values we calculated so far:

$$PFC \triangleq \Phi(T) \tag{5.15}$$

where the function $\Phi(\cdot)$ is any function that combines the table's elements into a single value. There are many ways to do so detailed in section 5.3.

An additional and optional step is to use a threshold over the PFC score to decide whether the examined feature F can be considered prosodic. This threshold reflects the amount of prosodic information a feature should carry in order to be called "prosodic feature."
This thresholding stage is not a must, as most of the time, we do not want to have a binary result - prosodic or not prosodic. We are usually looking for a continuous value, which shows how prosodic is a specific feature.

Let us summarize the PFC methodology's steps:

- **STEP 0: Features Extraction** - for each utterance extract the feature F values. $U \Rightarrow F$

- **STEP 1: Features Dissimilarity** - calculate dissimilarity between features values for all pairs of utterances. $F \Rightarrow D$

- **STEP 2: Partitioning to Subsets** - partitioning D set into subsets of "same" and "different" prosody. $D \Rightarrow D^{p}_{same}, D^{p,q}_{diff}$

- **STEP 3: Calculate PMFs** - for each subset, evaluate its PMF. $D^{p}_{same}, D^{p,q}_{diff} \Rightarrow g^{p}_{same}, g^{p,q}_{diff}$

- **STEP 4: PMFs Dissimilarity** - calculate dissimilarities between "same" and "different" PMFs $g^{p}_{same}, g^{p,q}_{diff} \Rightarrow v_{p,q}$

- **STEP 5: PFC Score Calculation** - arrange all of the $v_{p,q}$ values in one table and combine them to get a single PFC score. $v_{p,q} \Rightarrow T \Rightarrow PFC$

As mentioned before, the PFC gives us a quantitative way to compare the amounts of prosodic information each feature carries. It is important to note that when comparing PFC scores of a few features, we should use the same $DD(\cdot,\cdot)$ and $\Phi(\cdot)$ functions to calculate their PFC scores.

5.3 Suitable Functions for PFC Methodology

As a part of the PFC methodology, we use several functions. One of the strengths of the PFC is that it is a general workflow and does not depend on a specific function. The user who applies the PFC methodology can choose the functions that best fit the examined feature F. We will now elaborate the meaning and how should we choose each function:

5.3.1 Feature's Dissimilarity Function, $d_F(\cdot,\cdot)$ (STEP 1)

This is the basic dissimilarity function of the PFC methodology. As illustrated in figure 5.2, it measures how far are the feature's values of two different utterances.

The choice of $d_F(\cdot,\cdot)$ function is highly related to the examined feature. Different choices can be equivalent to testing a totally different feature. In other words, the final dissimilarity-values are always a combination of the $d_F(\cdot,\cdot)$ function and the feature itself. Meaning that the combination of feature A with function B can lead to the same values, such as the combination of feature C with function D.

An example: suppose we have some vector feature $\tilde{F}$. Let us define the following $d_F(\cdot,\cdot)$ functions:

$$d_F^1\left(F^i,F^j\right) \triangleq \frac{1}{D}\sum_{n=1}^{D}\left|F_n^i - F_n^j\right| \tag{5.16}$$

$$d_F^2\left(F^i,F^j\right) \triangleq \frac{1}{D}\sum_{n=1}^{D}\left|log\left(F_n^i\right) - log\left(F_n^j\right)\right| \tag{5.17}$$

where D is the vector feature's dimension.
Now we will define a new feature:

$$\bar{F} = log\left(\tilde{F}\right) \tag{5.18}$$

We can easily see that:

$$d_F^1\left(\bar{F^i},\bar{F^j}\right) = d_{\tilde{F}}^2\left(\tilde{F}^i,\tilde{F}^j\right) \tag{5.19}$$

There are many dissimilarities functions we can use. The considerations that should be taken into account when choosing the function to use, are related of the feature properties and requirements. As a general state - the function should reflect the feature.
Some examples:

- The speech signal is temporal and therefore there is a different semantic meaning for each element in the feature vector.
  An example: in question prosody, most of the speakers tend to raise their pitch at the end of the utterance. Therefore, for pitch feature, we may want to choose $d_F(\cdot,\cdot)$ function that would refer more significantly to the end of the utterance.

- The function should fit to the dimensionality and the space of the feature.
  An example: many features change their length according to the length of the utterance (e.g., MFCC, which is calculated per frame). Therefore, we may need to use a function that can compare vectors with different dimensions, e.g., Dynamic Time Warping (DTW) [20].

- Different features can have different scales what should affect the choice of the $d_F(\cdot,\cdot)$ function.
  An example: The F0 feature family which measure the pitch. We recall that the humans hearing and voice producing systems are working on a logarithmic scale, i.e., changing the pitch in one octave (increasing by a factor of 2) is perceived by the human ear as changing the pitch in one unit only. Therefore we may want to measure pitch related features using logarithmic dissimilarity function.

5.3.2 Distribution Dissimilarity Function, $DD(\cdot,\cdot)$ (STEP 4)

In this step, we measure the dissimilarity between two distributions. One is g_p - the dissimilarity between instances with same prosody p. The other distribution is $g_{p,q}$ - the dissimilarity between instances with different prosodies p and q, when $p \neq q$.

Measuring this dissimilarity can indicate how strong is the dependency between the feature's values to a change in prosody class. As long the distributions of "same" and "different" groups are more separable, there is a higher dependency on prosody change.

Many functions can measure how far are two distributions from each other and can serve as $DD(\cdot,\cdot)$ function. It is important to note that the choice of function affect dramatically the value of the PFC score.

In this work we use Helinger distance as the $DD(\cdot,\cdot)$ function because it is a simple probabilistic analog of the Euclidean distance that satisfy our three requirements. It is also bounded and does not require the same function's support which simplify the calculations even more.

For the discrete probability distributions $P = (p_1, ..., p_n)$ and $Q = (q_1, ..., q_n)$ where n is the vectors' dimension, the Helinger distance is defined as:

$$H(P,Q) = \frac{1}{\sqrt{2}} \sqrt{\sum_{i=1}^{n} (\sqrt{p_i} - \sqrt{q_i})^2} \tag{5.20}$$

5.3.3 PFC Score Function $\Phi(\cdot)$ (STEP 5)

In this step of the PFC methodology, we arrange the $v_{p,q}$ values in the prosodic table T which is the full relationship map between all prosody classes of the examined feature.
We want to combine the table's values to a single score, which would serve as the PFC. There are various functions we can use; each represents different requirements from a prosodic feature. Here are some possible examples:

- The maximum value in the prosodic table T, i.e., the best element of the table:

$$\Phi_1(T) = \max(T) \tag{5.21}$$

  Meaning that even if this feature can distinguish only between one pair of prosodic classes, we would consider it prosodic.
  An example: suppose we have three prosodic classes: anger (P1), panic (P2), and sadness (P3). The examined feature F can distinguish just between anger to sadness. The prosodic table for that scenario may look like this:

$$\mathbf{T^1} = \begin{array}{c} \begin{matrix} \text{P1} & \text{P2} & \text{P3} \end{matrix} \\ \begin{bmatrix} / & 0.23 & 0.65 \\ 0.15 & / & 0.33 \\ \mathbf{0.87} & 0.51 & / \end{bmatrix} \end{array} \begin{matrix} \\ \text{P1} \\ \text{P2} \\ \text{P3} \end{matrix}$$

  If we use this max function, we actually ignore the fact that the feature does not distinguish between prosodies $P1$ and $P2$ or between $P2$ and $P3$, so the PFC score is:

$$PFC = \Phi_1\left(T^1\right) = \max\left(T^1\right) = 0.87$$

- A weighted average over the off diagonal elements of the prosodic table:

$$\Phi_2(T) = \sum_{p \neq q} \alpha_{p,q} \cdot T_{p,q} \tag{5.22}$$

In this case, the score takes into account how well this feature distinguishes between all possible prosodic classes pairs. Using this function implies that when a feature distinguishes between more pairs of prosodies, its PFC score is higher.
An example: using the same scenario described in previous bullet, with three prosodies (anger, panic, and sadness) we will receive that $\Phi_2\left(T^1\right) \neq \Phi_1\left(T^1\right)$ because it takes into account all the off diagonal elements of the table.

5.4 Relations Between PFC and Feature Selection

The PFC was created to measure whether a feature carries prosodic information and to what degree. The output of the PFC is a numerical value that can be compared to other features (under some conditions that were mentioned before). Using this comparison, we can rank a set of features by their prosodic manifestation. One of the usages of that ranking method can be feature selection, i.e., if we want to reduce the number of features to use in a specific learning problem that is related to prosody, we can rank the full set of features by their PFC scores, and choose the features that received the highest values.

Even though it is possible to use the PFC as a feature selection method, initially, it was not developed for that purpose. The PFC methodology looks at the problem from the **feature** point of view, while taking into account the separation task between different prosodic classes.

The "feature point of view" means that we are looking at the feature's statistical properties. Even though the PFC calculates the dissimilarity between features values, implicitly it is affected by:

1. The distribution of the features values within each class. For example, in the binary classification case - when the standard deviation of features values within a specific class increases, PFC score decreases. In figure 5.9, you can see this example: PFC of A is smaller than PFC of B, despite the fact that the two classes are well separated.

2. How far are two distributions from each other. In figure 5.9 when calculating PFCs of B and C, we receive higher values for C than for B because these distributions are farther away.

These two examples of "the feature point of view," come from the definition of a prosodic feature - it should depend on prosody. When the standard deviation of a feature is larger and when two prosodic classes are closer to each other, it means that this feature is less dependent on prosody.

We can understand it better by analyzing two extreme cases of distribution of feature's values for some prosodic class P: (1) Uniform distribution - means that the feature is independent of prosody P . This is because it does not matter what the feature's values are; the probability that the prosody class is P is the same. (2) Dirac delta distribution - means that the feature totally depends on prosody P. This is because there is only a single value that indicates that the instance is related to prosody P. By having any other value, we can say that the prosodic class is different than P.

To summarize, PFC evaluates how well a feature separates between prosodic classes as well as the dispersion of the feature values within these classes. The PFC can definitely be used for feature selection as well.

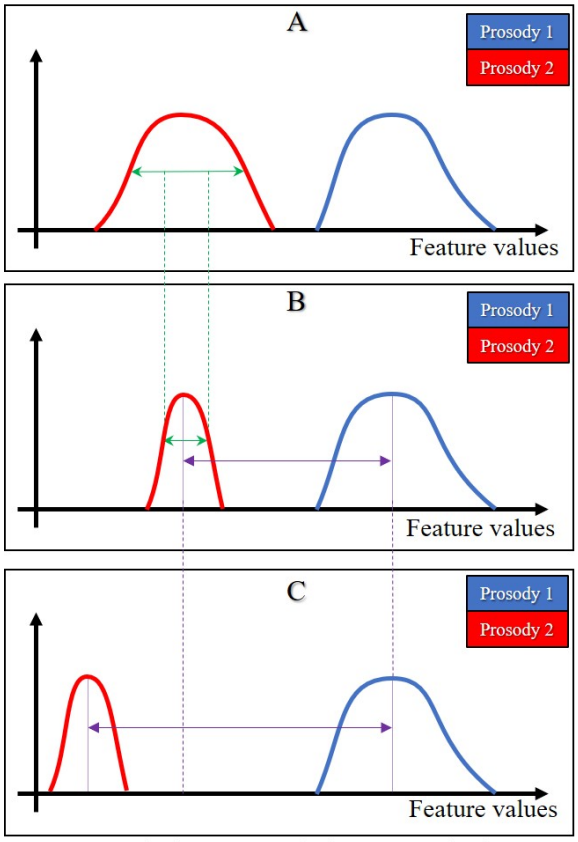


Figure 5.9: The relations between PFC to feature selection methods. In all these cases, feature selection methods would rank the feature relative high, because the two classes are separated. PFC also tests how far the instances of different classes are from each other. Therefore it will provide different scores for each of these cases: $PFC(A) < PFC(B) < PFC(C)$**.**

5.5 A Possible Extension of PFC

In STEP 4 of the PFC methodology we use $DD(\cdot,\cdot)$ function, in order to measure the dissimilarity between PMFs of "same" and "different" prosodies. We noticed that the distribution of "same" prosody dissimilarity (i.e., g^p_{same}) itself also contains information regarding the prosodic nature of the feature.
Therefore, we suggest to extend STEP 4 of the methodology by defining another function that measures this distribution:

$$s_p \triangleq SD(g^p_{same}) \tag{5.23}$$

$$S \triangleq \{s_p\} \forall p \tag{5.24}$$

where the set S contain the s_p values for all prosodies, and the function $SD(\cdot)$ measures some property about each g^p_{same} distribution. This property should measure how far different instances are within the same prosodic class p (and not with respect to other prosodies). It should provide additional information regarding the prosodic nature of this feature with respect to this prosodic class. Different choice of functions are possible, for example, the standard deviation or entropy.

This extension requires modifying STEP 5 as well, where we use $\Phi(\cdot)$ function to combine all of the $v_{p,q}$ values to a single PFC score. We suggest to use a different function that can also get the set S as an input and combine all these values together:

$$PFC^* \triangleq \Phi^*(T,S) \tag{5.25}$$

In that case, PFC^* might be a pair of values where the first represents the combination of $v_{p,q}$ values, and the other represents the combination of s_p values.

6 Datasets

We demonstrated the PFC applicability using two recognition tasks that are related to prosody: question detection and emotion recognition. For each task, we used a different dataset. We will start this chapter by explaining each task, and proceed to discuss a few general requirements a dataset should fit. Then we present the datasets we used in this work.

6.1 The Tasks That Were Used

6.1.1 Question Detection Task

Question detection is a basic, important, interesting, and one of the most researched tasks in the field of dialogue-acts detection[1]. As other dialogue-acts detection tasks, it is also an important step towards artificial systems and a better understanding of natural languages and speech [142], which can serve as a basis for human–computer dialog systems [99]. The use of question detection can improve multiple speech–related applications, such as (1) speech understanding where it can help model the speech structure [28] and sentence modality [73, 79, 88], (2) transcription where it can enrich transcription by punctuation marks [88, 125], and (3) speakers separation where it can provide useful clues for identifying a speakers role in a dialog [122].
Different applications can also use question detection as one of the components, for example: indexing and summarizing lectures or meetings [99, 79, 96], or systems which support communication with deaf and hearing impaired people [154, 73]. This task has been researched in many languages, e.g. French [73, 59, 162], English, German and Arabic [21].

Many works have noted the improvement that can be achieved by using prosody in addition or instead of lexical content [88]. Two examples of why lexical information can sometimes actually impair a question detection task: (1) some questions are called

[1] section 3.2.1 discusses dialogue-acts

declarative questions (e.g., "John is here?"). These kinds of questions share the same word order with its statement form. They may be distinguished as a question only by their prosody [32]. (2) In many spoken dialog systems, automatic speech recognition (ASR) is a preliminary step whose performance will have a significant impact on the following question detection steps [79]. Incorrect recognition of the lexical content during the ASR stage will lead to failure of the next steps. The use of prosody eliminates these problems.

In this work, we use the question detection task, as in addition to it's being a common, useful, and important task, it is also a basic, relatively well-defined task. For humans, it is quite easy to recognize a question in comparison to other dialogue-acts.

6.1.2 Emotion Recognition

Over the last few years, the recognition of emotions has become a multi-disciplinary research field which has received great interest [126]. People express emotions in a variety of ways, such as through body gestures, facial expressions [53, 126], changing of various biological signals like muscle activity and sweat [41], etc. Nevertheless, methods that try to measure some of these signals are sometimes invasive, complex, and cannot be used in certain real-life applications. This leads to a more feasible option - the use of speech [126]. Speech is another way to express emotions by human beings [141]. It can be expressed explicitly by the lexical context (e.g., "I'm angry with you") or implicitly through prosody. Therefore, speech signals can be analyzed to indicate the emotional state of the speaker - a task that is called *Audio Emotion Recognition* (AER).

AER is a current research topic with a very wide range [141]. This topic is important and can serve in many real-life applications. A few examples are: (1) human-computer interaction, where the ability of the system to sense the emotional state of the speaker in the dialogue is crucial [33, 126, 53, 144]. (2) Speech To Text systems, where the meaning of a sentence can be completely different when it is said using different emotions and cannot be understood only by the words [33], e.g., the expression "Yeah, right" has a positive literal meaning, but when said sarcastically has a negative meaning [76]. (3) Speech Synthesis systems, which include the usage of emotions for a more realistic output [33, 26]. (4) Call centers [70] like emergency services which can detect fear and stress, or customer service which manages customer requests.

The emotional state of a speaker has been widely studied in psychology, psycholinguistics, and speech [33], as also mentioned in 3.1.2. [33] shows that in addition to the lexical information in speech (the words), the emotional state is also encoded in the acoustic level [9, 23, 105, 84] and even more specifically by prosodic information [53, 14]. Many

additional works show the significant effect of emotions on prosodic parameters and the significance of using prosodic parameters in emotion recognition [126]. People can often evaluate the emotional condition of another person, only by listening to their voice [53], and not to the lexical content, e.g., understanding the emotion of a person speaking in a foreign language [80].

Even though speech contains a lot of information about the emotional state of the speaker, the task of recognizing and extracting this information is still very challenging [141]. Part of the reasons are that emotions have a complex nature, which is hard to model. It can be described by discrete labels [75] or continuous dimensions [4]. The number of emotions in human interactions is huge, e.g., [63] claims that it is infinite, subtle, and often mixed. Another challenge is data collection on which we will elaborate in section 6.3.2.

6.2 Dataset General Requirements

In data-related tasks, using the right dataset (either by choosing or collecting it), is challenging and has a significant effect on the system's results. The correct dataset composition is crucial for the algorithm's success. There are many considerations which should be taken into account; a few important ones are:

1. Size of dataset - it is important to have a large enough dataset, i.e., enough samples so that the sample group will correctly represent the population, and the results will be statistically reliable. The more dimensions or parameters we have in our model, the more data we need, otherwise our model could overfit the training set, or we could suffer from the curse-of-dimensionality (see Appendix A for more details).

2. Variability of the data - our dataset should not only be large enough, it also has to be diverse, i.e., the samples should be well distributed across the real population. This is to estimate the real distribution, so our results and conclusions are more accurate. The data should also be collected using various recording equipment, as different parameters (e.g., signal to noise ratio, frequency response, etc.) change the way the signal sounds.

3. Relevancy - the data should be relevant to the task and research question. This requirement sounds almost trivial, but in many cases, especially when it is hard to find or collect relevant data, we may have to use data that is only partially related to the task or research question. In this work, at first, we chose to collect our own dataset

as we could not find a dataset that has the required combination of different prosodic classes and different lexical content. Another example is emotion recognition sets, which many times consist of acted scenarios and do not always sound like natural emotions. This is because of the difficulty in collecting real-life speech utterances of different emotions.

In speech processing, data variety can be expressed through several aspects, such as: (a) the speaker: use speakers from different genders and various ages (as they have different speech properties). (b) The content: use different kinds of phrases and, in our case, different types of prosodies. Sometimes even different languages, length of utterances, etc. (c) Recording equipment and acoustic environments - use different microphones, room environments, etc. as they have different acoustic properties.

