ABSTRACT

Title of Thesis	DETECTION OF IRREGULAR PHONATION IN SPEECH
	Srikanth Vishnubhotla, Master of Science, January 2007
Directed By:	Prof. Carol Y Espy-Wilson Department of Electrical Engineering, University of Maryland

The problem addressed in this work is that of detecting and characterizing occurrences of irregular phonation in spontaneous speech. While published work tackles this problem as a two-hypothesis problem only in those regions of speech where phonation occurs, this work also focuses on trying to distinguish aperiodicity due to frication from that arising due to irregular voicing. In addition, this work also deals with correction of a current pitch tracking algorithm in regions of irregular phonation, where most pitch trackers fail to perform well, as evidenced in literature. Relying on the detection of such regions of irregular phonation, an acoustic parameter is then developed in order to characterize these regions for speaker identification applications. The algorithm builds upon the Aperiodicity, Periodicity and Pitch (APP) detector, a system designed to measure the amount of aperiodic and periodic energy in a speech signal on a frame-by-frame basis. The detection performance of the algorithm has been tested on a clean speech corpus, the TIMIT database, and on telephone speech corpus, the NIST 98 database, where regions of irregular phonation have been labeled by hand. The detection performance is seen to be 91.8% for the TIMIT database, with the percentage of false detections being 17.42%. The detection performance is 89.2% for the NIST 98 database, with the percentage of false detections being 12.8%. The corresponding pitch detection accuracy increased from 95.4% to 98.3% for the TIMIT database, and from94.8% to 97.4% for the NIST 98 database, on a frame basis, with the reference pitch coming from the ESPS pitch tracker. The creakiness parameter was added to a set of seven acoustic parameters for speaker identification on the NIST 98 database, and the performance was found to be enhanced by 1.5% for female speakers and 0.4% for male speakers for a population of 250 speakers. These results lead to the conclusion that the creakiness detection parameter can be used for speech technology. This work also has potential applications in the field of non-intrusive diagnosis of pathological voices.

DETECTION OF IRREGULAR PHONATION IN SPEECH

by

Srikanth Vishnubhotla

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Master of Science
2007

Advisory Committee
Dr. Carol Y Espy-Wilson, Chair and Advisor
Dr. Shihab Shamma
Dr. Rama Chellappa

ACKNOWLEDGEMENTS

"Mathru Devo Bhava, Pithru Devo Bhava, Acharya Devo Bhava"

Mother is equivalent to God, Father is equivalent to God, and Teacher is equivalent to God.

I would like to express my undying and infinite gratitude towards my parents, who have given me everything I wished for, and have made me what I am. I owe everything to them.

This is the best opportunity to express my sincere thanks and immense gratitude to my teacher and guide – Prof. Carol Espy-Wilson. She has been my inspiration and Guru during my work in the Speech Lab. She has taught me the philosophy and method of research, and has helped me shape my research career.

I would like to express my thanks to the members of my thesis committee, Prof. Rama Chellappa and Prof. Shihab Shamma, for sparing their invaluable time in reviewing the manuscript of the thesis, and for providing valuable comment and suggestions.

My thanks to all my colleagues of the Speech Communication lab. Discussions with Om, Tarun, Xinhui, Sandeep and Daniel have always proved to be fruitful and insightful. I would especially like to thank Om for all his help and suggestions at various stages of my research, especially with the APP detector. Sandeep has been a very good friend & colleague, and has helped in acquainting me with the speaker ID setup. Daniel has provided valuable advice and comments with the speaker ID experiments. I would like to thank Olakunle for his help with verifying manual transcription of creakiness in the TIMIT database.

I would like to acknowledge the timely help and support from the ECE and ISR IT staff.

My thanks go to all my friends here at UMD, who have made the last two years really memorable and have helped and supported me at various stages. Abhinav, Anuj and Pavan will always find a place among the best of my friends, and they have my ever-lasting gratitude. Ramya & Swati – thanks for all the good times outside and at home. And

everybody - thanks for all the participation in the innumerable games of AoM! Rahul, Sudarshan, Gunjan – thanks for the wonderful food at 143, Westway! Jishnu & Sravya – thanks for the late night rides. They've made a huge difference, I assure you! Bargava – thanks for all the philosophical (and all else!) discussions. Pooja – thanks for participating in the birthday bashes and the get-togethers. And for all the feedback!

Last but not the least – thanks to my family. My mom, dad and sister – I love you and miss you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION & BACKGROUND

Speech is the primary mode of communication in humans. Indeed, the most obvious trait that sets the human race apart from the rest of the living world is the ability of humans to communicate with each other with richness, efficiency and variety. The speech signal produced by humans contains much more information than purely the message to be passed: it also gives the listener an idea of the gender and emotional state of the speaker, the language of communication, as well as the speaker's identity in some cases. Each of these components of information is processed by the listener, and then put together in the brain to generate a complete picture of the particular conversation.

Automatic methods have been developed for a long time to capture these various levels of information, and researchers have contributed to different areas of speech processing by machines, like Automatic Speaker Recognition (ASR), Language Identification, Natural Language Processing (NLP), Speaker Identification (Speaker ID), Emotion Recognition etc. In addition, research has also focused on giving machines the ability to speak, or Speech Synthesis. Further, there have also been major contributions to the area of Speech Enhancement, with the intention of rendering speech signals more usable for machines and humans. However, while there have been significant advances in all of these components of speech technology, each of them is still far from perfect or even near-human performance, and a lot needs to be done before the problem of "teaching machines to converse like humans" can be declared solved. Indeed, for example, in the technology of speech recognition, human speech recognition beats ASR by an order of magnitude in both clear and noisy speech [1] – a gap that has not been bridged over the past nearly ten years.

There are several reasons for the marked difference between the performance of humans and machines. In particular, for the problems of ASR and speaker ID, the performance difference is significantly high and for most cases, unacceptable, and this might be attributed to the approach taken towards the problem [1]. Most ASR and speaker ID systems rely on the statistical pattern recognition framework, wherein the front-end contains the Feature Extractor, and the back-end consists of one of several class-discriminating modeling techniques. Figure 1.1 shows a typical statistical pattern recognition system akin to those used for ASR and speaker ID applications:

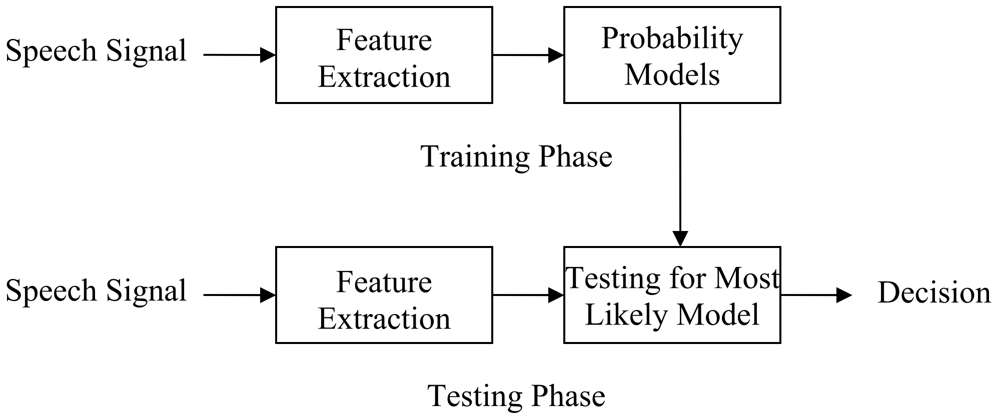

Figure 1.1 : A Typical Pattern Recognition System for ASR Speaker ID

The speech signal is first processed in order to extract certain features that enhance the discrimination between the different classes that are required to be identified (phonemes for ASR and speaker characteristics for speaker ID). The features used in current state of the art systems are the Mel-Frequency Cepstral Coefficients (MFCCs), spectral features of speech that have been processed by auditory-inspired techniques in order to mimic human auditory processing [2]. In order to minimize the effect of channel distortions, these features are smoothed using Relative Spectral Transform (RASTA) filtering [3], a technique that applies a band-pass filter to the energy in each frequency band in order to smooth short-term noise and to remove any constant offset in the band. To capture implicitly some of the temporal information in speech, these MFCCs are complemented by their first and second derivatives (ΔMFCCs and ΔΔMFCCs) which give the rate of change of these features with time.

Once the features are extracted, they are then fed into a modeling back-end, which characterizes their statistical and temporal behavior by means of marginal and conditional probability distributions, respectively. In state-of-the-art ASR systems, the back-end is a Hidden Markov Model (HMM) [4] which characterizes temporal behavior of the features, as well as their statistical description, by their conditional and marginal probability distributions. For speaker ID applications, the Gaussian Mixture Model (GMM), which accurately describes the statistical distribution of the extracted features in terms of a set of Gaussian distributions with varying means and variances, has been found to be most useful [5].

The reason why machine recognition has not lived up to expectations and still demands new developments is that the modeling techniques, as well as the feature extraction methods, have not been (i) accurate and (ii) appropriate to the tasks at hand. As an illustration, the speaker ID technology uses a purely statistical mechanism to model speaker characteristics, without taking temporal information into account. It is a known fact [6] that speakers differ in the way they articulate the same phoneme, and therefore, the characteristics of speech they produce should show different dynamic behavior. However, the GMMs used for Speaker ID do not explicitly capture this temporal variation of the features, and rely on the ability of features to model temporal behavior. This is unlike the HMMs which characterize conditional distributions to capture temporal behavior, but which are not used for speaker ID because of the inherent complexities in modeling multimodal distributions, and the lack of enough training data for each of the marginal and conditional probabilities.

More importantly, the choice of features is also a bottleneck in pushing the limits of current speech technology. The MFCCs that have been traditionally used for speech and speaker recognition are not very intuitively motivated, and more significantly, do not provide a clear picture of why they should work or fail in certain circumstances. These features implicitly code the human vocal tract characteristics, and their temporal behavior does not clearly reflect the vocal tract activity that would have generated such a MFCC structure. Further, the traditional set of MFCCs used are low time liftered, capturing only the vocal tract information and are thus not fully efficient, since for example, the source information of the speaker is also rich in speaker-specific information like voice quality and could be used for speaker ID. Thus, it is useful to explore beyond the MFCCs and look for efficient features that explicitly capture specific information that would be useful for the particular task at hand. For example, recent work has shown that for speaker ID applications, the performance of a set of eight acoustic parameters matches that of the traditional 39 MFCCs for varying population sizes, typically beating the MFCCs in the case of female speakers [6]. Thus, it is a promising enterprise to embark on the search for parameters that would explicitly capture specific information from the speech signal.

1.1 The Speech Production Process & The Acoustic Parameters

If it were possible to extract reliable features that explicitly capture the activities of the speech-producing mechanisms, then it would be possible to improve the performance of current speech technology. A brief look at the speech production process explains how:

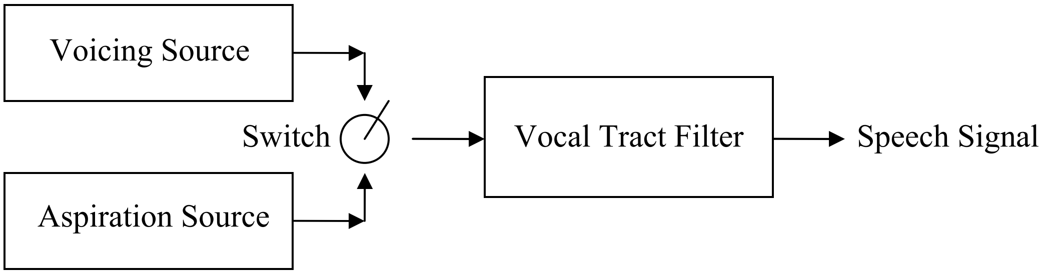

Figure 1.2 : The Speech Production Process modeled as an electrical system

The speech production process can be approximated by an electrical system, where a source signal excites a filter [7]. The behavior of the glottis is modeled by the source – when the glottis opens and closes periodically, as in the case of vowels, the source is represented by a periodic or quasi-periodic signal. When the glottis remains partially open allowing air to flow from the lungs and there are no major constrictions in the vocal tract, aspiration noise created just above the glottis produces the sound /h/. When the glottis remains open and a major constriction is created along the vocal tract, a second source is produced, either by creating a close constriction like in the case of fricatives (/f/, /s/ etc.) or by allowing the articulators to excite the vocal tract with sudden impulses, like in the case of stops (/p/, /t/ etc). In such cases, the source is represented by a noise source. In some phonemes, like the voiced fricatives (/z/), both the periodic and aperiodic sources are simultaneously active, in which case the periodic energy typically manifests itself in the lower frequencies, and the aperiodic energy in the higher frequencies. Each of the different vowels is produced when a periodic source is modified by a different configuration of the filter representing the vocal tract; these configurations are represented by the resonant frequencies of the filter. The harmonic spectrum of the periodic source thus gets modified by the vocal tract frequency response, giving the resultant speech a spectrum that has peaks at certain frequencies. These frequencies, called the Formant frequencies, are different for different vowels (and are in fact characteristic of them).

In particular, taking the example of ASR for vowels, a simple way to identify vowels automatically would then be to estimate the resonant frequencies from the speech spectrum, and then identify the vowel using the formants thus obtained. Though the variability in speech, both due to medium as well as due to speaker identity, renders the problem not so simple to solve, the principle is more or less the same. Thus, it comes naturally to use the formants as parameters for recognition of vowels – a subset of the parameters motivated by the speech production process and the acoustics involved. These parameters that are called Acoustic Parameters (APs) in the remainder of this thesis. For the case of speaker ID, the motivation of using acoustic parameters is that each speaker has a characteristic pattern of movement of articulators [6]. Thus, when acoustic parameters are used to characterize the speaker's speech, their behavior (both temporal and spectral) can help a lot in identifying the speaker from others. Further, the physical attributes of a speaker also limit the range of values that each of the acoustic parameters derived can take, and this can provide a valuable clue to speaker identity. Putting together this potpourri of acoustic parameters, that captures various characteristics of the speaker and speech generated, in an adequately capable modeling system, can give significant results for speech technology. For example, it was recently shown in Espy-Wilson et al, [6] that the speaker ID performance of a set of eight acoustic parameters either matches or beats that of the traditional 39 MFCCs.