When focusing on prosody, the naturalness of the data becomes an issue. This is because when speakers are asked to talk in specific prosody, it is not trivial that they will succeed in doing so. Most of the speakers may exaggerate to "achieve" the requested prosody, in a way that will not sound natural to the listener (e.g., when trying to imitate happiness or sadness). Due to this issue, it is better to have a dataset that contains only natural speech. However, it is difficult and time consuming to collect a large enough dataset, in which a specific prosody class will also be large enough. It also requires a long pre-processing and tagging process.

In the next section we describe our two datasets. When collecting and choosing these datasets, we tried to take most of the above considerations into account.

6.3 The Datasets We Used

In this work we used two different datasets. The first is a self-collected dataset in the Hebrew language. The second is in English and was used by previous works. Our Hebrew dataset is smaller than the other one, as it was designed with specific requirements and therefore was harder to collect.
As mentioned in chapter 5, a prosodic feature should be dependent on prosody class and independent of other speech parameters such as content. For that reason, both our datasets include multiple content classes (i.e., different phrases) and prosody classes. This composition of the dataset makes it relevant for our task and can emphasize how the PFC behaves in different scenarios.

6.3.1 Hebrew Dataset - Question & Neutral Prosodies (Hebrew Q&N)

It is hard to find a dataset that exactly fits our research requirements, therefore we decided to begin our work with a simple, self-collected dataset.

The dataset is freely available for research purposes. It was recorded by non-actors in Hebrew and was designed specifically for prosody research. Because this was our initial dataset, we made it very simple, i.e., containing only one language, with two prosodic classes and short phrases with the same number of syllables.

We recorded three different phrases (content classes):

- /BO/ /LA/ /TSA/ /BIM/ = Come to the turtles
- /EN/ /KAN/ /BO/ /NIM/ = There are no beavers here
- /GAM/ /PO/ /AR /MON/ = There is a palace here too

For each phrase we recorded two prosody types: Question and Neutral. 36 speakers were recorded (males: 47%, females: 53%) of various ages (20-30: 22%, 30-40: 33%, 40-50: 8%, 50-60: 20%, 60-70: 17%). Each speaker uttered the same three short phrases, which consisted of four syllables each. All phrases were syntactically correct, and contained mostly voiced phonemes. Each phrase was recorded in two different prosodies (Neutral: 46%, Question: 54%). The data was recorded using various personal cellular phones, in a quiet room environment but not in a lab. In total there are 252 short phrases. Dataset details are presented in figure 6.1. In order to validate the data labels, two experienced listeners tagged all utterances in a random blind test. The manual tagging was 97% correct; therefore we consider the prosody labeling to be accurate.

We exclude a few variability parameters which are related to unique populations with different speech pattern and properties, which can affect the results. To the best of our knowledge, the speakers in our dataset are native speakers with a standard speaking style and do not include: (1) speakers with speech disorders, (2) speakers who suffer from medical conditions that affect their speech or prosody (e.g., Parkinson's Disease, Alzheimer's Disease, Autism Spectrum Disorder, respiratory diseases, etc.).

6.3.2 English Dataset - Emotional Prosodies (English Emotions)

Emotion recognition is a more complex task than question detection and therefore may be considered as the second level of PFC validation.

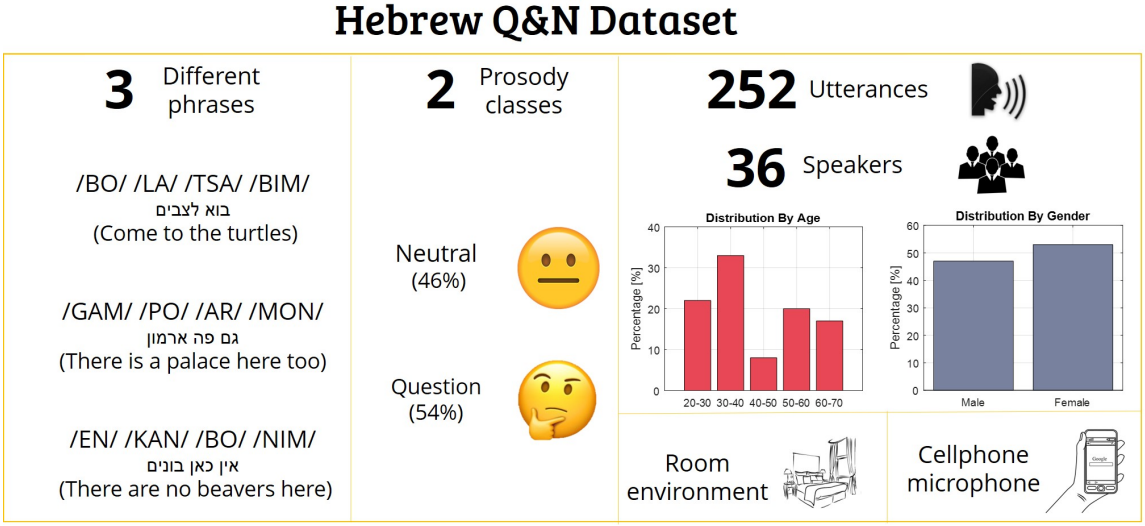


Figure 6.1: Specification of the Hebrew Q&N dataset.

Finding a proper dataset for this learning task is not easy. Most of the corpora are small, not diverse enough, and do not reflect real-life scenarios. Emotional speech is very different when it is acted in comparison to the natural, real-life scenario. The difference is so significant that lots of AI systems that were trained over artificial acted datasets fail in real life scenarios [43]. It is not only that real-life vs. acted emotions are different, but also that acted-emotion-recognition is a much easier task.
Many works deal with this issue and attempt to collect real-life data, which will be robust enough and include diverse speakers, acoustics, and emotional states classes.

Another difficulty is tagging or annotating the data in an accurate way. This task is even more challenging in real-life scenarios [124]. Although humans are usually quite good at recognizing emotions just by listening to a speaker, there are some emotions that contain very subtle nuances, making it difficult to differentiate between them, e.g., it is hard to distinguish between boredom and sadness.

Taking all of the above considerations, we used the LDC2002S28 - "Emotional Prosody Speech and Transcripts" corpus [164]. This dataset is in English, it was developed by Linguistic Data Consortium (LDC) and was recorded in 2000-2001.
It was designed to support research in emotional prosody, and it is relevant to our work, as prosody is one of the ways to expresses emotional states through speech (as discussed in section 6.1.2).

The speakers are 8 professional actors reading a series of dates and numbers, that are semantically neutral. Each utterance contains 4 syllables, they use 15 emotional cate-

gories, and in total there are 9 recorded hours. There are 5 Females and 3 Males, 7 of them in their mid-20s and 1 speaker is in his late-30s.

In our work we used two subsets of this dataset. For experiment 3, whose goal is to validate the PFC over different datasets, and not to change the number of classes, we used only two classes out of the full LDC dataset: Hot Anger and Neutral. For experiment 4, whose goal is to test the non-binary case of the PFC, we used five prosodies: Boredom, Sadness, Elation, Hot Anger, and Panic. The full explanations about the different experiments is presented in chapter 8. Figure 6.2 summarizes all the details about this dataset and the subsets that have been used.

The differences between this dataset and the Hebrew Q&N dataset are (1) different languages, (2) different speakers, (3) different prosody and content classes, (4) multiple prosody classes (as opposed to two classes).

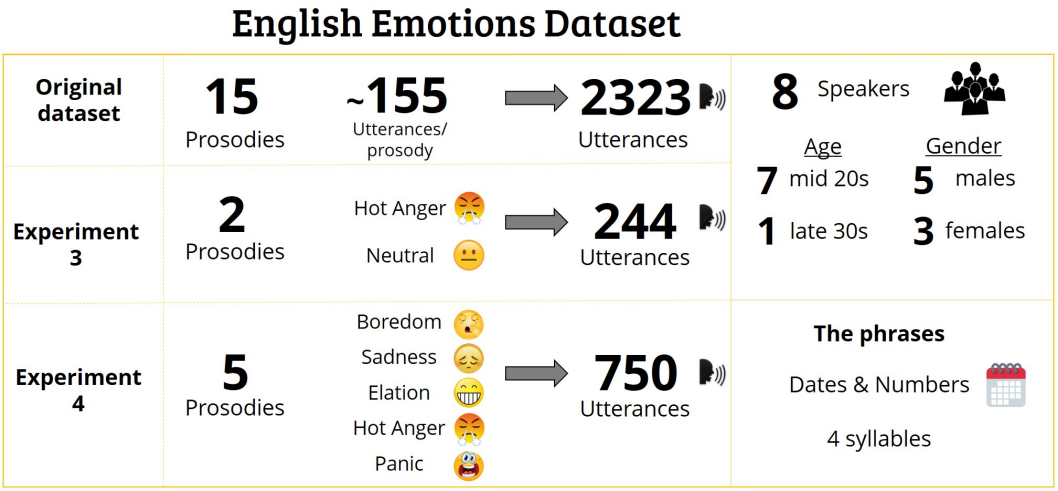


Figure 6.2: Specification of the English Emotions dataset.

7 Feature Sets

As we develop the PFC in order to measure features, we have to test and analyze its performance over multiple and diverse feature sets.
In this chapter, we explain some basic principals of speech features, review the most common features in the world of speech processing, and finally we describe the two feature sets we used.

7.1 Features for Speech Processing

A feature is a descriptor that represents some observed phenomenon within the signal and many times captures both long and short term phenomena. In signals that are sampled in a high-frequency sample rate, such as speech signals, the features are expected to be of lower dimension than the number of samples they represent. Features are measurable properties that are extracted from the signal; In most cases, they have numerical values. Some features are related to the concept of explanatory variables used in statistical techniques.

In the world of speech processing, a feature can be extracted in various granularities. The smallest one is called a "frame" - typically a short time window that does not have a semantic meaning. As the speech signal is quasi-stationary, we refer to each frame as if it is stationary. Most other common granularities have semantic meaning, and are usually associated with phonemes, syllables, words or sentences.
Features are sometimes influenced by the way humans produce and perceive audio and speech signals. For example, the feature *average-power* which is mathematically defined by: $\frac{1}{t_2-t_1}\int_{t_1}^{t_2}|x(t)|^2\,dt$, has also some linguistic description: "the average speech loudness between the times t_1 and t_2".
Features can also be associated with linguistic properties that are related to speech; for example, there are several features which can provide clues as to a certain sound being a consonant or a vowel, or to distinguish between sounds of different phonemes.

When focusing on prosody, many works show that people produce different prosodies by modifications in the spectral energy distribution [50], fundamental frequency, loudness, speaking rate, stress distribution [25, 42, 46, 50, 126] and more. Influenced by that, previous works use both acoustic and spectral features as features which are related to prosody.

7.1.1 Levels of Features

We can relate to two levels of features that can be extracted from a speech signal:

1. Lower Level Descriptors (LLD): this level is usually calculated directly over the raw speech signal, and in most cases is evaluated separately for each frame or for each block of frames. Such features can be F0, Energy, Voicing, etc. As speech is quasi-stationary over short periods, the length of each frame or block of frames should not be too long. On the other hand, it should not be too short in order to have enough samples to calculate the features reliably.

2. Functionals: this is the next level of features, calculated over the LLDs and constitutes higher-level descriptors, e.g., functions such as min, max, std, and more. These functionals can be calculated using segments of various lengths: arbitrary segments lengths such as frames, or semantic segments lengths by setting the start and end points according to phonemes, syllables, or words boundaries.

Features can be either LLD or functionals. For example, the standard deviation of F0 max values that were calculated over syllables, is a functional-based scalar feature that was calculated using the F0 LLD.
It is important to note that for each LLD we can extract a large number of possible features by applying many kinds of functionals. It is common to choose a subset of features to avoid issues like overfitting or the curse of dimensionality (see appendix A for further explanations).

7.1.2 Forced Alignment

In order to extract features over semantic units that are longer than the frame level (e.g. phonemes, syllables or words), we first need to find their boundaries (i.e., start and end points).
We do this using *Forced Alignment* procedure, which is the process of identifying which

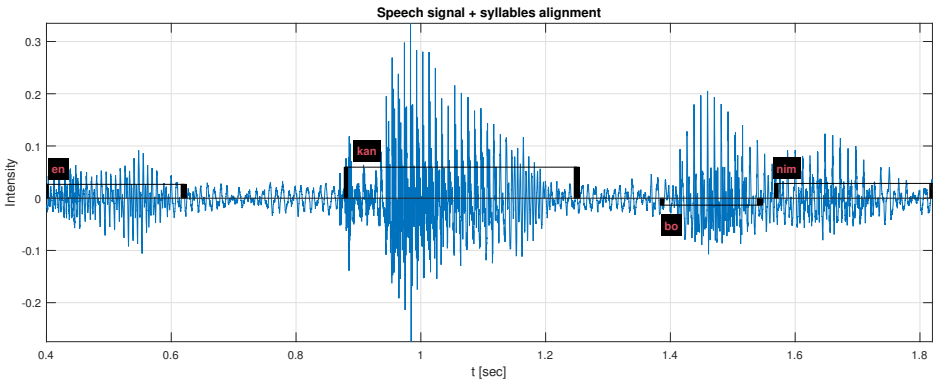


Figure 7.1: Speech signal taken from our Hebrew Q&N dataset, including the corresponding uttered text, aligned to the syllable level.

segments in a speech signal correspond to which part of an uttered text. The input to the system is a speech signal and the text that was uttered. The forced alignment algorithm finds the exact boundaries of each unit. This process can be done for different semantic units - phonemes, syllables, or words, as shown in figure 7.1. In this work, we used phoneme level forced alignment using acoustic models trained with the Kaldi engine [120], to produce syllable boundaries.

7.1.3 Common Features

Next, we describe some of the important and common features in speech processing; many of them have been used in this work:

F0

F0 is the fundamental frequency of a speech signal. In a periodic signal, the fundamental frequency is the inverse of the period, i.e., $F0 = \frac{1}{T0}$, where $F0$ is the fundamental frequency, and $T0$ is the period. The period is defined as the smallest positive number which we can shift the signal and it will remain the same, i.e. $x(t) = x(t+T0)\,\forall t$. Figure 7.2 shows a comparison between high and low F0.
A periodic signal is composed of multiple harmonies. We can present signals using the Fourier series, which decomposes the signal to its different harmonies and weighs every harmony differently. In Fourier series presentation of a signal, F0 is the lowest harmony.

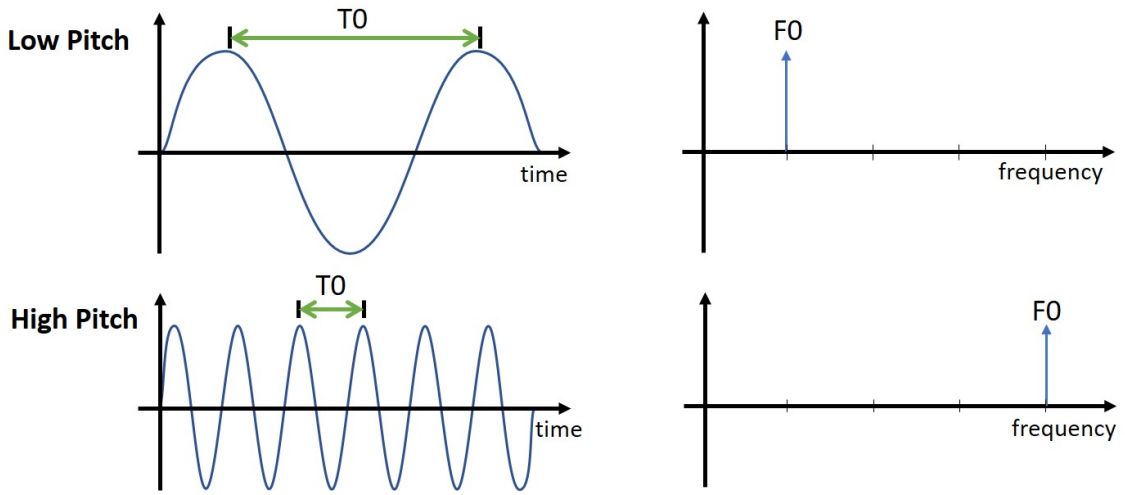


Figure 7.2: Illustration of F0 and pitch. Top - low frequency of the speech signal is perceived as low pitch to human ears. Bottom - high frequency is perceived as high pitch.

When dealing with a speech signal, we can define F0 as the rate of vocal folds vibration. The speech signal is quasi-stationary as the produced sound changes throughout the speech utterance; therefore F0 needs to be estimated over short segments. Voiced phonemes (all of the vowels and some consonants like 'z,' 'b,' etc.) are characterized by periodicity, so they have F0 values. Unvoiced phonemes (e.g., the sounds 'p', 't', 'k', 's') which are uttered without vibration of the vocal cords are not periodic at all, so they don't have F0 values. Due to the above properties of the speech signal, it is sometimes difficult to estimate F0.
Pitch Detection Algorithms (PDA) are the set of algorithms that try to estimate pitch/F0 contours. There are three common types of PDA methods which are based on time-domain analysis, frequency-domain analysis or a combination of both [159].
In this work, we extracted the F0 contour using auto-correlation which is one of the time-based methods [5], which is known to be robust to noise. In addition, we used two post-processing stages: (1) removed frames which are known as no-speech, using our forced alignment output. (2) Corrected F0 values on frames that "jumped" in octave multiples.

Pitch is a psycho-acoustic characteristic, which is the perceptual way humans understand F0. It is a subjective measure and can be perceived differently between people. F0, on the other hand, is a physical measure and therefore is objective. Even though the pitch is not exactly equivalent to F0, most of the time these terms have the same meaning. Humans can perceive sound as if it has "high" or "low" pitch, therefore tracking the changes of the pitch over time. This is also called *intonation* of the speech, and can have different

semantic meanings. Two examples for semantic meaning of pitch: (1) raised pitch at the end of a sentence is usually associated with a question, (2) a relatively high pitch is usually associated with a feminine voice.
Having said that, pitch and F0 are not equivalent, we should note that in most works (and in this thesis as well) it is conventional to use the term pitch to actually mean F0 and vice versa.
Figure 7.3 shows an example of F0 calculated from speech signals in our Hebrew Q&N dataset (see section 6.3.1) comparing Neutral and Question prosody.