While capturing features, it is necessary to characterize both the vocal tract filter as well as the source signal exciting the vocal tract. The set of acoustic parameters that have been used in Espy-Wilson et al, to characterize the vocal tract are the four formants. The degree of periodicity (and aperiodicity) in the source signal exciting the vocal tract, as well as the spectral tilt of the speech produced, are used to characterize the source information. The motivation for selection of these source features is elaborated on in [6].

The aim of this thesis is to develop an acoustic parameter that characterizes a particular type of source information (i.e., voice quality). In particular, this work deals with automatically detecting regions in speech where the source signal exciting the vocal tract shows irregular behavior – the signal is neither periodic nor noise-like, but has a different structure as a consequence of irregularity in the pattern of vibration of the vocal folds [8].

1.2 Phonation & Voice Quality

While it is the vocal tract that determines the phoneme to be produced, it is the mechanism of production of the glottal pulses exciting the vocal tract that determines the quality and perceptual attributes of the speech signal produced [7,9]. In this section, a brief description of the source mechanism during voiced speech is given [7]. The glottal pulses exciting the vocal tract are actually produced by the building of air pressure just behind the glottis. When the air pressure (coming from the lungs) is adequately high, the vocal folds that close the glottal opening are pushed apart to allow the air to escape. The volume velocity of

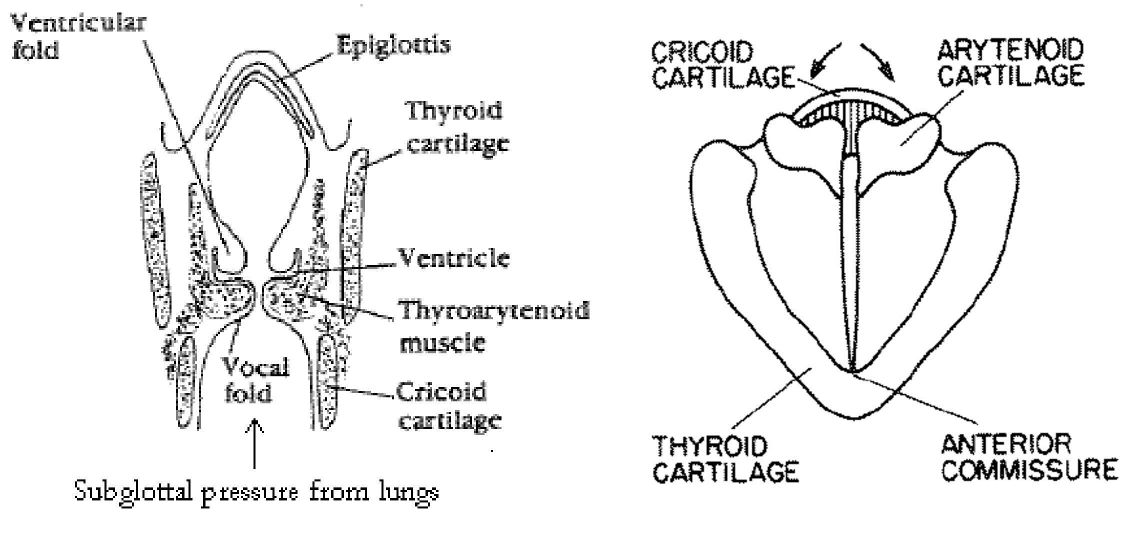

Figure 1.3 : A front view (left) and top view (right) of the vocal folds and the associated muscles (adapted from [7])

this air excites the vocal tract, causing it to resonate at certain frequencies according to its configuration.

Figure 1.3 shows the front view of the glottis and vocal folds. Air pressure from the lungs causes the glottis to open by pushing the vocal folds apart and leaking through the opening thus formed. The figure on the right shows a top-view of the vocal folds and associated muscles. Because of the close (but importantly, not exact) symmetry between the muscles controlling the two vocal folds, the vocal folds usually open and close in synchrony to each other, and thus the resultant volume velocity is a smooth function of time, with both the vocal folds modulating the air velocity equally and in a balanced way. It is noteworthy that the modulation by the vocal folds shapes the volume velocity that excites the vocal tract, and thus influences the nature of the speech produced. The actual modulation of the air by the vocal folds is shown in Figure 1.4.

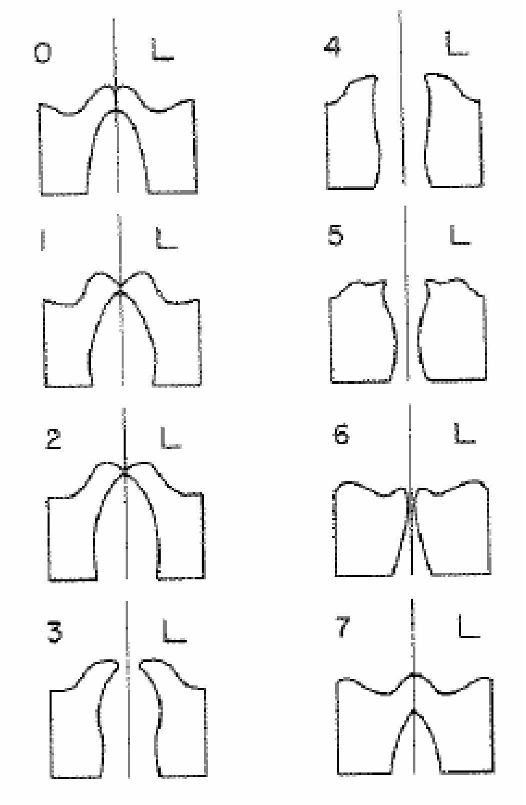

Figure 1.4 : Modulation of the air flow by the vocal folds (adapted from [7])

This is a view of the glottis and vocal folds from the front. It can be seen that the lower part of the vocal folds first separate out due to build up of sub-pressure, and then there is a gust of air that pushes through the vocal folds. Thus, the upper muscles now open apart and the air escapes from the glottis to the vocal tract. As the gust of air moves through the glottis, the lower muscles of the vocal folds start pulling them back together and the vocal folds start closing, from lower part to upper part. Thus, a Bernoulli Effect is seen in the opening and closure of the vocal folds. Further, the closure of the vocal folds is faster than the opening. This is because the vocal folds that are inherently closed have to be pushed apart and therefore take a sufficiently large sub-glottal pressure to be opened. Once they are opened, there is a large volume of air traversing upwards and thus, the vocal folds are held apart. In contrast, once the air rushes out, there is no pressure to keep the vocal fold muscles apart and they close shut more rapidly. These mechanisms are reflected in the structure of the volume velocity that excites the vocal tract. The source signal that excites the vocal tract is seen to have a slower and more gradual opening phase, and a faster, more rapid closing phase (Figure 1.5).