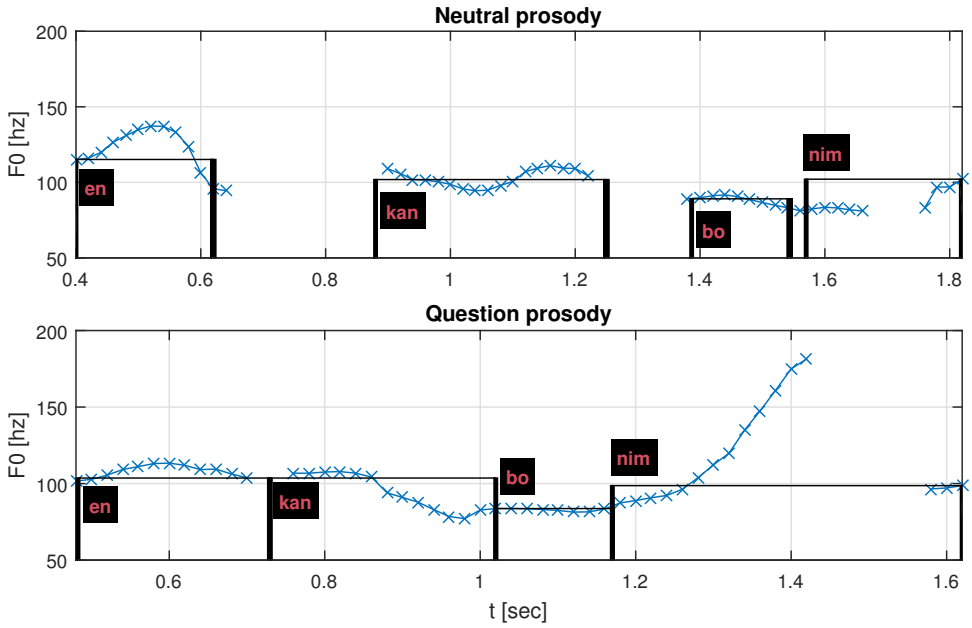


Figure 7.3: Fundamental Frequency - F0 of Neutral prosody (top) and Question prosody (bottom). In Question prosody, the pitch tends to rise at the end of the utterance.

MFCC

Another well-known and widely used features family is the Mel Frequency Cepstral Coefficient (MFCC). It extracts linear and non-linear features and captures important properties of the speech. It is used in tasks such as automatic speech and speaker recognition.

The MFCC features are based on the human auditory system, that perceives pitch by a unique and non-linear scale, according to psychological studies. The most popular approximation of that scale is called the Mel scale [52].

There is more than one Mel scale, most of them are split into two regions: linear below a certain point, usually around 1 [kHz], and logarithmic above this point. Eq. 7.1 shows one of the most common formulas to convert between Hertz and Mel units. Figure 7.4 shows the relations between the Hertz and Mel scale.

$$M(f) = 2595 \cdot \log_{10}(1 + f/700) \tag{7.1}$$

The MFCC is a variation of the cepstrum transformation [2]. The term cepstrum comes from the word spectrum when we inverse the letters "spec" to be "ceps." This term indicates it is the inverse transformation of the spectrum.

In practice, the real-cepstrum is the inverse Fourier Transform over the log Power-Spectrum of a signal, while the MFCC is the coefficients of the Discrete Cosine Transform (DCT) over log power of the signal for each filter. We show the MFCC calculation process as a block diagram in Figure 7.5, and describe it below:

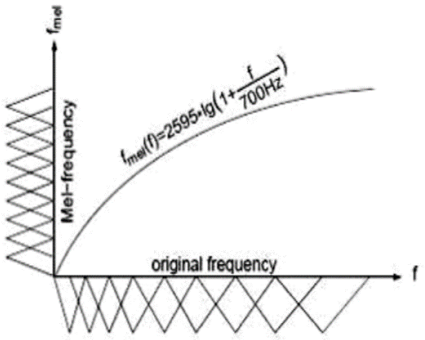


Figure 7.4: The relation between frequencies in linear Hertz scale and Mel scale. On the f axis we can also see the filter bank which is used for the conversion (taken from [117]).

1. **Framing + Windowing**: divide the signal into short frames (usually between 20-40 [ms]). Apply window filter over each frame (usually by Hamming window), in order to decrease discontinuities of the signal on the frames' edges.

2. **FFT/ DFT**: take the Fourier Transform of each frame of the signal, in order to convert time domain into frequency domain. We apply square magnitude to get the Power Spectrum.

3. **Mel Filter-Bank**: change the scale from linear to Mel scale, usually by using filter banks. One triangular band-pass filter for each Mel frequency as shown in figure 7.4. Now we have the power of the signal at the output of each filter.

4. **Log**: take the log of each Mel scaled frame. Now we have the log power of the signal for each filter.

Figure 7.5: MFCC calculation process – framing the signal and windowing each frame, then converting the frames from time to frequency domain using FFT/DFT. After that, converting it from linear to Mel scale and applying the logarithm. Finally, applying DCT and take its 13 coefficients as the MFCC.

5. **DCT**: convert each Log-Mel-Spectrum back to the time domain using the Discrete Cosine Transform (DCT) and use the transform coefficient as the MFCC vector. In most cases 13 coefficients are kept.

Energy

In signal processing, the energy of a signal can be defined as the area under the squared amplitude of the signal (see eq. 7.2). The humans' perception of the sound loudness, affected by the pressure of the sound wave, is related to the energy of the signal. Therefore, when dealing with a speech signal, we can use the energy as a feature, as it is associated with some meaning expressed by the speech. A few examples for usages are: (1) distinguish between different emotions, e.g., sadness or boredom will usually be quieter than anger or happiness, (2) different meanings of a sentence with the same word order, where emphasis on some words is expressed also by increasing the loudness for these words.

We should note that as the sound wave decays in space, the distance between the wave source (i.e., the speaker's mouth) and the microphone can dramatically change the energy values, therefore different recordings should be done under the same conditions.
Energy is an LLD level of a feature, so we can apply more transformation over it, e.g., use log-energy instead of energy.

$$E = \int_{-\infty}^{\infty} |x(t)|^2 \, \mathrm{d}t \tag{7.2}$$

Duration

Duration features can be associated with rhythm, which is the rhythmic pattern of the sentence (similar to rhythmic patterns in music such as 4/4 or 7/8). These features are also associated with the tempo, which indicates how fast or slow the sentence is. We can measure these duration features over different units of the sentence, i.e., phoneme,

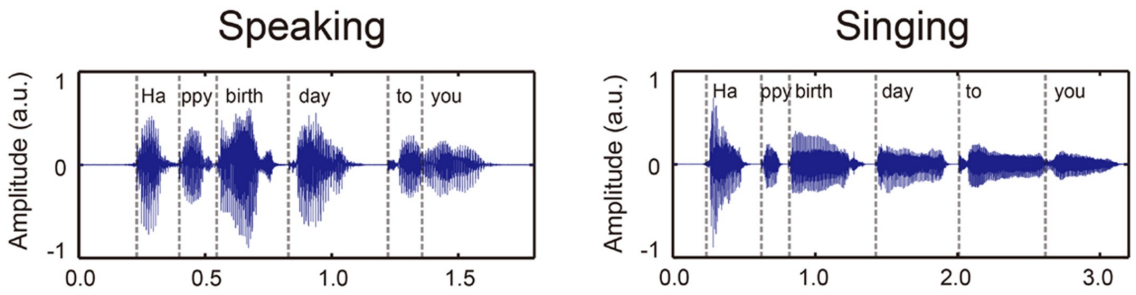


Figure 7.6: The rhythm and tempo of the voice changes between speaking and singing the same words (taken from [137]).

syllable, word, or sentence level. Duration features can indicate speech properties and can be seen in figure 7.6. Other examples are: (1) different emotional states may use different rhythm or tempo, (2) some dialects and cultures usually have their own unique tempo and rhythm.
We should note that some words have different duration than others, simply because they have a different number of phonemes. This fact itself is not related directly to prosody. It is the relative difference between uttering the same units with different duration, which is what really matters for our prosodic analysis. Therefore we usually analyze the relative duration of a segment in comparison to the standard duration most of the people uttered it.

Jitter & Shimmer

Jitter and Shimmer are two features that measure temporal variations of the speech signal. Jitter measures F0 variations, and Shimmer measures amplitude variations. These features can be considered as prosodic, as they deal with F0 and amplitude (related to the energy of the signal). They are widely used in various types of tasks, such as evaluation of pathological voice quality [72]. That is because when a person utters a long sustained vowel, and there is significant variation in pitch or amplitude (measured by Jitter and Shimmer), it is considered to indicate some pathology. Other persons can perceive this pathological sound as breathiness, roughness or hoarse voices.
In addition, Jitter and Shimmer can be used for other tasks like speaker verification [91], speaking style [93] and to classify genders, tones, and vowels [136].

There are several definitions for these features but the basic ones are:
(1) Jitter (see equation 7.3) is the cycle-to-cycle variation of F0, i.e. "the average absolute

difference between consecutive periods" [123].

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1}|T_i - T_{i+1}| \tag{7.3}$$

where T_i are the $T0$ period lengths and N is the number of periods.

(2) Shimmer [dB] (see equation 7.4) is the variability of the peak-to-peak amplitude in decibels, i.e. "the average absolute base-10 logarithm of the difference between the amplitudes of consecutive periods, multiplied by 20" [123]. Illustration can be seen in figure 7.7.

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1}|20\log(A_{i+1}/A_i)| \tag{7.4}$$

where A_i are the extracted peak-to-peak amplitude data and N is the number of extracted fundamental frequency periods.

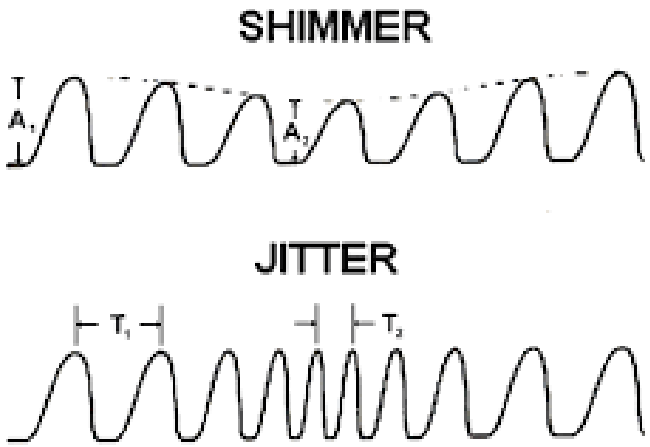


Figure 7.7: Illustration of how to calculate Jitter (cycle to cycle F0 variation) and Shimmer (peak to peak amplitude variability in dB), image taken from [169].

7.2 Feature Sets Used

We used two types of feature sets. The first was an initial, small, self-implemented standard feature set, while the second was an open-source toolkit for feature extraction, called OpenSMILE.

7.2.1 Initial Feature Set

To demonstrate and test our methodology, we reviewed many works in speech processing while focusing on prosody research looking for the most common features in use. At first, we wanted to prove our concept, so we chose a small subset of 48 features and implemented them.

Most of these features are considered standard in prosodic research, e.g., F0 and its derivatives. Some of the features are hand-crafted, such as duration-tilt and amplitude-tilt [97]. In addition, we used some features that are not considered prosodic, e.g., MFCC, in order to examine the difference between features that are considered to carry prosodic information and features that are usually related to other aspects of the speech signal.

We extracted the features in three types of "segments length":

1. "Per-frame" - evaluated over single frames, yielding a vector whose length is the number of frames in the utterance.

2. "Per-syllable" - evaluated over a single syllable. For example, F0_max of the x syllable is the maximum value of the F0 LLD over the segment length of the x^{th} syllable. Utterance with s syllables will have s F0_max instances.

3. "Accumulated" - evaluated over a segment starting at the beginning of the syllable and ending at the end of the utterance. For example, F0_max$^{\text{a}}$ is the maximum value of the F0 LLD over the segment, starting at the beginning of syllable x and ending at the end of the utterance. The number of instances is identical to the number of syllables.

All features are listed in table 7.1, grouped by their LLD, functionals and their "segment length". Each row also shows the number of features derived. For example the first row has 3 LLDs (F0, dF0 and Energy), 4 functionals (min, max, mean, var) and 2 segment lengths (per-syllable, accumulated), so in total there are $3x4x2 = 24$ features.

7.2.2 OpenSMILE Feature Set

In order to extend the PFC validation, we decided to use a broader feature set. We chose OpenSMILE [108], which is an open-source toolkit for extracting many types of acoustic and spectral features. OpenSMILE can be used either offline or online. This tool is widely

LLD	Functionals	Segments Length	# features
F0, dF0, Energy	max, min, mean, var	per-syllable, accumulated	24
F0, dF0	max-range		4
F0	peak-position, ampTilt, durTilt		6
MFCC (1-13)	–	per-frame	13
Duration	–	per-syllable	1
		Total	48

Table 7.1: List of Initial feature set. Including the LLD, functional and segment-length that were used to extract each feature.

used and has been cited over 1,300 times, mainly in the areas of speech recognition, emotion recognition, affective computing, and music information retrieval. The OpenSMILE serves as a baseline acoustic feature set in many competitions, for example AVEC 2013 challenge [131] or at Interspeech challenges like: 2009 emotion challenges [103], the 2010 paralinguistic challenge [Schuller u.a.)Schuller u.a.], the 2011 speaker state challenge [119], etc.

In this work we focus on the 2011 speaker state feature set (*IS11_speaker_state.conf*). This challenge had two sub-tasks, including the classification of "Alcohol Language" and "Sleepy Language." These two classes are distinguished from regular speech, especially by the prosody, and this is why we chose this specific subset. The feature set includes 4,368 features composed of LLDs (Energy, Spectra, voice–related, etc.) and functionals applied over them. We chose this feature set configuration as it is large and widely used.

8 Experiments

In this chapter, we show a few of the experiments we performed during the PFC development process. The main goal is to ascertain that our criterion requirements make sense and are consistent with prior knowledge about the prosodic nature of the examined features.
In addition, we visualize interim results from the process of the PFC methodology as described in chapter 5.2.

We recall that according to the criterion, a prosodic feature should be dependent on prosody but independent of other speech parameters. In this work we refer to the speech signal as if it has only two components, one is prosody and the other is the lexical content. Therefore, in each visualization we show the feature's instances (values of the feature per frame/syllable, etc.), using two different graphs: one shows the data by its prosodic classes, and the other shows it by its content classes.

We use two types of features: those that are considered prosodic and those not considered to be prosodic. We also use two datasets: Hebrew - Q&N and English - Emotions.

8.1 Experiments Specification

In order to validate the results and the reliability of the criterion, we run a few experiments and show that even when changing the dataset, feature set or the number of classes, the criterion remains valid.
Our first experiment is a combination of a simple dataset and feature set. Then, we gradually complicate the experiments. In each experiment, we test only one aspect or parameter (dataset, feature set, etc.), by setting all parameters to be the same as in the previous experiment, and changing only one parameter.
Further descriptions of the datsets and feature sets can be found in chapters 6 and 7 respectively.
The experiments settings as follows, also shown in figure 8.1.

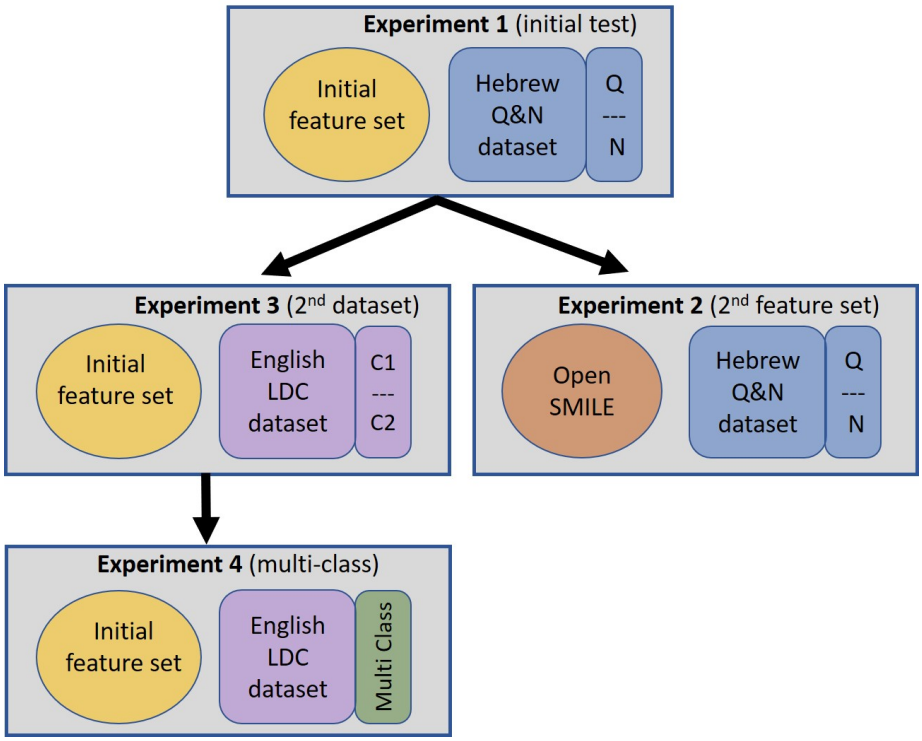


Figure 8.1: Description of our four experiments. We started with an initial experiment (1), then in order to validate our results, we changed only one aspect in each experiment: the feature set (2), the dataset (3), or the number of tested classes (4).

- **Experiment 1 (initial test)**: this is the proof of concept stage, so we chose to start with a relatively simple problem. We used the Hebrew Q&N dataset (see section 6.3.1) and the Initial feature set (see section 7.2.1). The dataset is diverse but not very big. It includes only two prosody classes - Neutral and Question, that are well separated. In addition, we controlled the recording protocol that was designed according to our needs.

  The features are a group of basic features-families for prosody research. Previous works gave us a good idea of which of them are considered to be prosodic in nature and which are not.

- **Experiment 2 (second feature set)**: this is a validation of the feature set. We kept the same dataset (Hebrew Q&N) but changed the feature set to OpenSMILE (section 7.2.2). That way we: (1) tested the PFC over many more features (~4300 features) thus increasing the reliability of the test. (2) Used a well-known, widely used, and reliable feature set that helped us verify our findings. (3) Verified results for features in the Initial feature set implementation as some features were the same in the two

feature sets.

- **Experiment 3 (second dataset)**: We kept the same feature set from experiment 1 (Initial feature set) and changed the dataset to be English - Emotions dataset 6.3.2. We use only two emotional classes out of many classes in this dataset, so this is still a binary problem. The new dataset is in a different language, with other speakers and new prosodic classes. It is also well-known and previously used by other papers, and we did not have any influence on the recording protocol. All these make this dataset reliable and fit for another validation of the PFC.

- **Experiment 4 (multi-class)**: in this experiment, we extended the last scenario and changed the task to include multi classes. We set the feature set and the dataset to be the ones used in experiment 3, but expanded the dataset to includes utterances with other classes as well, so in total, we had five prosodic classes.

8.2 Distributions Analysis

One of the basic analysis that can be done to test the prosodic nature of a feature, is to analyze its values' distribution when partitioning the data by both prosodic and content classes. This is not part of the PFC but it helped us in the development process.
In this analysis we show a Probability Mass Function (PMF) of a feature while not taking into account temporal information. For example, if a certain feature is calculated per syllable, we will have S different instances for each speech utterance with S syllables.
The process of creating the PMF is: (1) extract feature values for each utterance. (2) Split these values by their class (either prosodic or content class). (3) Calculate the histogram of each class separately and normalize the histogram's values of each class by the number of data points in the class, to get the PMF.

Example for the PMF of the feature F0_mean_per_syllable (mean of F0 values that were evaluated along a syllable) can be seen in figure 8.2. This feature's values were calculated over the Hebrew dataset.
This feature seems to be prosodic, as PMF values are significantly different for the prosodic classes P1 and P2, and separable. When looking at the same feature's values, partitioned by their content class, the PMFs look very similar, i.e., the distributions of the three content classes C1, C2 and C3 are not separable at all.
Indeed, previous works also show that F0_mean is considered to convey prosodic information [128].

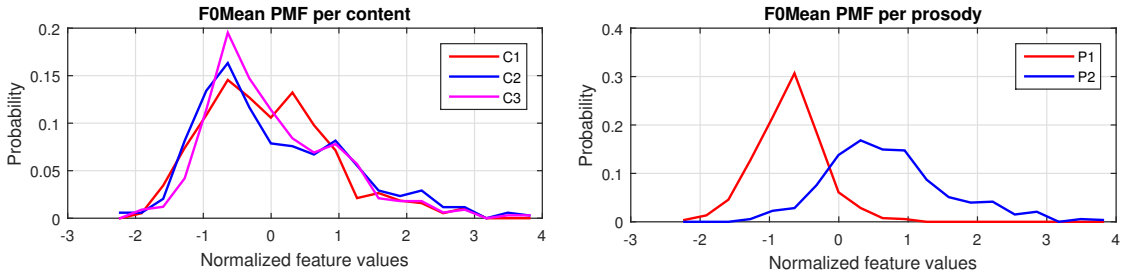


Figure 8.2: PMFs of F0_mean feature. Right - good separation between prosody classes. Left - no separation between content classes.

By using the same analysis over the 8^{th} entry of the MFCC vector, we can see in figure 8.3 a different behavior. The distributions of the feature's values when colored by prosody are not separable at all, while there is a slight separation between the class C2 to the other content classes. Using this analysis, we say that MFCC8 does not seem like a prosodic feature and indeed, it is known from previous works that MFCC features are not considered to represent significant changes in prosody [128].

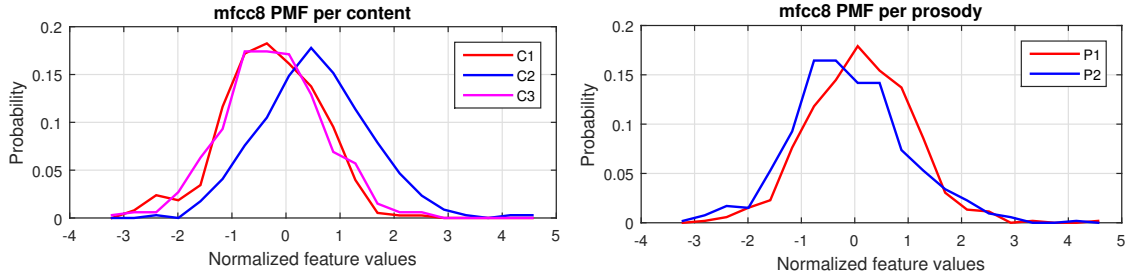
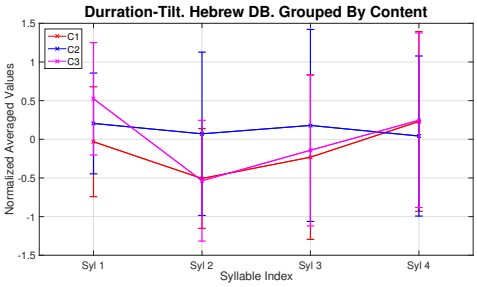


Figure 8.3: PMFs of MFCC8 feature. Right - no separation between prosody classes . Left - some separation between content classes.

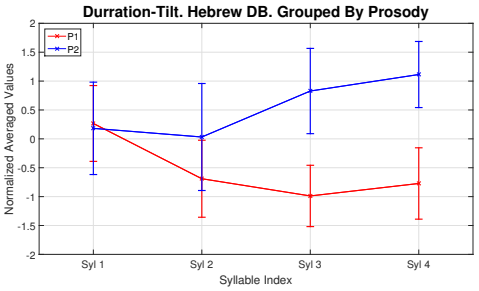
8.3 Temporal Analysis

As we are dealing with speech utterances, the nature of the signal is that it changes over time. Sometimes the essential cue indicating the utterance's prosodic class is related to the temporal information. For example, in most cases of question prosody, the pitch will rise at the end of the utterance. Neutral prosody in contrary usually keeps the same pitch value for the whole utterance.

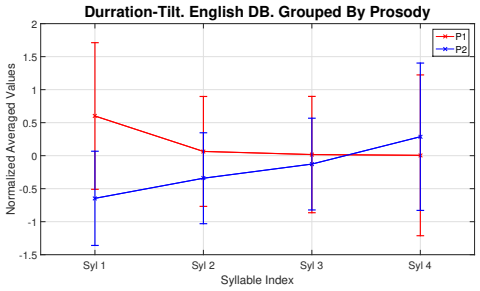
For that reason, our current analysis examines the dynamics of a feature's values over time.



(a) The feature Duration-Tilt does not separate between different **content** classes in the Hebrew dataset.



(b) The same feature as in figure 8.4a, separates the two **prosodic** classes of the Hebrew dataset. The feature depends on prosody but is independent of content (figure 8.4a), therefore it may be considered prosodic.



(c) The same feature that separates prosodic classes in the Hebrew dataset (figure 8.4b), does not separate two prosodic classes in the English dataset.

Figure 8.4: Temporal visualization of the average values and standard deviation of two features over both datasets.

We first extract features for all utterances; then, we split the feature's values by the utterance's class (either prosodic or content classes). We then calculate the average and standard deviation of the feature's values per syllable for each class separately.

In figure 8.4 we can see a temporal analysis of the features Duration-Tilt and F0_min. Sub-figure 8.4b shows that the Duration-Tilt feature can be considered prosodic, since we can see that the values of the features (both average and STD) distinguish between Question and Neutral prosodies. As we progress towards the end of the utterance, the differences between these classes become larger.
On the other hand, in sub-figure 8.4a we can see that this feature does not separate well between the different content classes, as the averaged values of the three different classes are almost the same for the whole utterance.

Sub-figure 8.4c shows the analysis of the same feature over the English - Emotions

dataset with the prosodies Neutral and Hot Anger. In this case, the feature does not show prosodic manifestation in regards to these two prosodies, as it does not distinguish between the two classes.

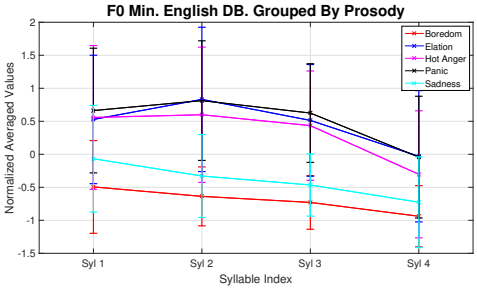


Figure 8.5: The feature F0_min splits five prosodies in the English dataset into two groups: (1) Boredom and Sadness, (2) Elation, Anger, and Panic. Therefore, it is considered as carrying some prosodic information.

Experiment 4 is more complicated and contains multiple prosodic classes. In figure 8.5, we show an analysis of this experiment using the feature F0_min. We can see that the feature can distinguish between two groups of features: (1) Boredom and Sadness, and (2) Elation, Anger and Panic.

The fact that a single feature does not distinguish between all of the five classes is totally reasonable. In fact, most of the time we need more than a single feature to distinguish even between two classes.

The reason that we show this analysis is to stress that there are a few levels of prosodic information a feature can carry. For example, if we were to examine another feature that would split these five classes into more than two groups, it might be considered as one who carries more prosodic information than F0_min regarding these five classes.

Recall that in STEP 5 of the PFC methodology we combine the table values into a single PFC score. It can be done in many ways, such as: (1) taking the highest value in the table, (2) taking the average. Therefore, even if a feature only distinguishes between two prosodic classes, its PFC score may still be high, especially when using option (1).

8.4 PMFs Dissimilarity & T table (STEP 4 + STEP 5)

8.4.1 Experiment 1

In experiment 1, we tested the Hebrew Q&N dataset, which has two prosodic classes and therefore we used the binary case PFC.

Figure 8.6 shows the PMFs of dissimilarity values of the Duration-Tilt feature (see chapter 7).

On the left side, we can see good separation between the sets of "same" and "different" prosody. It means that statistically, pairs of utterances with the same prosody have lower dissimilarity values than pairs of utterances with different prosody. In other words, we can say that statistically, utterances with different prosody are "farther" from each other and utterances with the same prosody are "closer" to each other.

On the right side we can see the opposite behavior for the **content** analysis. To create this figure we performed the similar process, but compared "same" and "different" content classes, instead of prosodic classes.
In the figure we can not see separation between the dissimilarity's PMFs for both "same" and "different" content classes. That means this feature responds in a similar way to "same" and "different" content, so it does not carry content/lexical information.

To summarize, this feature satisfies our two requirements of a prosodic feature, as it is (1) **dependent on prosody** changes ("same" prosody and "different" prosody pairs have different dissimilarity values), and (2) **independent of content** changes ("same" content and "different" content pairs have statistically similar values).

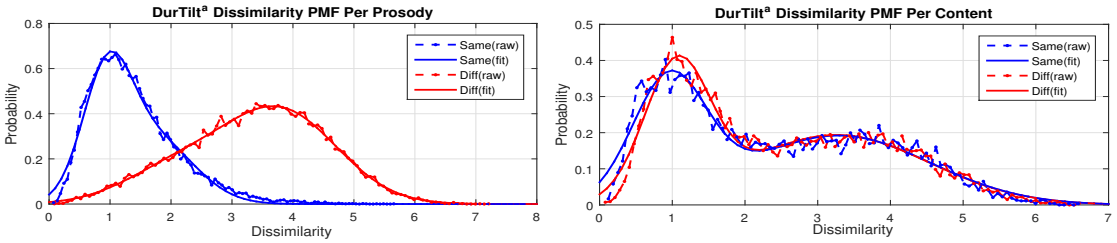


Figure 8.6: PMFs of dissimilarity values of the feature Duration-tilt. Left - same and different prosodies. Right - same and different content.

8.4.2 Experiment 4

In experiment 4, we used the PFC over the English - Emotions dataset, and over five prosodic classes out of the full set. We tested the features of the Initial feature set.
Figure 8.7 shows the final T table (STEP 5) of F0-Min.
We can see that the table has high values on some of the cells, e.g., the cells (1,2) - (1,4) and (2,1) - (4,1). On the other hand, there are cells with a very low score, e.g., the cells in the center (2,2) - (4,4).

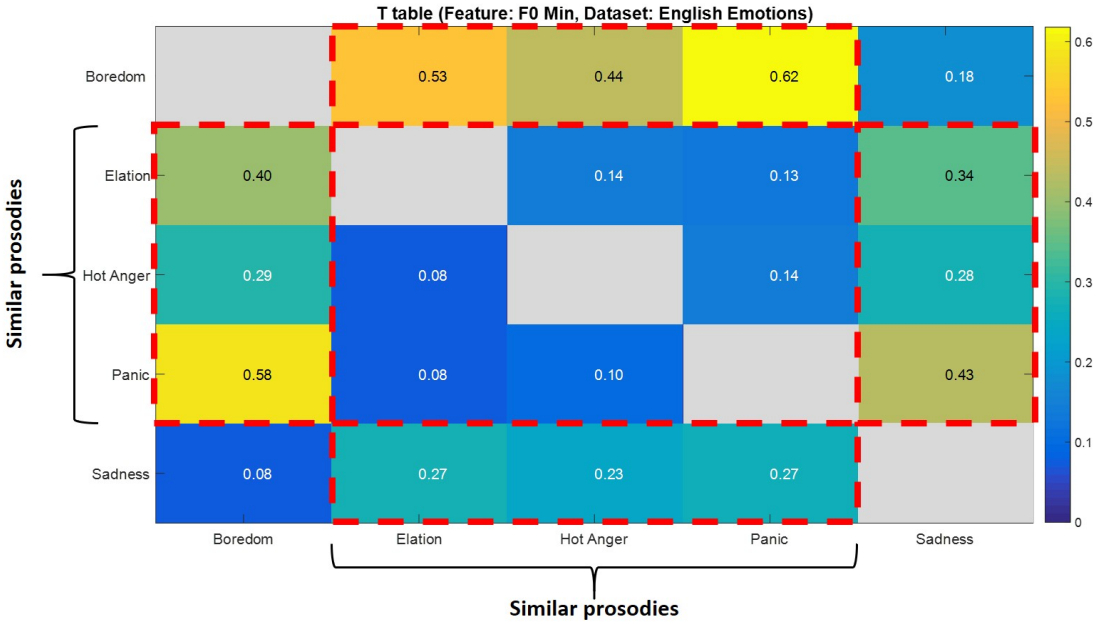


Figure 8.7: T table of F0_min for the English dataset. High scores (cells in red) between the two groups of prosody: (1) Boredom and Sadness, (2) Elation, Hot Anger and Panic. The rest of the cells have lower scores, reflecting that this feature does not distinguish between all prosodies within this set.

If we compare between cell (1,2) with the value $v_{1,2} = 0.53$ and cell (3,4) with the value $v_{3,4} = 0.14$, we conclude that this feature better separates between the prosodies Boredom and Elation, than between Hot Anger and Panic.

When looking globally at the table, we can see a pattern. The five prosodic classes are split into two groups: the first contains the emotions Elation, Hot Anger, and Panic. The second group includes the emotions Sadness and Boredom.
This is not surprising, as even humans perceive these two groups differently. It is obvious that distinguishing between Panic and Boredom is easier than between Boredom and Sadness.
This analysis shows that the PFC succeeded to analyze which classes are separable by this feature and which are not.

We can see that the dissimilarity between g^{1}_{same} to $g^{1,5}_{diff}$ (i.e., $v_{1,5}$ value) is much smaller than the dissimilarity between g^{1}_{same} to $g^{1,4}_{diff}$ (i.e., $v_{1,4}$ value) and it is reflected in the the cells values (1,4) and (1,5).

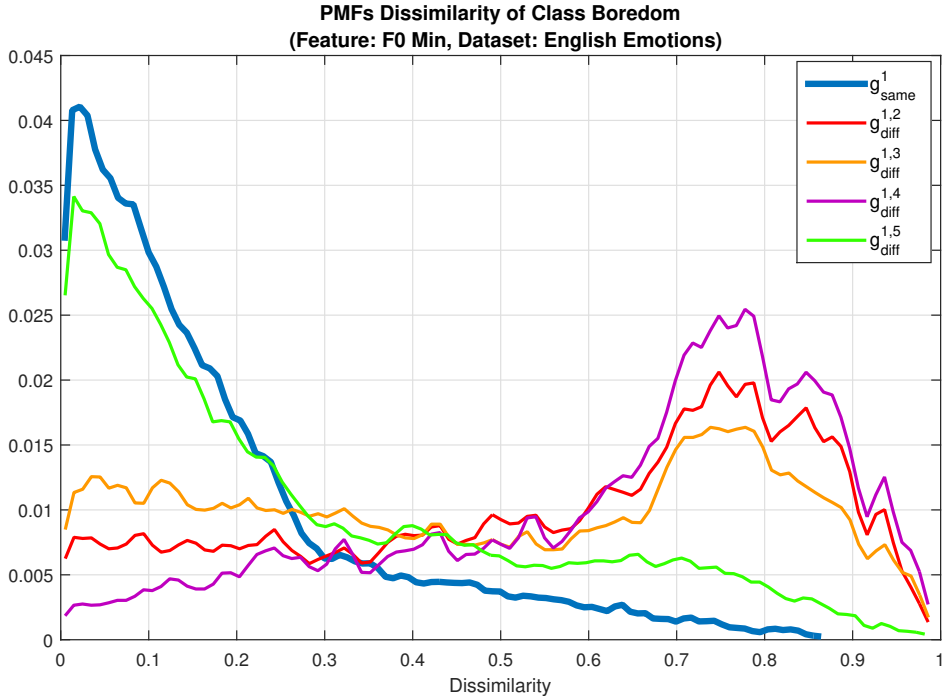


Figure 8.8: PMFs of dissimilarity values of F0_min feature in the English dataset. The same partitioning of these five classes into two groups, as in figure 8.7.

9 Data Analysis and Results

In order to validate our criterion, we use several methods to observe and analyze the PFC results. The following sections will elaborate on each validation method.

9.1 Validation 1: Comparing Between Features' Families

One way to verify the PFC results is to compare it to previous knowledge in the field. As we explained in chapter 4, even though there is no common definition for prosodic features, there are features that are considered to carry prosodic information. We will use this knowledge as our ground truth.

The comparison process includes the following steps: (1) calculate PFC for each feature in the feature set, (2) group features into features-families, (3) sort features-families in descending order by their PFC scores, (4) sort each family members in descending order by PFC scores.

Grouping into feature-families was done according to the features' LLD category (F0, MFCC, etc.). For example, the features F0_max, F0_mean, F0_amplitude_tilt, and F0_duration_tilt are all part of the F0 family.

9.1.1 Experiment 1: Initial Feature Set + Hebrew Q&N Dataset

Figure 9.1 shows results of the Initial feature set (section 7.2.1) for our Hebrew Q&N dataset (section 6.3.1). A few conclusions from this figure:

- The F0 and F0-derivative families are the highest scored families. This fact makes sense as: (1) these features are considered to be prosodic [69, 128] and (2) our two prosodic classes are Neutral and Question, known to be distinguished by F0 [99].

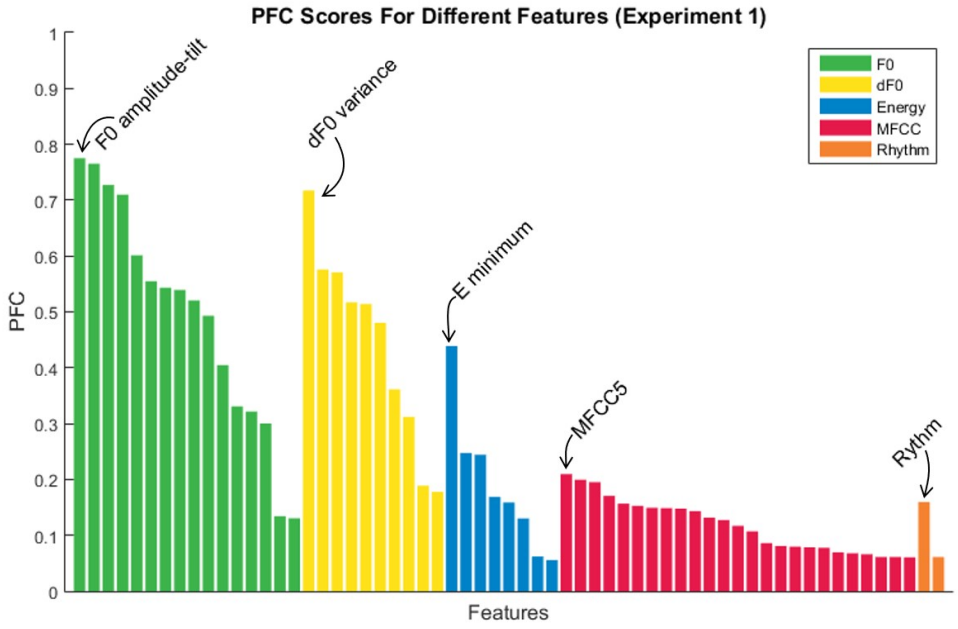


Figure 9.1: PFC scores of the Initial feature set over the Hebrew Q&N dataset.