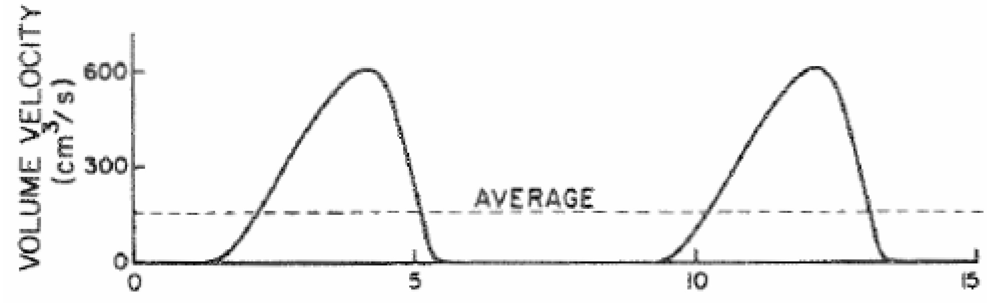

Figure 1.5 : Volume velocity of the air flow through the glottis during modal (normal) voicing (adapted from [7])

It is obvious that the physical characteristics of the speaker greatly determine the source signal modulating the vocal tract. Indeed, the typical durations of opening and closing of the glottis are a function of the acoustic mass of the upper and lower part of the vocal folds, as well as their acoustic compliances and the coupling compliance between them. (Mathematical details can be found in [7]). Thus, different speakers are expected to have different excitation patterns for their vocal tracts. Further, speakers are not consistent in the way they excite their vocal tract – the source signal for the same vowel can significantly change within a period of 10-15 msec for the same speaker during spontaneous speech. To complicate matters, the two vocal folds are not perfectly symmetric in terms of their properties like acoustic mass, hardness etc. and this causes an imbalance in the modulation of the volume velocity by the two vocal folds. All these phenomena result in the source signal exhibiting characteristics that differ so greatly from one phoneme to another, that it sometimes becomes difficult to define what can be called "normal" and "abnormal" for a specific speaker.

Perceptually, however, one can broadly classify "what speech sounds like" (i.e., the perceptual quality of speech, or *Voice Quality*) into three classes – Modal, Breathy and Creaky. Modal voicing is simply the normal voicing of a speaker – perceptually, it can be said that modal voicing is what a speaker usually sounds like. In this case, the vocal folds vibrate in the mode as described above, with the glottis opening and closing at regularly intervals and the vocal folds behaving symmetrically offering no greater resistance than they usually do [9]. In the case of Breathy voicing, the vocal folds do not always close completely and there is always a leakage of air through the slightly-open glottis (figure 1.6). Consequently, the speech produced has a slightly whispery or breathy quality to it, in the sense that there is always an underlying noisy element even during the vowels. Because of the leakage of air, the volume velocity has a non-zero DC component. This constant leakage of air also causes the resistance of the vocal fold muscles to be less cumbersome to air-flow, and thus, the opening phase is usually as free and the closing phase – giving rise to a symmetric glottal pulse. This, along with the fact that there is coupling between the sub-glottal and supra-glottal regions, explains the more sinusoidal nature of the glottal source pulse. Spectrally, the glottal source signal has a steeper slope (-18 dB/octave) than that of the modal case (-12 dB/octave). This is because of two reasons – the DC offset given by the leakage of the air, and the sinusoidal nature of the signal (as a signal approaches a sinusoid, its spectrum approaches that of an impulse and thus becomes steeper). This also causes the first harmonic (H1) of the source spectrum to be significantly higher than the second harmonic (H2). The H1-H2 difference for breathy vowels is usually about 6 dB more than that of the modal vowels. It must be noted that in the case of breathy voicing, the frequency of vocal fold vibration does not change and the pitch (F0) remains the same as it should in the modal case. Physiologically, the difference only exists in that the vocal folds allow a leakage of air during closure. Finally, this kind of voicing should also be noted as another example where a voiced source and aperiodic source act simultaneously, with the former dominating the lower frequencies and the latter, the higher frequencies.

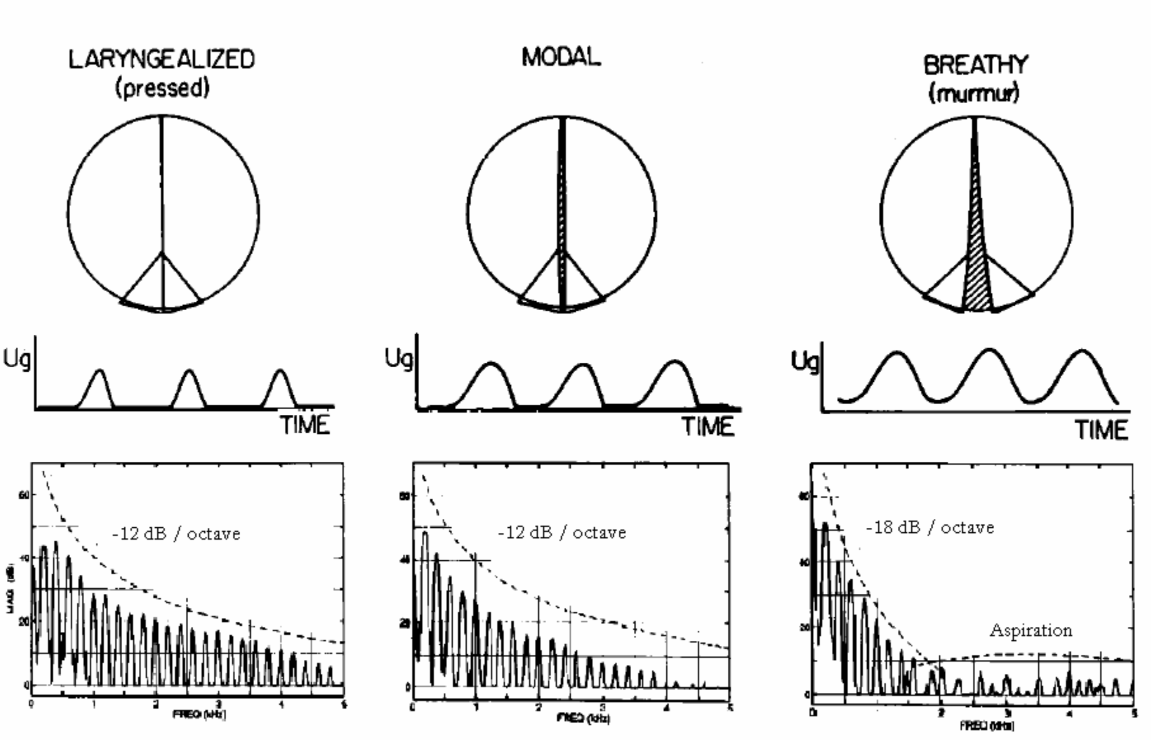

Figure 1.6 : Glottal Configuration (top pane), Volume Velocity (middle pane) and Glottal Spectrum (bottom pane) for a creaky (left column), modal (middle column) and breathy (right column) vowel. (adapted from [9])

In the third voice quality class, i.e., creaky (or laryngealized) voicing, the vocal folds do not open completely, and close more abruptly than in the modal case. This could be due to one of several reasons – physically, creakiness is usually attributed to heavier or strained vocal folds [7]. The vocal folds do not open completely and therefore, the volume of air flowing out is also usually less as compared to the other two voicing modes. The frequency of glottal opening and closure also reduces (i.e, the period, or pitch, of the source signal reduces) because of the difficulty in opening the vocal folds. The opening and closing phases are significantly shorter than the modal case, for the same reason. Spectrally, this translates to a spectrum that is flatter than the modal or breathy case. This is because of the abruptness in the opening and closing, which makes the source more impulse-like, rendering it a flatter spectrum. The typical spectral roll-off is -12 dB/octave, and the H1-H2 difference is actually negative, and less than that of the modal or breathy case. In this thesis, this mode of voicing is also classified as being irregular because the glottal opening and closure is not as it normally should be, and the pitch also does not fall in the same range it usually does.