- The MFCC family receives low scores. In general, MFCC is known to be less sensitive to prosody changes [128] and not sensitive to pitch changes. Since the differences between Neutral and Question prosody are mainly in pitch changes, it makes sense that MFCC would receive a low PFC score.

- The Energy and Rhythm families also receive low PFC scores, despite the fact they are considered to be prosodic [69]. The explanation for these results is that our two prosodic classes are not very different in their energy and rhythm behaviors. In other words, it is difficult to distinguish between a sentence uttered with Question or Neutral prosody, based solely on the energy or the rhythm of the sentence.

These conclusions bring us back to the discussion about the prosodic nature of a feature. In the general case, we can say that a feature has a different sensitivity to different prosodies. Thus, a certain feature may separate between one pair of classes much better than between another pair of classes.
When focusing on prosody, a feature can carry prosodic information for some of the prosodic classes (or maybe none of them). Obviously, there is no single feature that carries prosodic information for all classes. That includes the traditional prosodic features - F0, energy, and rhythm which do not always carry prosodic information for all classes.

The third point, which was mentioned above about energy and rhythm families, illustrates this distinction and shows us the strength of the PFC. In that case, two feature families (energy and rhythm) that are considered to be prosodic, received low PFC scores, as they cannot distinguish between Question and Neutral prosodies.

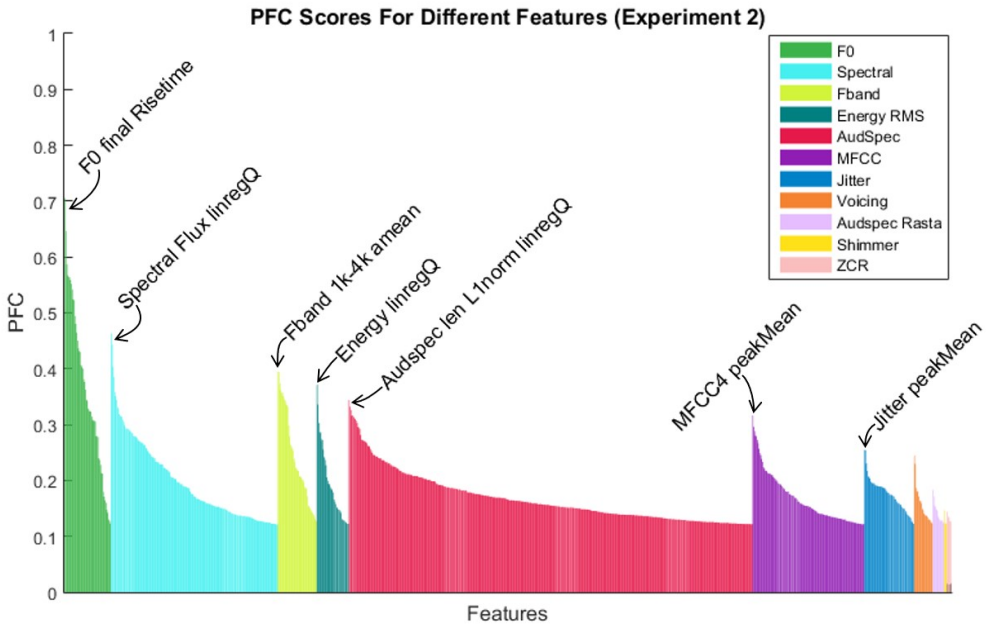


Figure 9.2: PFC scores of the best 1,000 features out of OpenSMILE feature set over the Hebrew Q&N dataset. The feature families from the Initial feature set have the same ranking as seen in figure 9.1

9.1.2 Experiment 2: OpenSMILE Feature Set + Hebrew Q&N Dataset

Figure 9.2 shows results of the OpenSMILE feature set over the Hebrew Q&N dataset (section 6.3.1). As there are thousands of features, we decided to show the 1,000 features with the highest PFC scores. A few conclusions out of this figure:

- PFC values of F0, Energy, and MFCC features families are ranked in the same order as in the Initial feature set. This is important and validates the PFC as these features were calculated differently: in the Initial feature set, we implemented the features ourselves, while in this dataset, we used the OpenSMILE implementation.

- Jitter and Shimmer families receive low scores, even though they are considered to be prosodic. These features track local changes in pitch, but in the examined

prosodic classes (Neutral and Question), changes are global. Therefore Jitter and Shimmer are less relevant for these specific prosodies.
Again we see the same phenomenon - some features that are considered to be prosodic for some classes do not convey prosodic information for all classes. The PFC successfully distinguishes between these cases.

9.1.3 Experiment 3: Initial Feature Set + 2 classes English Emotions Dataset

Figure 9.3 shows PFC results over the Initial feature set, using the Emotions dataset. We can see that F0 family receives high PFC scores. The differences of F0 between these two classes can be heard when listening to the different utterances.
This is another validation of the PFC as we get similar behavior for a totally different dataset and classes. We were surprised to see that MFCC2 received a relatively high score. This is unexpected as MFCC family is not considered to convey prosodic information. This result requires deeper investigation, which is beyond the scope of this work.

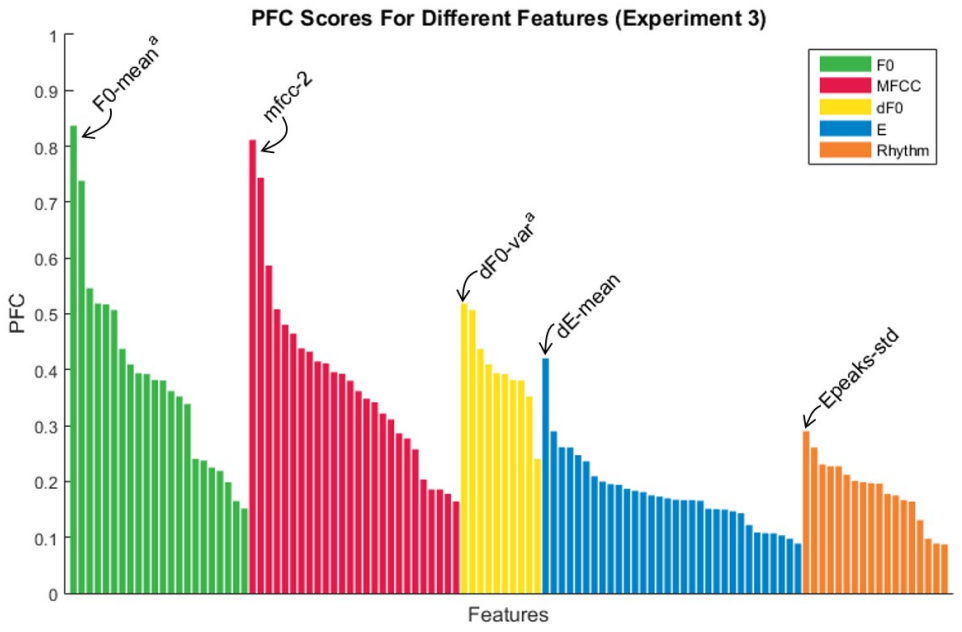


Figure 9.3: PFC scores for the Initial feature set over the English 2 classes dataset.

9.2 Validation 2: Comparison to a Classification Task

One of the goals in grading and estimating features' quality using the PFC is to choose the best features to be used in a classification task. Hence, another way to validate the PFC is to compare two methods of features ranking: by PFC scores and by classification performance.

In section 9.2.1, we describe some standard methods used to estimate classifiers' performances. Then, in section 9.2.2, we show the results for each experiment.

9.2.1 Classification Performances

In order to evaluate the performance of a classification model, we can compare the model's predicted class for each sample to the actual tagged class (i.e., ground truth). We focus on a binary classification task.

For each instance in the dataset, there are four possible combinations between the classifier's prediction and the ground truth: (1) true-positive (2) true-negative (3) false-positive and (4) false-negative.
The terms true or false refer to the prediction being correct or incorrect, while the terms positive or negative refer to classifier prediction. E.g., false-positive means that the classifier predicts this instance as a positive, but it is wrong, and the actual tagging is negative.

It is common to arrange these four values in a 2X2 matrix, called a confusion matrix. We can normalize them by the size of the marginal total (either the classifier prediction or the actual tagging). This normalization leads to two 2X2 matrices and a total of 8 different ratio measures as can be seen in figure 9.4).

A few common measures shown in figure 9.4 that we use are:
(1) Precision - when the classifier predicts a certain instance class as positive, what is the probability that this instance's label is indeed positive.
(2) Recall - what is the probability to classify a certain instance as positive, when it is known that this sample is labeled positive.

There is a trade-off between these two measures. For a specific classification model, increasing the recall, decreases precision and vice-versa. We can completely control the value of only one of the above measures, e.g., setting recall to be 100% by classifying all samples as positive, but as a result, the precision of this model will be very low.
A standard measure that combines these measures into a single value is the F1-score (see figure 9.4). This is the harmonic mean between precision and recall.

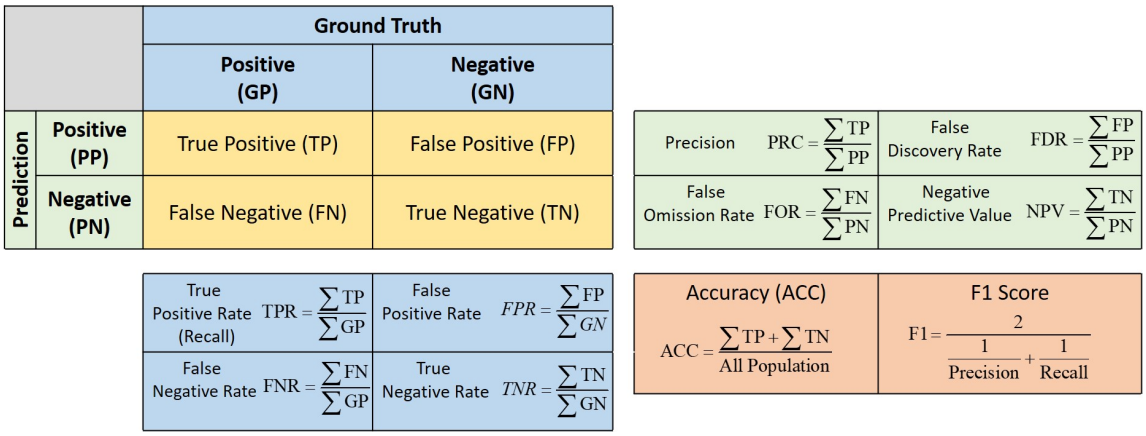


Figure 9.4: Confusion Matrix of the binary classification problem (yellow), with the common classifier's measurements (in green, red, and blue).

9.2.2 Comparison Process

Now we explain how we compare the two ranking methods: PFC ranking and classifier's performance. We decided to use the F1 score as the classifier's performance measure as it is a standard measure that combines both recall and precision.

For each feature separately, we train a classifier. We train with a single feature as the PFC is calculated over single features as well. We use a simple logistic regression classifier using 66% of the data for training while making sure train and test sets do not contain the same speakers.
For each feature, we use a threshold that yields the maximum F1 score over the train set. The "classifier performance" on the other hand is the F1 score obtained by applying the above threshold to the test set.

To visualize the correlation between the PFC and F1 scores, we arrange all data in a 2D graph. Every point represents a single feature with its F1 score on the x-axis and its PFC score on the y-axis.
In addition, we want to have a numerical value, so we calculate the Pearson's correlation coefficient between all data points:

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{9.1}$$

Where: $\operatorname{cov}()$ is the covariance and σ_X, σ_Y are the standard deviations of X and Y respectively. In our case X and Y are the PFC and F1 scores.

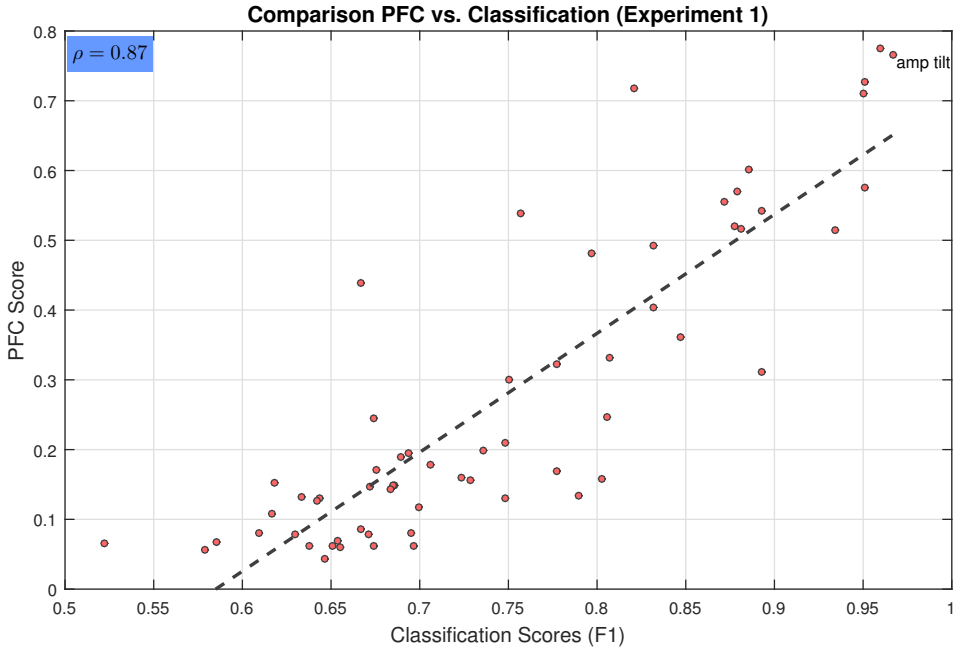


Figure 9.5: Comparison between F1 scores to PFC scores of experiment 1 showing positive correlation between PFC and classification. Table 9.1 presents additional details about the best PFC features.

9.2.3 Comparison Results

Next, we show this validation over experiments 1-3, as described in section 8.

Experiment 1: Initial Feature Set + Hebrew Q&N Dataset

Figure 9.5 shows the graph of PFC vs. F1 scores. We can easily see the positive correlation between the two methods. We also calculated Pearson's correlation and found that $\rho_{=}0.87$, which shows a significant positive correlation. We should note that it is still a small feature set and dataset, and therefore in the next sections we will show validation over a larger dataset and feature set.

Table 9.1 compares between the two ranking methods. We chose to look at the 15 features with the highest PFC scores. The table is sorted by the PFC score of each feature. It shows the PFC score and corresponding ranking, together with F1 score and ranking. Note that out of these 15 best PFC features, only two ranked below 15 in the F1 score.

For both comparison methods, either by a 2D graph or by a table, we should note that the PFC scores and the classification performance values are not on the same scale.

Feature name	PFC		F1	
	Ranking	Score	Ranking	Score
Amp tilt	1	0.78	2	0.96
Durr tilt	2	0.77	1	0.97
Amp tilt[a]	3	0.73	3	0.95
dF0 var[a]	4	0.72	17	0.82
Durr tilt[a]	5	0.71	5	0.95
F0 mean[a]	6	0.60	9	0.89
dF0 mean[a]	7	0.58	4	0.95
dF0 max[a]	8	0.57	11	0.88
F0 mean	9	0.55	13	0.87
F0 max[a]	10	0.54	8	0.89
F0 var	11	0.54	25	0.76
F0 max	12	0.52	12	0.88
dF0 max	13	0.52	10	0.88
ddF0 mean	14	0.51	6	0.93
F0 var[a]	15	0.49	15	0.83

Table 9.1: Comparison between two ranking methods: PFC and F1 scores. We can see that the F1 ranking of only two features is below 15. Features' naming notation: F[a] denotes accumulated features as explained in section 7.2.1

Therefore we compare the ranking of the features and not the actual values.

Experiment 2: OpenSMILE Feature Set + Hebrew Q&N Dataset

We perform the same process over the 4,368 OpenSMILE features. Figure 9.6 shows the results, and again we can see the positive correlation. In general, for low PFC and F1 scores, results are noisy, and there is no real correlation between the methods. We can see an example for that around the area of PFC < 0.2 and F1 < 0.75.
Still, when calculating Pearson's correlation coefficient, we receive a positive correlation of $\rho = 0.72$ between these two methods. It is another result that supports the relevancy of the PFC method.

We should note that the correlation coefficient is smaller than one; This means that ranking features by the F1 score or by the PFC is still not an identical process. This is significant and emphasizes the importance of the PFC as not merely a feature selection method. It conveys more information about a feature and not just whether a feature is suitable for a specific classification task.

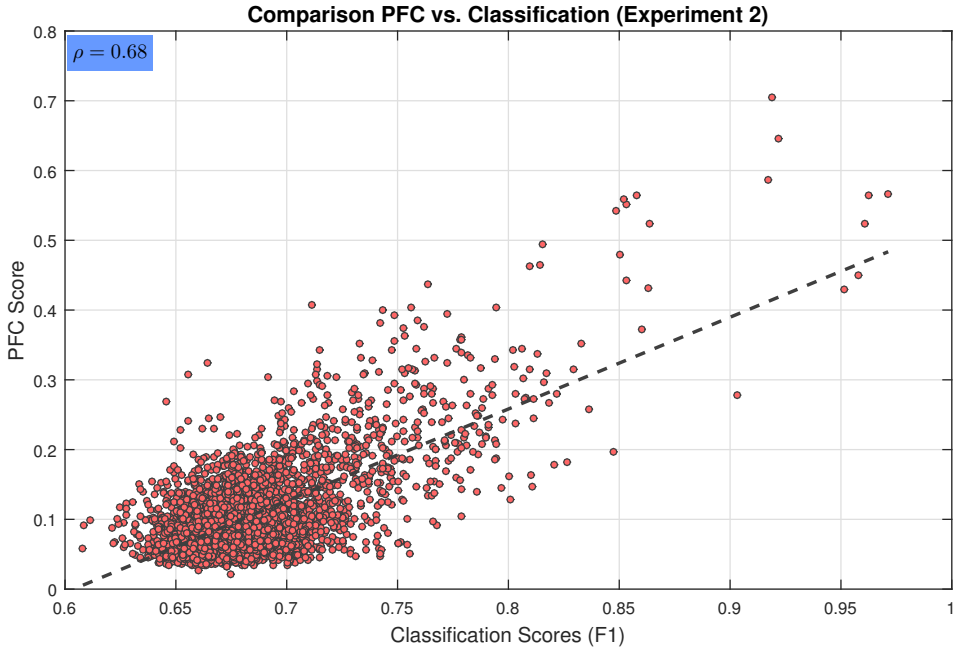


Figure 9.6: Comparison between F1 and PFC scores of experiment 2 showing positive correlation.

Experiment 3: Initial Feature Set + English 2 Classes Emotions Dataset

In this experiment we ran the same validation test shown in figure 9.7. The dataset is more complicated as it contains utterances of different length, in another language, and the classes we chose (Neutral and Hot Anger) are less significant and more variable than the classes in experiment 1 (Neutral and Question).
Still, we can see positive correlation of $\rho = 0.71$ between the two methods. This test validates the PFC using a different independent dataset.

9.3 Validation 3: Dimensionality Reduction

In section 9.2 we analyzed the performance of classifiers that were trained by single features. In this section we extend this analysis by looking over a few features together. We choose a subset of features, that received the highest PFC scores, and test how well these can separate between different prosodic classes.
We deal with more than three features so obviously we cannot visualize more than three dimensions on the paper. Therefore we chose to reduce the dimensionality of the problem. In section 9.3.1 we explain how we reduce dimensionality and how we can use this method

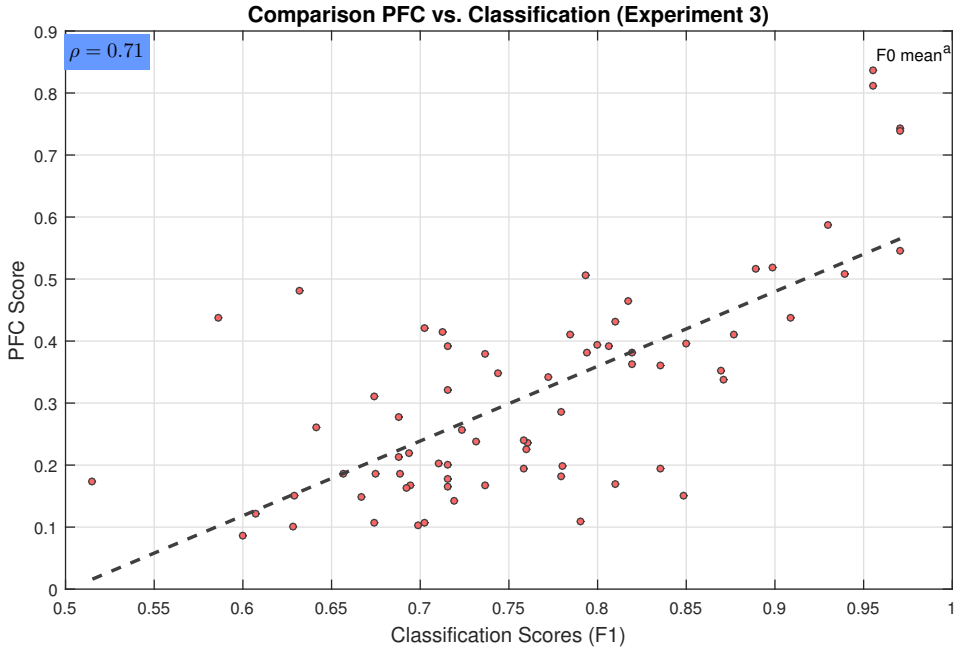


Figure 9.7: Comparison between F1 and PFC scores of experiment 3 showing positive correlation for a different dataset.

to visualize the data. Then, in section 9.3.2 we show the results of a few experiments.

9.3.1 What is Dimensionality Reduction

Dimensionality reduction is a field that contains a set of methods that project data points from their original space onto a new lower dimensionality space. The projection is usually made using a combination or a function of the original features. The transformation of the original features can either be linear or non-linear and can make use of all or just some of the original features. When using these techniques, we cannot always interpret the meaning of the new features, as they lose their original physical meaning.

Dimensionality reduction methods are in use for both: (1) solving some of the issues that arise by using a large number of features, such as overfitting. From this perspective, it can be considered as a features creation method. (2) For visualization purposes, when there are more than three features. We use dimensionality reduction for this purpose.

There are many dimensionality reduction methods; two famous ones are (1) Principal Component Analysis (PCA) [110], which performs a linear mapping of the features such that the variance of the data in the lower dimension space is maximal, and (2) Linear Discriminant Analysis (LDA) [35] which finds a linear combination of features in a way

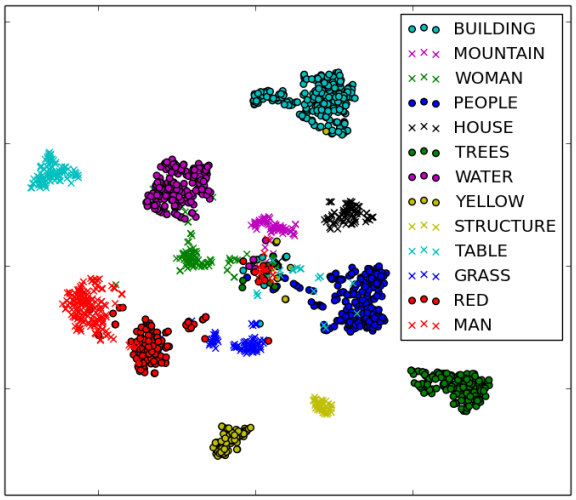


Figure 9.8: Example of t-SNE visualization of a DNN's output [145]. This DNN analyzed speech signals and projected the data onto 2D space. The visualization shows that the DNN separates between 14 different classes.

that separates between different classes.

In our work, we chose a third method, which became popular in recent years, as it is dedicated for visualization purposes: T-distributed Stochastic Neighbor Embedding (t-SNE) [98]. The t-SNE method was specifically developed to reduce multiple dimensions into two or three dimensions. It is a non-linear algorithm that maps each data point in such a way that similar points will be mapped to be closer in the lower dimension space.

Figure 9.8 shows an example taken from [145]. This work used neural networks to convert speech signals into lower dimension representation; then, they reduced the dimensionality of the network output into 2D using t-SNE. We can see that similar words are mapped to the same area in the 2D graph.

9.3.2 Validating the PFC Using Dimensionality Reduction

As mentioned above, we want to analyze the whole feature set, i.e., look at many features together, instead of at each feature separately.

In order to validate the PFC we examine several subsets of features (e.g., (a) the full feature set, or (b) the subset of features that received the best PFC scores). We reduce the dimensionality of each subset into 2D space to visualize it. We also color the instances by their classes (either prosodic or content class). Finally, we compare these 2D graphs of different features' subsets and draw some conclusions.

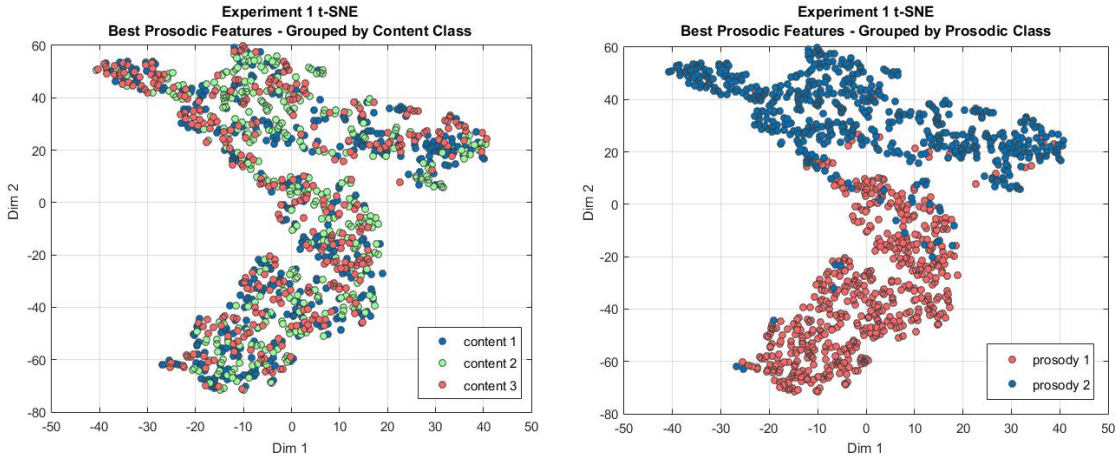


Figure 9.9: Experiment 1 - dimension reduction of the best prosodic features obtained using the PFC. There is a good separation between prosodic classes (right) and there is no separation between content classes (left).

Experiment 1: Initial Feature Set + Hebrew Q&N Dataset

Out of the Initial feature set we chose a subset of 15 features that received the highest PFC scores. Let's call them the "prosodic features subset." For each instance in the dataset (for each speech utterance), we used a feature vector that contains the above "prosodic features subset". Using the t-SNE algorithm we reduced the dimensionality from 15-dimensions into 2D space, in order to visualize the data.
Figure 9.9 (right) shows this dimensionality reduction, while splitting the data points by their prosody class. We can see that there is a good separation between the two prosodies, as can be expected from the "prosodic features subset".
In figure 9.9 (left), on the other hand, we split the same data points by their content class. We can see that there is no separation at all between the classes.
This further validates the PFC, as out of the full feature set, the PFC successfully found the best prosodic features, i.e., a subset of features that separates between prosody classes.

To examine it from a different perspective, we ran another test. We used a similar scheme like the PFC, but instead of using the probability distribution of "same" and "different" **prosodic** classes (as was explained in chapter 5), we used **content** classes. We can think about it as a criterion for "content features".
Figure 9.10 shows the best "content features" (after dimensionality reduction), while different colors represent different prosodies (and not content) classes. As expected, these features can not separate well between prosodic classes.

At this point, it is important to emphasize that when a dimensionality reduction algo-

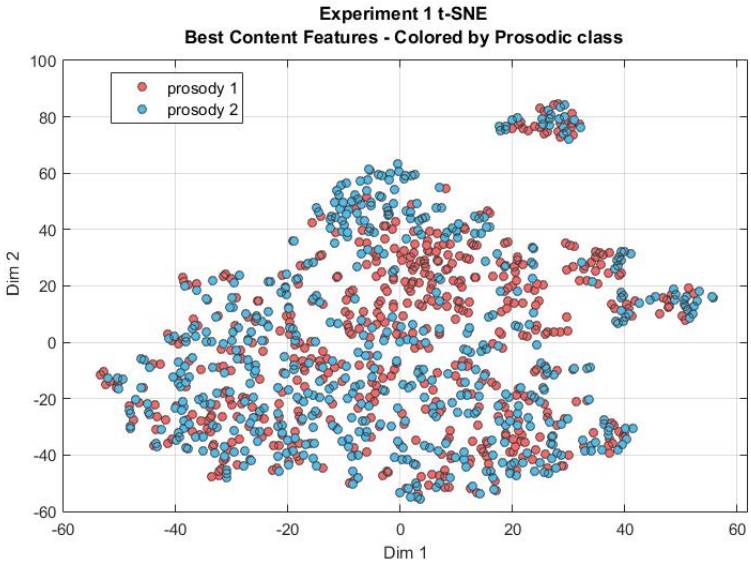


Figure 9.10: Experiment 1 - dimension reduction of the best content features. There is no separation between prosodic classes.

rithm shows a good separation between classes, we can definitely trust the results and state that these features can separate well between different classes. On the other hand, when a dimensionality reduction algorithm does not show any separation, we cannot deduce the opposite conclusion (i.e., that these features cannot separate the classes). The fact that we cannot see separation may just indicate that we could not find the right combination of the features or that the relevant dimension should be higher.

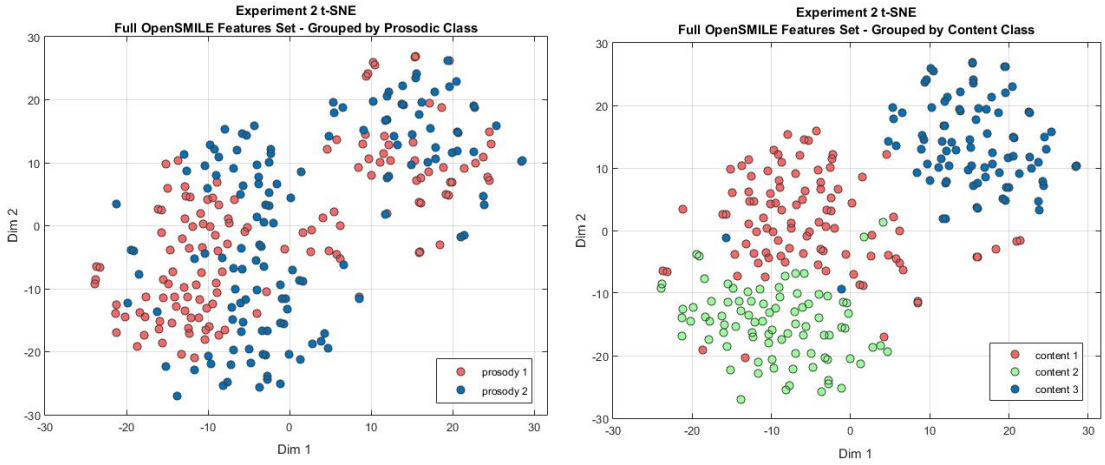


Figure 9.11: Experiment 2 - dimension reduction of the full OpenSMILE feature set shows separation between content classes (right) and not between prosodic classes (left).

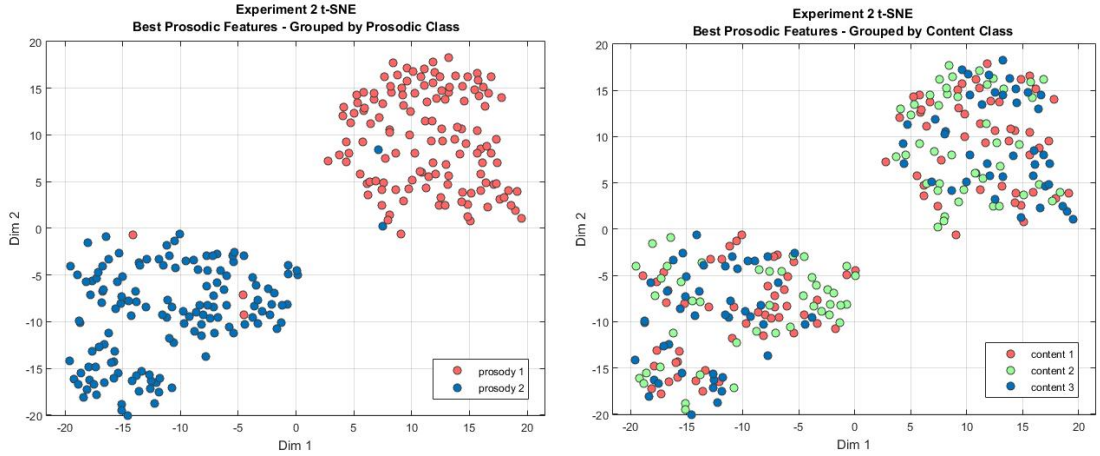


Figure 9.12: Experiment 2 - dimension reduction of the best prosodic features shows the opposite from figure 9.11. A good separation for prosodic classes (left) and not for content classes (right).

Experiment 2: OpenSMILE Feature Set + Hebrew Q&N Dataset

For the OpenSMILE features set, we visualize two different subsets of features. For each subset we apply the t-SNE algorithm and create two figures, one groups the instances by their prosodic classes while the other groups by the content classes.

Figure 9.11 shows that when looking at the full feature set, we do not see good separation between prosodic classes (left side of the figure). On the other hand, we found out that this full feature set provides good separation between the three content classes (right side of the figure). This means that the full feature set is biased towards content presentation and we can use it 'as is' for that type of task.

Figure 9.12 shows the opposite: dimensionality reduction over a subset of the 15 features that received the best PFC values, shows clear separation between the prosodic classes (left), but not between content classes (right). This means that a small subset of the highest scoring 15 features indeed satisfies the definition of a prosodic set, i.e., dependent on prosodic classes but independent of content classes.

These two tests show us the strength of the PFC, as out of thousands of features that are biased towards content classes, it succeeded in pinpointing a small subset of features that indeed carry prosodic information.

10 Summary

10.1 Discussion and Conclusions

The original purpose of this project was to research the psycho-acoustic phenomena of human speech. We decided to focus on prosody as it carries information that is related to both producing and understanding the speech signal by humans.

During the initial stages of our exploration regarding prosodic features, we realized that even though prosody was previously researched and is widely used in many fields, and despite the fact previous works used features that were considered prosodic – there is no clear definition of what makes features prosodic.

In this work, we proposed a simple definition of what prosodic features are. We mathematically formalized it into a Prosodic Feature Criterion (PFC) and presented a practical and numerical methodology to calculate this PFC score.
To the best of our knowledge, this is the first work that tries to quantify the quality of a prosodic feature. We believe that this is just a first step towards better understanding of the prosody field.

Our methodology measures the amount of prosodic information a single feature carries with respect to multiple prosodic classes. We implemented this methodology, ran it over two datasets and two feature sets, and validated our results using several experiments.

The PFC is currently limited to non-tonal languages. In addition to that, it can only currently examine one single feature at a time (and not a group of features).

10.2 Future Work

The methodology we presented can be extended. Future work includes validating the criterion with additional experiments, extending the criterion to overcome its current limitations, or using it for practical applications.

Here are a few extensions that we thought of:

- Multiple-Features: the PFC currently analyzes the prosodic nature of a single feature. It is known that a single feature is usually not enough to distinguish between different classes. Therefore it is necessary to extend the PFC to analyze the prosodic nature of a group of features.

- Tonal languages: we excluded tonal languages from this work, as changing the prosody in these languages also has lexical meaning. It can be interesting to conduct a research that analyzes the effect of prosodic features in these languages and how the PFC may assist in the evaluation of prosodic features.

- Further validations: even though we validated the PFC using many features and two datasets with different classes and languages, additional validations can be done using larger datasets, over more languages, more complicated classes and with additional speakers.

- Prosodic features generation: in this work, we analyzed the prosodic nature of existing features. As we now have a numerical way to represent the prosodic nature of a feature, we can try generating new prosodic features by applying feature construction process that maximizes the PFC.

- Practical applications: there are some applications which involve prosody that can rely on the PFC, such as:

  - An application for training users to achieve a desired prosody, based on biofeedback methods. That is, by illustrating to them using different visualizations "how far" they are from the desired prosody.
  - An automatic tool for objective assessment and monitoring of the progress of a neurological condition. This application can be useful for creating a better treatment plan for some neurological conditions that affect prosody, e.g. Parkinson's Disease.

To summarize, there are many ways we can further develop the PFC and use it for practical applications. We believe this work is just the first step towards more advanced and accurate prosodic research.

References

[Schuller u.a.)Schuller u.a.] Schuller, Bjorn / Steidl, Stefan / Batliner, Anton / Burkhardt, Felix / Devillers, Laurence / Muller, Christian / Narayanan, Shrikanth*The INTERSPEECH 2010 paralinguistic challenge*2794–2797.

[2] Noll, A Michael(1967): *Cepstrum pitch determination*, 2: 293–309.

[3] Lehiste, Ilse(1970): *Suprasegmentals.*

[4] Russell, James A / Mehrabian, Albert(1977): *Evidence for a three-factor theory of emotions*, 3: 273–294.