1.3 Irregular Phonation & Its Types

In this study, we focus on one of the variations of voice quality, namely irregular phonation. This particular category is a super-class comprising of sounds that various researchers from different disciplines have called creak, vocal fry, diplophonia, diplophonic double pulsing [10, 11], glottalization [12], laryngealization [9], pulse register phonation [10], vocal fry [10] and glottal squeak [10]. While it is clear that each of these terms is well-defined in literature, there still remains some difficulty in defining the exact perceptual correlates of each of these phenomena. This is because of some common source mechanisms occurring in these various forms, which do not give a characteristic clue that identifies them from other types. However, the main characteristics of all these various forms of phonation are that the vocal folds do not vibrate as they would for a modal case, and the difference comes either in asymmetry of the behavior of the vocal folds, or a significant difficulty in opening the two vocal folds to the extent that would happen in modal case. As an example, in the case of diplophonia (or diplophonic double pulsing), the vocal folds do not vibrate simultaneously and are out of phase. Thus, the volume velocity is modulated in a two-fold way, and there are two glottal pulses in each cycle, one being delayed and damped with respect to the other. The corresponding speech signal exhibits a waveform that shows two distinct pulses exciting the vocal tract, and a spectrum that shows interharmonics arising due to the interaction of these phase-inconsistent source signals. Most pitch trackers fail to track the pitch correctly in such cases, as was tested by the author on a few samples of such speech. An extreme case of diplophonia is diplothongia, where the vocal folds do not vibrate in synchrony at all and thus, two distinct tones are produced, one due to each vocal fold.

In this work, we therefore avoid the confusion due to literary definitions, and instead rely on the work in [10] that has shown that in spite of their varying definitions, some of these phenomena are similar to each other, and are classified together by listeners to fall in the same perceptual space, thus narrowing the variety to very few classes. We thus clearly define our task as the automatic detection of all sounds that will fall in the above perceptual category, namely irregular phonation. Irregular phonation is defined for the scope of this thesis as "those sounds which fall into one of the categories of creak, vocal fry, diplophonia, diplophonic double pulsing, glottalization, laryngealization, pulse register phonation, glottal stop or any mode of phonation that arises from either asymmetric vibration of the vocal folds or due to a relatively abrupt closure of the vocal folds".

1.4 Why This Thesis?

Speakers have a certain characteristic quality to their voice, which is a consequence of their style of phonation, as well as their individual source properties. The variety in voice quality is a well-studied phenomenon, and researchers from various disciplines like speech processing, voice pathology, phonetics, linguistics and music have examined the various aspects of phonation. Detection of irregular phonation by visual inspection, and description of the characteristic features that identify it, as well as those that distinguish it from modal or breathy speech, has opened the doors to approaching one of the most important problems in speech processing – capturing the source information. This is especially true in present speaker identification systems, as studies have shown that different speakers can glottalize at different rates that are characteristic to them [12]. Traditional speaker identification and speech recognition systems have, however, focused on modeling the vocal tract, and there have been relatively few studies that try to incorporate source information into the system. One of the probable reasons for this is that there is still no set of parameters that one can use to explicitly and automatically arrive at the source information. The goal of research in the Speech Communication Lab at the University of Maryland is, in addition to other applications, arrive at a set of parameters that can identify automatically the degree and type of irregular phonation in a speaker. Since voice quality is a characteristic of the speaker, such a set of parameters would certainly boost the performance of a speaker identification system. Identification of creaky regions has been one of the tasks in this long-term goal.

In addition to speaker identification, the task of automatic detection can contribute to ASR as well. Most ASR systems rely on the traditional MFCCs, but recently, research is also being pushed into the direction of landmark-based ASR using acoustic-parameters [c.f. 13]. Identifying the source information and removing it from the speech signal to arrive at the vocal tract behavior could give insight into the actual articulators involved during the production of that phoneme. This could well lead to the recognition of the phoneme being spoken. However, in order to remove the source information, it is important to get an idea of what kind of voicing is involved and what the glottis configuration has been – the irregular phonation detector could give insight into these issues and thus help in ASR.

Automatic identification of irregular phonation can also aid in the study of voice pathology. This is an important potential application, as it can lead to non-intrusive diagnosis of laryngeal problems. It may be possible to identify the exact problem by simply processing the speech file with an algorithm that can identify how the vocal folds are vibrating and what physical activity is going on inside the larynx. While this thesis does not really focus on this particular problem, it is a step in that direction and gives sufficient thought to that possibility as being a possible future application, in terms of design of the pitch detection algorithm and other internal working details.

This work can contribute to phonetics by identifying regions of speech such as turn-taking etc., and also to the task of language identification, as there are a variety of languages that exploit creakiness and breathiness to articulate certain sounds [11]. Analyzing the rate of occurrence of certain creaky vowels can help in identifying the language of communication. Lastly, this work can also aid in automatic emotion recognition. Speakers under stress and those who have spoken for a long time, as well as physically exhausted speakers, show a tendency to go creaky in their speech. This fact can be exploited to detect if a certain speaker is under a stress of some kind. It also seems possible to use this system to detect voice imitation, as certain speakers may have certain patterns of irregularity in their phonation, and this may not be possible to mimic by the imitator.

While a number of such potential applications have been listed, this work also supports one of them– speaker identification – numerically, as an illustration. Thus, it is indeed a worthwhile task to design a system that can automatically detect irregular phonation.

1.5 Literature Survey & Contributions of this Thesis

The problem of trying to identify certain source characteristics has long been investigated by various researchers for different potential applications, and are too numerous to list with just a few references. However, there has been significantly little work done on actually arriving at irregular vocal fold activity during spontaneous speech. Further, developing this phenomenon as a means to characterize speakers is an idea that has not been investigated previously.

Published attempts at explicitly detecting creaky voicing by automatic methods seem to have been made in parallel by three other authors, all in the period of 2004-2005. In the first of these works, published by Ishi in 2004, [14], the author tries to classify segments of clean speech as belonging to either vocal fry or modal voicing. Both voiced and unvoiced regions of speech are investigated. The acoustic cues that have been proposed for the purpose are the variation of short-time power over consecutive glottal cycles, and features called intra-frame periodicity and inter-pulse similarity, based on the properties of the auto-correlation function of the voiced segments. The detection rate reported was 73.3%, with an insertion rate of 3.9%.

Independent and parallel work was published by Slifka [15] in 2005, where the task was to classify tokens of clean speech from the TIMIT corpus containing modal voicing from those containing creaky voicing. The acoustic features extracted from the voiced segments are the pitch, normalized root mean square amplitude, smoothed energy difference and shift-difference amplitude. These features have been used with a Linear SVM as a back-end, and have yielded a detection accuracy of 91.25%, with insertion error rate 4.98%.

Parallel work was also published by Yoon et al [16] in the same year, with the task being the same as the other two, namely, classification of segments of speech into modal or creaky voicing. However, the task was tougher in the sense that the corpus used was telephone speech, which exhibits significant channel distortion, and where pitch detection algorithms have been reported to fail even in voiced regions of speech [17]. Preprocessing steps were to boost the speech in order to equalize the channel effects, and then use a robust pitch tracking algorithm to detect creakiness. Fundamental frequency (F0) and spectral cues were used with an SVM to perform classification, and the classification accuracy was reported to be 75%.

The work done in this thesis, independent from and parallel with the above works, is a significantly different task because of various reasons. The first is the data that has been used to evaluate the algorithm. The idea of developing the Irregular Phonation Detector was to be able to detect and numerically characterize creakiness in spontaneous speech, which includes regions of speech with no voicing and where the speech need not be clean. Indeed, this algorithm has been tested on both clean and noisy speech corpuses, and has been found to give results approaching the best of the above approaches ([15]) in clean (no results are reported in [15] for noisy telephone speech). Specifically, in the case of noisy speech that the algorithm was tested on, various degrading elements like channel distortion, laughter and coughing, etc. have been worked on. Also, while the above works have been tested on regions of voiced speech that have been predetermined and extracted manually, this algorithm operates on speech and non-speech from all sources. While it is a trivial affair to identify regions of voicing in the case of clean speech like from the TIMIT corpus, using either the transcription labels or a sufficiently good pitch detection algorithm, the same is not the case with spontaneous speech. In the latter case, transcriptions are not available for detecting sonorant (voiced) regions, and pitch detection algorithms have been reported to fail detection of voicing [17] and detect voicing in regions where there is no voicing in reality (observed by author on various occasions). Even other simple and well-established acoustic features like the zero crossing rate, energy thresholds etc. fail when they are used to identify voiced from unvoiced regions in spontaneous speech from telephone data. All these facts are illustrated in the Figure 1.7 (next page), which demonstrates the temporal behavior of some of these features in telephone speech.

The figure 1.7 shows a region of laughter (/ehehe/, boundary marked by red lines), where the region is partitioned into five regions: the first, third and fifth regions are voiced and breathy (/e/) and the second and fourth are aspirated. These two regions should be separated from each other by an appropriate voiced/unvoiced detector. Further, the detector must be able to separate unvoiced regions from creaky regions (which are voiced). It may be noted that the pitch detector of the ESPS software does not do a good job at detecting all the creaky regions, and thus cannot be used for purposes of this thesis. Further, the zero crossing rates and the energy are also not very useful in separating out these regions. For example, in regions 1,2 and 3, wherein 1 and 3 fall in the voiced category and 2 in the unvoiced, the zero crossing rate it seen to be in the same range. Further, even the short-time energy parameter in regions 1 and 2 are in the same range. Thus, no statistical framework can be used in this case to separate the voiced regions from unvoiced. Further, it may also be seen that the energy parameter in the unvoiced regions (which is typically expected to be low compared to voiced regions) is significantly higher than the voiced regions towards the end, where creakiness sets in. This kind of behavior is very common in spontaneous speech, and it is simply impractical to use any of these measures in a simple way and expect good separation between the voiced and unvoiced regions.

In this work, not only has the creakiness detection algorithm been designed to handle noisy speech containing distortions, as well as non-voiced parts of speech that would be identified as voiced by most current voicing-detection approaches, but an accurate pitch detection algorithm has also been incorporated into an Aperiodicity, Periodicity and Pitch (APP) detector [18]. An illustration of its performance may also be seen in figure 1.7. The ESPS pitch tracker clearly fails to track the correct pitch in creaky regions, and in regions with modal regions followed by creaky, it gives wrong pitch estimates for one of these tow modes of voicing. The APP detector pitch tracker, however, is seen to perform a decently better job at tracking the pitch, and fails to find only one voiced region – the breathy region 5. It does not give the correct pitch value in some frames in one of the creaky regions, and shows a doubling error, but still, it could be said that on an overall basis, this modified pitch tracker shows greater accuracy than the ESPS pitch tracker. This modified pitch detection algorithm has been found to give pitch detection accuracy of 98.3%, compared to detection accuracy of 95.4% of the earlier version of the APP detector. The irregular phonation detector gave a detection accuracy of 91.8% with an insertion rate of 17.4% instances, or frame-wise, 5.6%. The irregular phonation detector was incorporated into the APP detector system. Further, as a demonstration of the speaker-distinguishing capability of the parameter thus developed, speaker identification experiments have been performed.