[5] Rabiner, Lawrence(1977): *On the use of autocorrelation analysis for pitch detection*, 1: 24–33.

[6] Fujisaki, Hiroya / Hirose, Keikichi(1982): *Modelling the dynamic characteristics of voice fundamental frequency with application to analysis and synthesis of intonation*In: Proceedings of 13th International Congress of Linguists57–70.

[7] Scott, Sheila / Caird, FI(1983): *Speech therapy for Parkinson's disease.*

[8] Scott, Sheila / Caird, FI(1984): *The response of the apparent receptive speech disorder of Parkinson's disease to speech therapy.*, 3: 302–304.

[9] Scherer, Klaus R / Ladd, D Robert / Silverman, Kim EA(1984): *Vocal cues to speaker affect: Testing two models*, 5: 1346–1356.

[10] Ross, Elliott D / Edmondson, Jerold A / Seibert, G Burton(1986): *The effect of affect on various acoustic measures of prosody in tone and non-tone languages: A comparison based on computer analysis of voice.*

[11] Childers, Donald G / Wu, Ke / Hicks, DM / Yegnanarayana, B(1989): *Voice conversion*, 2: 147–158.

[12] Shen, Xianonan Susan(1990): *Ability of learning the prosody of an intonational language by speakers of a tonal language: Chinese speakers learning French prosody*, 2: 119–134.

[13] Abe, Masanobu / Nakamura, Satoshi / Shikano, Kiyohiro / Kuwabara, Hisao(1990): *Voice conversion through vector quantization*, 2: 71–76.

[14] Scherer, Klaus R / Banse, Rainer / Wallbott, Harald G / Goldbeck, Thomas(1991): *Vocal cues in emotion encoding and decoding*, 2: 123–148.

[15] Silverman, Kim / Beckman, Mary / Pitrelli, John / Ostendorf, Mori / Wightman, Colin / Price, Patti / Pierrehumbert, Janet / Hirschberg, Julia(1992): *ToBI: A standard for labeling English prosody*In: Second international conference on spoken language processing.

[16] Hirschberg, Julia(1993): *Pitch accent in context predicting intonational prominence from text*, 1-2: 305–340.

[17] Wightman, Colin W / Ostendorf, Mari(1994): *Automatic labeling of prosodic patterns*, 4: 469–481.

[18] Hirst, Daniel / Ide, Nancy / Veronis, Jean(1994): *Coding fundamental frequency patterns for multi-lingual synthesis with INTSINT in the MULTEXT project*In: The Second ESCA/IEEE Workshop on Speech Synthesis.

[19] Arnfield, Simon(1994): *Prosody and syntax in corpus based analysis of spoken English*.

[20] Berndt, Donald J / Clifford, James(1994): *Using dynamic time warping to find patterns in time series.*In: KDD workshop, 16: 359–370.

[21] Strom, Volker(1995): *Detection of accents, phrase boundaries and sentence modality in German with prosodic features.*

[22] Garner, Stephen R / others u.a.(1995): *Weka: The waikato environment for knowledge analysis*In: Proceedings of the New Zealand computer science research students conference57–64.

[23] Banse, Rainer / Scherer, Klaus R(1996): *Acoustic profiles in vocal emotion expression.*, 3: 614.

[24] Zissman, Marc A(1996): *Comparison of four approaches to automatic language identification of telephone speech*, 1: 31.

[25] Dellaert, Frank / Polzin, Thomas / Waibel, Alex(1996): *Recognizing emotion in speech*In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP1970–1973.

[26] Murray, Iain R / Arnott, John L(1996): *Synthesizing emotions in speech: Is it time to get excited?*In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP1816–1819.

[27] Garner, Philip N / Browning, Sue R / Moore, Roger K / Russell, Martin J(1996): *A theory of word frequencies and its application to dialogue move recognition*In: Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP1880–1883.

[28] Jurafsky, Daniel u.a.(1997): *Automatic detection of discourse structure for speech recognition and understanding*In: IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings88–95.

[29] Shriberg, Elizabeth / Bates, Rebecca / Stolcke, Andreas(1997): *A prosody only decision-tree model for disfluency detection*In: Fifth European Conference on Speech Communication and Technology.

[30] Campbell, Joseph P(1997): *Speaker recognition: A tutorial*, 9: 1437–1462.

[31] Taylor, Paul / King, Simon / Isard, Stephen / Wright, Helen / Kowtko, Jacqueline(1997): *Using intonation to constrain language models in speech recognition*In: Fifth European Conference on Speech Communication and Technology.

[32] Shriberg, Elizabeth u.a.(1998): *Can prosody aid the automatic classification of dialog acts in conversational speech?*, 3-4: 443–492.

[33] Polzin, Thomas S / Waibel, Alex(1998): *Detecting emotions in speech*In: Proceedings of the CMC.

[34] Jurafsky, Daniel / Shriberg, Elizabeth / Fox, Barbara / Curl, Traci(1998): *Lexical, prosodic, and syntactic cues for dialog acts*In: Discourse Relations and Discourse Markers.

[35] Balakrishnama, Suresh / Ganapathiraju, Aravind(1998): *Linear discriminant analysis-a brief tutorial*1–8.

[36] Kain, Alexander / Macon, Michael W(1998): *Text-to-speech voice adaptation from sparse training data*In: International Conference on Spoken Language Processing.

[37] Taylor, Paul(1998): *The tilt intonation model*In: Fifth International Conference on Spoken Language Processing.

[38] Hirschberg, Julia / Nakatani, Christine(1998): *Using machine learning to identify intonational segments*In: Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing.

[39] Kjellin, Olle(1999): *Accent addition: Prosody and perception facilitates second language learning*, 2: 373–98.

[40] Conkie, Alistair / Riccardi, Giuseppe / Rose, Richard C(1999): *Prosody recognition from speech utterances using acoustic and linguistic based models of prosodic events*In: Sixth European Conference on Speech Communication and Technology.

[41] Amir, Noam / Ron, Samuel / Laor, Nathaniel(2000): *Analysis of an emotional speech corpus in Hebrew based on objective criteria*In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

[42] Erickson, Donna / Abramson, Arthur / Maekawa, Kikuo / Kaburagi, Tokihiko(2000): *Articulatory characteristics of emotional utterances in spoken English*In: Sixth International Conference on Spoken Language Processing.

[43] Batliner, Anton / Fischer, Kerstin / Huber, Richard / Spilker, Jorg / Noth, Elmar(2000): *Desperately seeking emotions or: Actors, wizards, and human beings*In: ISCA tutorial and research workshop (ITRW) on speech and emotion.

[44] Schulz, Geralyn M / Grant, Megan K(2000): *Effects of speech therapy and pharmacologic and surgical treatments on voice and speech in Parkinson's disease: a review of the literature*, 1: 59–88.

[45] Wightman, Colin W / Syrdal, Ann K / Stemmer, Georg / Conkie, Alistair / Beutnagel, Mark(2000): *Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis*In: Sixth International Conference on Spoken Language Processing.

[46] Paeschke, Astrid / Sendlmeier, Walter F(2000): *Prosodic characteristics of emotional speech: Measurements of fundamental frequency movements*In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion.

[47] Shriberg, Elizabeth / Stolcke, Andreas / Hakkani Tur, Dilek / Tu, Gokhan(2000): *Prosody-based automatic segmentation of speech into sentences and topics*, 1-2: 127–154.

[48] Noth, Elmar / Batliner, Anton / Kieling, Andreas / Kompe, Ralf / Niemann, Heinrich(2000): *Verbmobil: The use of prosody in the linguistic components of a speech understanding system*, 5: 519–532.

[49] Cano, Pedro / Loscos, Alex / Bonada, Jordi / Boer, Marteen de / Serra, Xavier(2000): *Voice morphing system for impersonating in karaoke applications*In: Proceedings of the International Computer Music Conference, ICMC.

[50] Polzin, TS(2000): *Waibel. A. Emotion-sensitive humancomputer interfaces*In: ICSA Workshop on Speech and Emotion: a Conceptual Framework for Research.

[51] Syrdal, Ann K / Hirschberg, Julia / McGory, Julie / Beckman, Mary(2001): *Automatic ToBI prediction and alignment to speed manual labeling of prosody*, 1-2: 135–151.

[52] Zheng, Fang / Zhang, Guoliang / Song, Zhanjiang(2001): *Comparison of different implementations of MFCC*, 6: 582–589.

[53] Yu, Feng / Chang, Eric / Xu, Ying Qing / Shum, Heung Yeung(2001): *Emotion detection from speech to enrich multimedia content*In: Pacific-Rim Conference on Multimedia550–557.

[54] Schroder, Marc(2001): *Emotional speech synthesis: A review*In: Seventh European Conference on Speech Communication and Technology.

[55] Pelecanos, Jason / Sridharan, Sridha(2001): *Feature warping for robust speaker verification*.

[56] Hastie / Poesio / Isard(2002): *Automatically predicting dialogue structure using prosodic features*63–79.

[57] Li, Aijun(2002): *Chinese prosody and prosodic labeling of spontaneous speech*In: Speech Prosody, International Conference.

[58] Fernandez, Raul / Picard, Rosalind W(2002): *Dialog act classification from prosodic features using support vector machines*In: Speech Prosody, International Conference.

[59] Auberge, Veronique(2002): *A gestalt morphology of prosody directed by functions: the example of a step by step model developed*In: Speech Prosody, International Conference.

[60] Bryant, Gregory A / Fox Tree, Jean E(2002): *Recognizing verbal irony in spontaneous speech*, 2: 99–119.

[61] Gutman, Dan / Bistritz, Yuval(2002): *Speaker verification using phoneme-adapted gaussian mixture models*In: 2002 11th European Signal Processing Conference1–4.

[62] Pierre Yves, Oudeyer(2003): *The production and recognition of emotions in speech: features and algorithms*, 1-2: 157–183.

[63] Scherer, Klaus R(2003): *Vocal communication of emotion: A review of research paradigms*, 1-2: 227–256.

[64] Campbell, Nick / Mokhtari, Parham(2003): *Voice quality: the 4th prosodic dimension*In: 15th ICPhS2417–2420.

[65] Chen, Ken / Hasegawa Johnson, Mark / Cohen, Aaron(2004): *An automatic prosody labeling system using ANN-based syntactic-prosodic model and GMM-based acoustic-prosodic model*In: 2004 IEEE International Conference on Acoustics, Speech, and Signal ProcessingI–509.

[66] Thompson, William Forde / Schellenberg, E Glenn / Husain, Gabriela(2004): *Decoding speech prosody: Do music lessons help?*, 1: 46.

[67] Gandour, Jackson / Tong, Yunxia / Wong, Donald / Talavage, Thomas / Dzemidzic, Mario / Xu, Yisheng / Li, Xiaojian / Lowe, Mark(2004): *Hemispheric roles in the perception of speech prosody*, 1: 344–357.

[68] Ananthakrishnan, Sankaranarayanan / Narayanan, Shrikanth S(2005): *An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model*In: Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)I–269.

[69] Tepperman, Joseph / Narayanan, Shrikanth(2005): *Automatic syllable stress detection using prosodic features for pronunciation evaluation of language learners*In: Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)I–937.

[70] Vidrascu, Laurence / Devillers, Laurence(2005): *Detection of real-life emotions in call centers*In: Ninth European Conference on Speech Communication and Technology.

[71] Hsia, Chi Chun / Wu, Chung Hsien / Liu, Te Hsien(2005): *Duration-embedded bi-HMM for expressive voice conversion*In: Ninth European Conference on Speech Communication and Technology.

[72] Kreiman, Jody / Gerratt, Bruce R(2005): *Perception of aperiodicity in pathological voice*, 4: 2201–2211.

[73] Kral, Pavel / Kleckova, Jana / Cerisara, Christophe(2005): *Sentence modality recognition in French based on prosody*In: International Conference on Enformatika, Systems Sciences and Engineering-ESSE185–188.

[74] Wang, Dagen / Narayanan, Shrikanth(2005): *An unsupervised quantitative measure for word prominence in spontaneous speech*In: Proceedings. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)I–377.

[75] Scherer, Klaus R(2005): *What are emotions? And how can they be measured?*, 4: 695–729.

[76] Tepperman, Joseph / Traum, David / Narayanan, Shrikanth(2006): *" Yeah Right": Sarcasm Recognition for Spoken Dialogue Systems*In: Ninth international conference on spoken language processing.

[77] Ananthakrishnan, Sankaranarayanan / Narayanan, Shrikanth(2006): *Combining acoustic, lexical, and syntactic evidence for automatic unsupervised prosody labeling*In: Ninth International Conference on Spoken Language Processing.

[78] Mareuil, Philippe Boula de / Vieru Dimulescu, Bianca(2006): *The contribution of prosody to the perception of foreign accent*, 4: 247–267.

[79] Quang, Vu Minh / Castelli, Eric / Yen, Pham Ngoc(2006): *A decision tree-based method for speech processing: question sentence detection*In: International Conference on Fuzzy Systems and Knowledge Discovery1205–1212.

[80] Thompson, William Forde / Balkwill, Laura Lee(2006): *Decoding speech prosody in five languages*, 158: 407–424.

[81] Liu, Yang / Shriberg, Elizabeth / Stolcke, Andreas / Hillard, Dustin / Ostendorf, Mari / Harper, Mary(2006): *Enriching speech recognition with automatic detection of sentence boundaries and disfluencies*, 5: 1526–1540.

[82] Beckman, Mary E / Hirschberg, Julia / Shattuck Hufnagel, Stefanie(2006): *The Original ToBI System and the*9.

[83] Demenko, Grażyna / Grocholewski, Stefan / Wagner, Agnieszka / Szymanski, Marcin(2006): *Prosody annotation for corpus based speech synthesis*In: Proceedings of the Eleventh Australasian International Conference on Speech Science and Technology460–465.

[84] Devillers, Laurence / Vidrascu, Laurence(2006): *Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs*In: Ninth International Conference on Spoken Language Processing.

[85] Campbell, William M / Campbell, Joseph P / Reynolds, Douglas A / Singer, Elliot / Torres Carrasquillo, Pedro A(2006): *Support vector machines for speaker and language recognition*, 2-3: 210–229.

[86] Bonafonte, Antonio u.a.(2006): *TC-STAR: Specifications of Language Resources and Evaluation for Speech Synthesis.*In: LREC142–160.

[87] Barrobes, Helenca Duxans(2006): *Voice Conversion applied to Text-to-Speech systems*.

[88] Quang, V Minh / Besacier, Laurent / Castelli, Eric(2007): *Automatic question detection: prosodic-lexical features and crosslingual experiments*In: Eighth Annual Conference of the International Speech Communication Association.

[89] Hsia, Chi Chun / Wu, Chung Hsien / Wu, Jian Qi(2007): *Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion*, 9: 1245–1254.

[90] Ananthakrishnan, Sankaranarayanan / Narayanan, Shrikanth(2007): *Improved speech recognition using acoustic and lexical correlates of pitch accent in a n-best rescoring framework*In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)IV–873.

[91] Farrus, Mireia / Hernando, Javier / Ejarque, Pascual(2007): *Jitter and shimmer measurements for speaker recognition*In: Eighth annual conference of the international speech communication association.

[92] Korpilahti, Pirjo / Jansson Verkasalo, Eira / Mattila, Marja Leena / Kuusikko, Sanna / Suominen, Kalervo / Rytky, Seppo / Pauls, David L / Moilanen, Irma(2007): *Processing of affective speech prosody is impaired in Asperger syndrome*, 8: 1539–1549.

[93] Li, Xi / Tao, Jidong / Johnson, Michael T / Soltis, Joseph / Savage, Anne / Leong, Kirsten M / Newman, John D(2007): *Stress and emotion classification using jitter and shimmer features*In: IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSPIV–1081.

[94] Paulmann, Silke / Kotz, Sonja A(2008): *Early emotional prosody perception based on different speaker voices*, 2: 209–213.

[95] Sridhar, Vivek Kumar Rangarajan / Bangalore, Srinivas / Narayanan, Shrikanth S(2008): *Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework*, 4: 797–811.

[96] Kathol, Andreas / Tur, Gokhan(2008): *Extracting question/answer pairs in multi-party meetings*In: IEEE International Conference on Acoustics, Speech and Signal Processing5053–5056.

[97] Mary, Leena / Yegnanarayana, Bayya(2008): *Extraction and representation of prosodic features for language and speaker recognition*, 10: 782–796.

[98] Maaten, Laurens van der / Hinton, Geoffrey(2008): *Visualizing data using t-SNE*2579–2605.

[99] Boakye, Kofi / Favre, Benoit / Hakkani Tur, Dilek(2009): *Any questions? Automatic question detection in meetings*In: IEEE Workshop on Automatic Speech Recognition & Understanding485–489.

[100] Turk, Oytun / Buyuk, Osman / Haznedaroglu, Ali / Arslan, Levent M(2009): *Application of voice conversion for cross-language rap singing transformation*In: IEEE International Conference on Acoustics, Speech and Signal Processing3597–3600.

[101] Diehl, Joshua J / Paul, Rhea(2009): *The assessment and treatment of prosodic disorders and neurological theories of prosody*, 4: 287–292.

[102] Hargrove, Patricia / Anderson, Amy / Jones, Jessica(2009): *A critical review of interventions targeting prosody*, 4: 298–304.

[103] Schuller, Bjorn / Steidl, Stefan / Batliner, Anton(2009): *The interspeech 2009 emotion challenge*In: Tenth Annual Conference of the International Speech Communication Association.

[104] Stylianou, Yannis(2009): *Voice transformation: a survey*In: IEEE International Conference on Acoustics, Speech and Signal Processing3585–3588.

[105] Devillers, Laurence / Vidrascu, Laurence / Layachi, Omar(2010): *Automatic detection of emotion from vocal expression*232–244.

[106] Ringeval, Fabien / Demouy, Julie / Szaszak, Gyorgy / Chetouani, Mohamed / Robel, Laurence / Xavier, Jean / Cohen, David / Plaza, Monique(2010): *Automatic intonation recognition for the prosodic assessment of language-impaired children*, 5: 1328–1342.

[107] Wagner, Michael / Watson, Duane G(2010): *Experimental and theoretical advances in prosody: A review*, 7-9: 905–945.

[108] Eyben / Wollmer / Schuller(2010): *Opensmile: The Munich Versatile and Fast Open-source Audio Feature Extractor*In: Proceedings of the 18th ACM International Conference on Multimedia1459–1462.

[109] Kinnunen, Tomi / Li, Haizhou(2010): *An overview of text-independent speaker recognition: From features to supervectors*, 1: 12–40.

[110] Abdi, Hervé / Williams, Lynne J(2010): *Principal component analysis*, 4: 433–459.

[111] Bryant, Gregory A(2010): *Prosodic contrasts in ironic speech*, 7: 545–566.

[112] Gaikwad, Santosh K / Gawali, Bharti W / Yannawar, Pravin(2010): *A review on speech recognition technique*, 3: 16–24.

[113] Vicsi, Klaa / Szaszak, Gyorgy(2010): *Using prosody to improve automatic speech recognition*, 5: 413–426.

[114] Ni, Chong Jia / Liu, Wenju / Xu, Bo(2011): *Automatic prosodic events detection by using syllable-based acoustic, lexical and syntactic features*In: Twelfth Annual Conference of the International Speech Communication Association.

[115] Dahl, George E / Yu, Dong / Deng, Li / Acero, Alex(2011): *Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition*, 1: 30–42.

[116] Deng, Li / Yu, Dong(2011): *Deep convex net: A scalable architecture for speech pattern classification*In: Twelfth Annual Conference of the International Speech Communication Association.

[117] Kasaei, S Hamidreza / Kasaei, S Mohammadreza / Kasaei, S Alireza(2011): *Development an Automatic Speech to Facial Animation Conversion for Improve Deaf Lives*, 2: 18–26.

[118] Li, Shang-wen / Wang, Yow-Bang / Sun, Liang-Che / Lee, Lin-Shan(2011): *Improved Tonal Language Speech Recognition by Integrating Spectro-Temporal Evidence and Pitch Information with Properly Chosen Tonal Acoustic Units*In: INTERSPEECH.

[119] Schuller, Bjorn / Steidl, Stefan / Batliner, Anton / Schiel, Florian / Krajewski, Jarek(2011): *The INTERSPEECH 2011 speaker state challenge*In: Twelfth Annual Conference of the International Speech Communication Association.

[120] Povey, Daniel u.a.(2011): *The Kaldi speech recognition toolkit*In: IEEE 2011 workshop on automatic speech recognition and understanding, CONF: .

[121] Alba Ferrara, Lucy / Hausmann, Markus / Mitchell, Rachel L / Weis, Susanne(2011): *The neural correlates of emotional prosody comprehension: disentangling simple from complex emotion*, 12: e28701.

[122] Bazillon, Thierry / Maza, Benjamin / Rouvier, Michael / Bechet, Frederic / Nasr, Alexis(2011): *Speaker role recognition using question detection and characterization*In: Twelfth Annual Conference of the International Speech Communication Association.

[123] Shirvan, R Arefi / Tahami, E(2011): *Voice analysis for detecting Parkinson's disease using genetic algorithm and KNN classification method*In: 2011 18th Iranian Conference of Biomedical Engineering (ICBME)278–283.

[124] Batliner, Anton u.a.(2011): *Whodunnit–searching for the most important feature types signalling emotion-related user states in speech*, 1: 4–28.

[125] Kolar, Jachym / Lamel, Lori(2012): *Development and evaluation of automatic punctuation for French and English speech-to-text*In: Thirteenth Annual Conference of the International Speech Communication Association.

[126] Koolagudi, Shashidhar G / Rao, K Sreenivasa(2012): *Emotion recognition from speech: a review*, 2: 99–117.

[127] Mohammadi, Seyed Hamidreza / Kain, Alexander / Santen, Jan PH van(2012): *Making conversational vowels more clear*In: Thirteenth Annual Conference of the International Speech Communication Association.

[128] Singh, Nilu / Khan, RA / Shree, Raj(2012): *Mfcc and prosodic feature extraction techniques: A comparative study*, 1: .

[129] Rakov, Rachel / Rosenberg, Andrew(2013): *" sure, i did the right thing": a system for sarcasm detection in speech.*In: Interspeech842–846.

[130] Godoy, Elizabeth / Koutsogiannaki, Maria / Stylianou, Yannis(2013): *Assessing the intelligibility impact of vowel space expansion via clear speech-inspired frequency warping*.In: Interspeech1169–1173.

[131] Valstar, Michel u.a.(2013): *AVEC 2013: the continuous audio/visual emotion and depression recognition challenge*In: Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge3–10.

[132] Abdel Hamid, Ossama / Jiang, Hui(2013): *Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code*In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing7942–7946.

[133] Alegre, Federico / Amehraye, Asmaa / Evans, Nicholas(2013): *Spoofing countermeasures to protect automatic speaker verification from voice conversion*In: IEEE International Conference on Acoustics, Speech and Signal Processing3068–3072.

[134] Wu, Zhizheng / Larcher, Anthony / Lee, Kong Aik / Chng, Engsiong / Kinnunen, Tomi / Li, Haizhou(2013): *Vulnerability evaluation of speaker verification under voice conversion spoofing: the effect of text constraints.*In: INTERSPEECH.

[135] Abdel Hamid, Ossama / Mohamed, Abdel rahman / Jiang, Hui / Deng, Li / Penn, Gerald / Yu, Dong(2014): *Convolutional neural networks for speech recognition*, 10: 1533–1545.

[136] Teixeira, João Paulo / Fernandes, Paula Odete(2014): *Jitter, Shimmer and HNR classification within gender, tones and vowels in healthy voices*1228–1237.

[137] Fujii, Shinya / Wan, Catherine Y(2014): *The role of rhythm in speech and language rehabilitation: the SEP hypothesis*777.

[138] Ling, Zhen Hua / Kang, Shi Yin / Zen, Heiga / Senior, Andrew / Schuster, Mike / Qian, Xiao Jun / Meng, Helen M / Deng, Li(2015): *Deep learning for acoustic modeling in parametric speech generation: A systematic review of existing techniques and future trends*, 3: 35–52.

[139] Sun, Lifa / Kang, Shiyin / Li, Kun / Meng, Helen(2015): *Voice conversion using deep bidirectional long short-term memory based recurrent neural net-*

*works*In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)4869–4873.

[140] Khan, Rubeena A / Chitode(2016): *Concatenative speech synthesis: A Review*.

[141] Taneja, Khusboo / others u.a.(2016): *Emotion Detection from Speech by using Neural Network*, 1: 1–7.

[142] Tang, Yaodong / Huang, Yuchen / Wu, Zhiyong / Meng, Helen / Xu, Mingxing / Cai, Lianhong(2016): *Question detection from acoustic features using recurrent neural network with gated recurrent unit*In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)6125–6129.

[143] Cernak, Milos / Asaei, Afsaneh / Honnet, Pierre Edouard / Garner, Philip N / Bourlard, Herve(2016): *Sound pattern matching for automatic prosodic event detection*.

[144] Tahon, Marie / Devillers, Laurence(2016): *Towards a small set of robust acoustic features for emotion recognition: challenges*, 1: 16–28.

[145] Harwath, David / Torralba, Antonio / Glass, James(2016): *Unsupervised learning of spoken language with visual context*In: Advances in Neural Information Processing Systems1858–1866.

[146] Oord, Aaron van den u.a.(2016): *Wavenet: A generative model for raw audio*.

[147] Lopez, Gustavo / Quesada, Luis / Guerrero, Luis A(2017): *Alexa vs. Siri vs. Cortana vs. Google Assistant: a comparison of speech-based natural user interfaces*In: International Conference on Applied Human Factors and Ergonomics241–250.

[148] Sotelo, Jose / Mehri, Soroush / Kumar, Kundan / Santos, Joao Felipe / Kastner, Kyle / Courville, Aaron / Bengio, Yoshua(2017): *Char2wav: End-to-end speech synthesis*.

[149] Arik, Sercan O u.a.(2017): *Deep voice: Real-time neural text-to-speech*In: Proceedings of the International Conference on Machine Learning195–204.

[150] Mohammadi, Seyed Hamidreza / Kain, Alexander(2017): *An overview of voice conversion systems*65–82.

[151] Stehwien, Sabrina / Vu, Ngoc Thang(2017): *Prosodic event recognition using convolutional neural networks with context information*.

[152] Sousa, Mariana / Trancoso, Isabel / Moniz, Helena / Batista, Fernando(2017): *Prosodic exercises for children with ASD via virtual therapy*In: Atas da Conferencia Jornadas SUPERA59–69.

[153] McCabe, Daniel J / Altman, Kenneth W(2017): *Prosody: An overview and applications to voice therapy*555719.

[154] Orosanu, Luiza / Jouvet, Denis(2018): *Detection of sentence modality on French automatic speech-to-text transcriptions*38–46.

[155] Fishman, Ben / Opher, Irit(2018): *Prosodic Feature Criterion for Hebrew Using Different Feature Sets*In: IEEE International Conference on the Science of Electrical Engineering in Israel (ICSEE)1–5.

[156] Fishman, Ben / Lapidot, Itshak / Opher, Irit(2018): *Prosodic Features' Criterion For Hebrew*In: proceedings of International Conference on Text, Speech, and Dialogue.

[157] Duanmu, San (2007): *The phonology of standard Chinese*. , OUP Oxford.

[158] Epstein, Melissa Ann (2002): *Voice quality and prosody in English*. .

[159] Gerhard, David / others u.a. (2003): *Pitch extraction and fundamental frequency: History and current techniques*. , Department of Computer Science, University of Regina Regina, Canada.

[160] Gussenhoven, Carlos / others u.a. (2004): *The phonology of tone and intonation*. , Cambridge University Press.

[161] Kelly, Finnian (2014): *Automatic Recognition of Ageing Speakers*. .

[162] Kompe, Ralf / Kompe, R (1997): *Prosody in speech understanding systems*. , Springer.

[163] Ladd, D Robert (2008): *Intonational phonology*. , Cambridge University Press.

[164] Liberman, Mark (2002): *Emotional Prosody Speech and Transcripts LDC2002S28* https://catalog.ldc.upenn.edu/LDC2002S28.

[165] Mary, Leena (2019): *Significance of Prosody for Speaker, Language, Emotion, and Speech Recognition*. Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition.Springer: 1–22.

[166] Nespor, Marina / Vogel, Irene (2007): *Prosodic phonology: with a new foreword*. , Walter de Gruyter.

[167] Prieto, Pilar / Esteve Gibert, Nuria (2018): *The Development of Prosody in First Language Acquisition*. , John Benjamins Publishing Company.

[168] Taylor, Paul (2009): *Text-to-speech synthesis*. , Cambridge university press.

[169] Titze, Ingo R / Martin, Daniel W (1998): *Principles of voice production* .

[170] Vaissiere, Jacqueline (1983): *Language-independent prosodic features*. Prosody: Models and measurements.Springer: 53–66.

[171] Wahlster, Wolfgang (2013): *Verbmobil: foundations of speech-to-speech translation*. , Springer Science & Business Media.

[172] Waibel, Alex (1988): *Prosody and speech recognition*. , Morgan Kaufmann.

[173] Whitenack, Daniel (2017): *Machine Learning with Go: Implement Regression, Classification, Clustering, Time-series Models, Neural Networks, and More Using the Go Programming Language*. , Packt Publishing Ltd.

[174] Yu, Dong / Deng, Li (2016): *AUTOMATIC SPEECH RECOGNITION.*. , Springer.

A Overview of Dissimilarity and Distance Functions

Our PFC deeply involves dissimilarity functions, and therefore we will now shortly review the topic. We use dissimilarity functions in STEP 1 and STEP 4 of the PFC, when calculating the dissimilarity between a pair of feature vectors, or between two distributions.

In this work, we use "dissimilarity function" as a general term that relates to all types of functions that quantify the resemblance, or measure how far two elements are from each other. These elements can be represented either by a scalar or by a vector.

A special case of dissimilarity functions is metric functions (section A.1) and "statistical distance" (section A.2):

A.1 Metrics

A metric function $d(\cdot,\cdot)$ is a type of dissimilarity function, which satisfies the following conditions:

1. non-negativity: $d(x,y) \geq 0$
2. $d(x,y) = 0 \Leftrightarrow x = y$
3. symmetry: $d(x,y) = d(y,x)$.
4. triangle inequity: $d(x,y) + d(y,z) \geq d(x,z)$

An example of a well-known and widely used metric family is the Minkowski function of order p (eq. A.1) between the points $X, Y \in \mathbb{R}^n$. This function satisfies the metric

conditions just if $p \geqslant 1$:

$$d(x,y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{A.1}$$

The Minkowski distance has a few known special cases, such as the Manhattan distance ($p = 1$), Euclidean distance ($p = 2$) and Chebychev distance ($p = \infty$). Figure A.1 shows a few illustrations of known metrics and dissimilarity functions.

The usage of metrics and dissimilarity functions in the field of machine learning is very broad. Some systems use dissimilarity functions instead of metrics to incorporate fewer restrictions. For example, in this work we do not use metrics, but we do require the dissimilarity functions to satisfy the first three conditions of a metric.

Many dissimilarity (and metric) functions exist, and each one of them has its advantages and disadvantages. A choice of a different dissimilarity function has a significant effect on the measurement. So, what is the "right" function one should choose?
In most cases, there is no single "right" function. Usually, several functions can fit each problem, and each one of them shows a different perspective of the same problem.

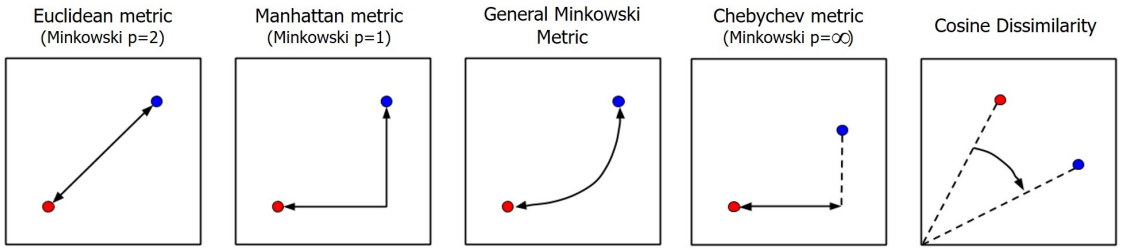


Figure A.1: 2D graphical illustration of different metric and dissimilarity functions (taken from [173]).

A.2 Statistical Distance

Real-world scenarios are, most of the time, stochastic; therefore, we use random variables and processes. We may also want to measure how far these random variables are from each other. We can do it by measuring the dissimilarity between two populations, i.e., between the distributions of these random variables.

A "statistical distance" measures how far two distributions are from each other. Most of the "statistical distances" do not satisfy all of the four metric conditions. Therefore, the term "statistical distance" is not accurate, even though it is widely used.

Divergence is a special case of "statistical distance," that satisfies only two of the metric conditions: (1) non-negativity and (2) $d(x,y) = 0 \Leftrightarrow x = y$. Obviously divergence is a weaker version than a metric, but still, these functions are used in many systems.
Examples of well-known and widely used "statistical distance" functions (some of them are divergences) are Kullback–Leibler, Jensen–Shannon, Bhattacharyya, and Helinger.

In this work, we use "statistical distances" in STEP 4 of the methodology when measuring how far the distributions of "same" and "different" prosodies are.

תקציר

המונח "פרוזודיה" מייצג את המידע שמועבר באות הדיבור, שאינו התוכן המילולי של המשפט. לדוגמא: האינטונציה, עוצמת הדיבור, הקצב והמקצב, צבע הקול ועוד.

פרוזודיה מהווה כלי חשוב ביותר בתקשורת יומיומית בין בני אדם, זאת בזכות המידע רב הערך אותו היא נושאת. מהפרוזודיה ניתן להבין את כוונתו האמיתית של הדובר (האם הוא צינית/ רצינית/ שואל שאלה), ניתן גם להבין את מצב רוחו (האם הוא עצוב/ שמח/ מפחד). כמו כן ישנם מספר מצבים רפואיים אשר ניתן לאתר באמצעות הקשבה לפרוזודיה.

המחקר בתחום זה הוא רחב, נערך שנים רבות ומופיע תחת דיסיפלינות ותחומים רבים - הנדסיים ולא הנדסיים. אך למרות המחקר הרב שנעשה, תשומת לב מעטה הוקדשה לניסוח ופורמליזם של מהם מאפיינים פרוזודים (Prosodic Features). כלומר, מאפיינים הקשורים לאות הדיבור ויכולים לייצג מידע פרוזודי. זהו פער משמעותי במחקר אשר מהווה בסיס לעבודה זו.

שאיפתנו במחקר זה היא להגדיר מהו מאפיין פרוזודי במובן של כימות האינפורמציה הפרוזודית שמאפיין מביע. במחקר נציג את הקריטריון אותו פיתחנו שנקרא Prosodic Feature Criterion (PFC). קריטריון אשר יכול לשערך את מידת הפרוזודיות ואת אופיו הפרוזודי של המאפיין אותו בוחנים. בנוסף לכך, נציג מתודולוגיה שבעזרתה ניתן לחשב את תוצאת הקריטריון.

את ה-PFC פיתחנו ובדקנו על גבי 2 סטים שונים של מאפיינים (Feature sets): (1) סט מאפיינים סטנדרטיים בתחום עיבוד הדיבור, (2) קבוצה מתוך סט מאפיינים ידוע, הנקרא OpenSMILE – זהו סט המכיל אלפי מאפיינים.

במחקר זה, השתמשנו ב-2 בסיסי נתונים (Datasets): (1) בסיס נתונים בשפה העברית אשר תוכנן ונאסף במיוחד לצורך מחקר פרוזודי. סט זה מכיל 2 סוגי פרוזודיות – "משפט שאלה" ו"משפט ניטרלי". (2) בסיס נתונים בשפה האנגלית של Linguistic Data Consortium (LDC) – סט זה מכיל 15 סוגי פרוזודיות הכוללות מצבי רוח שונים.

לצורך ולידציה ובדיקה של הקריטריון השתמשנו במספר שיטות: השווינו את תוצאות ה-PFC עבור מאפיינים שונים לידע קודם בתחום. השווינו את ה-PFC גם לתוצאות משימת סיווג אותה ביצענו עם אותם מאפיינים ועל גבי אותן המחלקות. בנוסף, הראינו ויזואליזציות של ציוני ה-PFC עבור אוסף מאפיינים, זאת בעזרת שיטות להורדת מימדים. ויזואליזציות אלו הראו יכולת הפרדה בין מחלקות פרוזודיות שונות של מאפיינים אשר קיבלו ציון PFC גבוה.

כל הניסויים שערכנו מראים תוצאות מעודדות וחיוביות, אשר מצביעות על כך ש ה-PFC אכן יכול לשמש למדידת מידת הפרוזודיות של מאפיינים שונים וזאת באופן כמותי ואובייקטיבי.

אוניברסיטת תל אביב

הפקולטה להנדסה ע״ש איבי ואלדר פליישמן

בית הספר לתארים מתקדמים ע״ש זנדמן סליינר

קריטריון למאפיינים פרוזודיים

חיבור זה הוגש כעבודת מחקר לקראת התואר ״מוסמך אוניברסיטה״ בהנדסת חשמל ואלקטרוניקה

על ידי

בן פישמן

העבודה נעשתה בבית הספר להנדסת חשמל

בהנחיית פרופ׳ חגית מסר ירון
ד״ר עירית עופר

שבט תש״פ

אוניברסיטת תל אביב

הפקולטה להנדסה ע״ש איבי ואלדר פליישמן

בית הספר לתארים מתקדמים ע״ש זנדמן סליינר

קריטריון למאפיינים פרוזודיים

חיבור זה הוגש כעבודת מחקר לקראת התואר ״מוסמך אוניברסיטה״ בהנדסת חשמל ואלקטרוניקה

על ידי

בן פישמן

שבט תש״פ