1.6 Acoustic Parameters for Speaker Identification

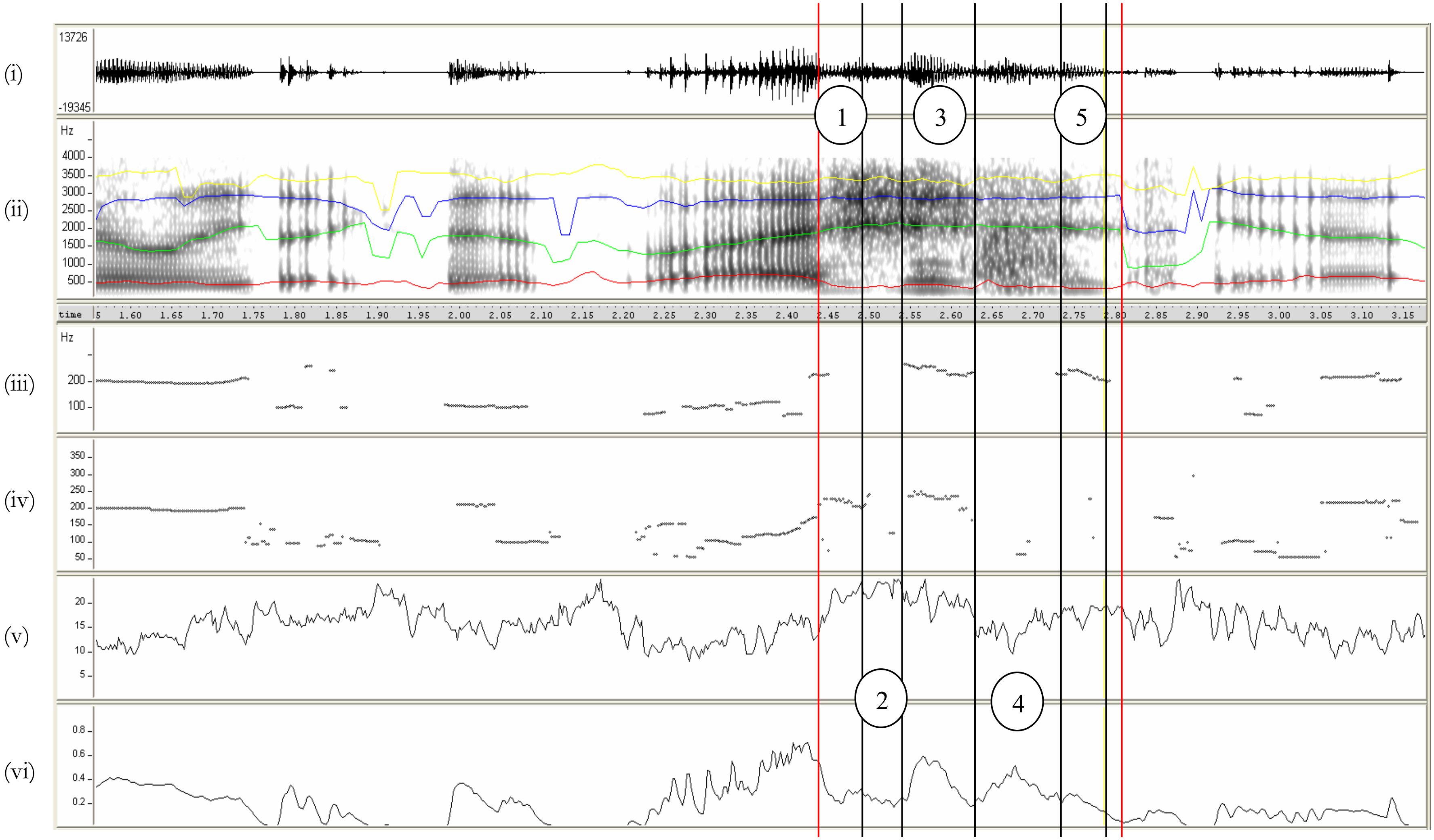

Figure 1.7 : Illustrative example showing the behavior of the pitch, zero crossing rate and energy in a speech file. (i) the time signal, (ii) spectrogram with the formants overlaid, (iii) pitch track from the ESPS software, (iv) pitch track from the modified APP detector, (v) zero crossing rate, (vi) short-term energy. The red lines delineate the boundaries of a region of laughter in the speech file. The black lines separate the breathy voiced regions of laughter from the aspiration regions.

While traditional speaker identification systems rely on the vocal tract dynamics and under-emphasize the significance of the source, more recent work has shown (c.f. [19]) that the addition of source information can prove to be valuable speaker-specific information. The set of knowledge-based APs proposed for speaker ID includes four parameters (the amount of periodic and aperiodic energy in the speech signal, the spectral slope of the signal, the amount of creaky energy in the speech signal) related to source information complementing four parameters characterizing the vocal tract behavior four formants (F1, F2, F3, F4) [6]. Text-independent speaker identification experiments were performed using this feature set and the standard MFCCs for populations varying from 50 to 250 speakers of same gender.

The relative amounts of periodic and aperiodic energy differ considerably depending upon the voice quality. For modal voice, sonorant regions (vowels and sonorant consonants) are strongly periodic with little if any aperiodicity. For breathy voice, sonorant regions are periodic at low frequencies with some aperiodicity in the region of F3 and the higher formants [9]. Finally, for creaky voice, there is irregular vocal fold vibration so that the irregular phonation detector (now incorporated into the APP detector) finds creaky energy. Speakers differ not only in the voice quality used to produce sonorant sounds, but also in the tradeoff between the supraglottal turbulent source and the glottal voicing source used to produce voiced fricatives. The differences in the amount of periodic, aperiodic and creaky energies, due to various factors, is captured by the APP detector. Another measure used to capture information about the source is the spectral tilt of the speech signal, as explained in an earlier section.

In addition to the various measures used to characterize the source, the formant frequencies F1 through F4 are used to characterize the vocal tract. The frequency range over which these formants vary is an indication of vocal tract size and shape. F3 characterizes the length of the vocal tract of the speaker, and hence an average measure of F3 helps in capturing this speaker-specific information. F4 provides the higher frequency characterization of the vocal tract. Finally, it has been speculated that F4 (sometimes called the singer's formant) in vowel regions is attributed to the resonance of the laryngeal cavity [20]. Given the relative stability of F4 in running speech for a particular speaker and its variance in frequency across speakers, it may be an additional measure of voice quality [21].

1.7 Outline of the Thesis

Chapter 1 provides a brief introduction to speech technology and the speech production process, and motivates the use of the acoustic parameters for various speech technology applications. This chapter describes in detail the source mechanisms in voiced speech, as well as irregular phonation. The motivation behind this work is then described, and the contributions of this thesis are highlighted.

Chapter 2 briefly describes the APP detector system, and explains the failure modes of the pitch detection algorithm in regions of irregular phonation, which also hinders the detection of such regions. Solution to this problem is then motivated, and the chapter ends with a description of appropriate signal processing parameters for the APP system.

Chapter 3 details the acoustic characteristics that are essential to detect creakiness, and motivates their use. The correction of pitch, and detection of creakiness, is then discussed.

Chapters 4 discusses the results of the creakiness detection and pitch detection experiments, and demonstrates the use of creakiness as an acoustic parameter for speaker identification. This chapter also discusses the reasons for errors and the influence that some of these errors may have on the other tasks.

Chapter 5 discusses ways to improve performance and to deal with channel variations. The interplay of the acoustic parameters, and possible extension to the realms of ASR and other applications is then discussed.

Chapter 2

MODIFICATIONS TO THE APP DETECTOR

The algorithm for the detection of irregular phonation can process an input speech file and automatically identify regions of the signal where the speaker exhibits irregular phonation. The algorithm is an extension of the Aperiodicity, Periodicity and Pitch (APP) Detector [18], a system that processes a speech file to give a spectro-temporal profile indicating the amount of aperiodicity and periodicity in different frequencies with time. The APP detector exploits the behavior of the Average Magnitude Difference Function (AMDF) to arrive at decisions of periodicity and aperiodicity for each frame. A complete description of the APP detector is given in [18]. In this chapter, the system is first briefly described for the sake of completeness, and then the reasons for the failure of the APP system to detect pitch in regions of irregular phonation are discussed. The output of the APP detector for different kinds of regions of speech is then discussed, and the current behavior of the APP detector for regions of irregular phonations is shown, highlighting the goals of this thesis. Finally, working parameters for using the APP system are recommended.

2.1 The APP Detector

The APP detector is a time domain algorithm that, at a broad level, makes a decision about how much voiced and aperiodic energy is present in a signal. To be more precise, this algorithm estimates 1) the proportion of periodic and aperiodic energy in a speech signal and 2) the pitch period of the periodic component. While most of the algorithms used to detect aperiodicity are passive, i.e., aperiodicity is considered as the inverse or lack of periodicity, they are prone to errors in situations where the signal has simultaneous strong periodic and aperiodic components. The APP detector is thus particularly useful in situations where the speech signal contains simultaneous periodic and aperiodic energy, as in the case of some voiced fricatives, breathy vowels and some voiced obstruents. Figure 2.1 shows the block diagram of the APP detector:

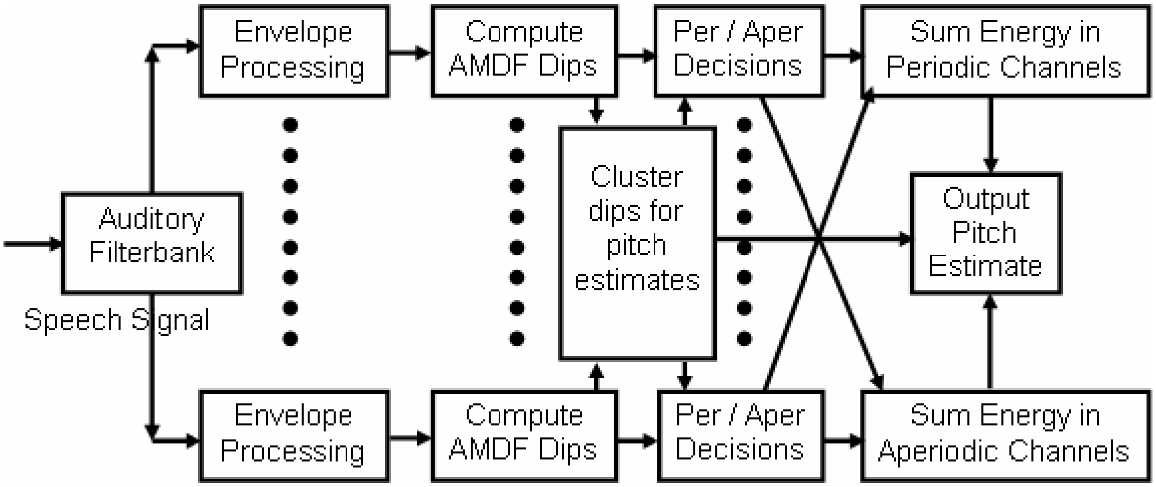

Figure 2.1 : Block Diagram of the Aperiodicity, Periodicity and Pitch Detector

The first block in the APP Detector is an auditory gamma-tone filter-bank that splits the channels into 60 frequency bands, with the upper-most center frequency being defined by the sampling rate. The choice of the filter-bank is in order to capture the perceptual importance of different frequencies. The outputs of the higher frequency channels (channels centered above 300 Hz) are then smoothed using the Hilbert transform to extract the envelope information and remove the finer structure. The next stage of the APP Detector incorporates frame-by-frame silence detection by applying a threshold to the signal energy in each channel. The energy is first normalized by the max energy in the speech signal, and then an empirically determined threshold of 0.0015 is used.

Following this, the signal in each channel is then processed to obtain the frame-wise Average Magnitude Difference Function (AMDF) to identify the amount of periodicity or aperiodicity in the signal. The signal is analyzed on a frame-by-frame basis. For each frame, a windowed portion of the signal, centered at the frame center, is used to compute the AMDF. The windowed signal is subtracted from a windowed portion of its neighbor located at a lag of *k* samples. The resultant signal is added to get a single number. For each such lag, the AMDF gives a number, and thus, for each frame, one obtains the AMDF as a function of the lag *k*. Mathematically, the AMDF $\gamma_n[k]$ of a signal $x[n]$ is defined as

$$\gamma_n[k] = \sum_{m=-\infty}^{\infty} \left| x[n+m]w[m] - x[n+m-k]w[m-k] \right|$$

where $w[n]$ represents a rectangular window centered at n and having a width as required. The conceptual idea behind the AMDF is the same as that of the Auto-Correlation Function (ACF) - signals which have periodicity content will yield an ACF that shows peaks at locations that are periodic with period equal to pitch frequency. However, the ACF involves multiplication, and is thus computationally expensive when calculated for a large number of lags. The AMDF is a computationally less expensive substitute for ACF, in that it incorporates subtraction instead of multiplication. The AMDF behaves like the ACF, with

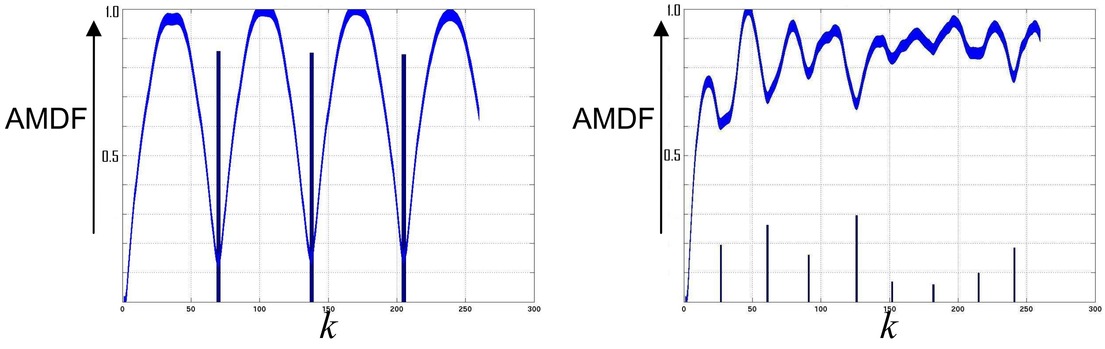

Figure 2.2 : Sample AMDF and dip strengths for a strongly periodic (left) and aperiodic (right) frame

one difference – with varying lag *k*, the periodic signal causes dips instead of peaks at locations corresponding to pitch period and its multiples. This is quantified in the APP detector by calculating the dip strengths using the convex hull of the AMDF.

Figure 2.2 above contrasts a sample AMDF function from a single channel of a strongly periodic signal with that of a strongly aperiodic signal. The vertical lines superposed on the figure represent the strength of the dips. The noteworthy features are that the AMDF shows strong dips at lags equivalent to the pitch period and its multiples in the former case, while the dips are significantly weaker and randomly distributed in the latter case. Decisions of periodicity and aperiodicity are made by summarizing this trend across all channels. For a periodic frame, it is expected that all non-silent channels will exhibit a similar trend in the AMDF dips, due to the underlying periodicity that occurs due to the glottal source. Thus, when the dips are summarized across all channels by addition, the dips will all cluster tightly together at lags equaling the pitch periods and its integer multiples, and give a significant strength of dips. If this is the case for a particular frame, then that frame is classified as being periodic *(per)*. Channels that contribute to this periodicity profile will then be called periodic channels, and other non-silent channels are called aperiodic *(aper)*. For an aperiodic frame, on the other hand, the dips, when summarized, will display a random behavior as a consequence of the individual channel behavior. The summary measure of periodic and aperiodic content is obtained by multiplying the frame *per/aper* decision by its energy and then adding it across channels. Thus, for each frame, a decision of *per/aper* is made for the frame, its individual channel-wise *per/aper* profile is produced, and the amount of aperiodic and periodic energy is also obtained. The entire speech file is processed this way, to get frame-wise output.

2.2 Dip Profiles in the APP Detector

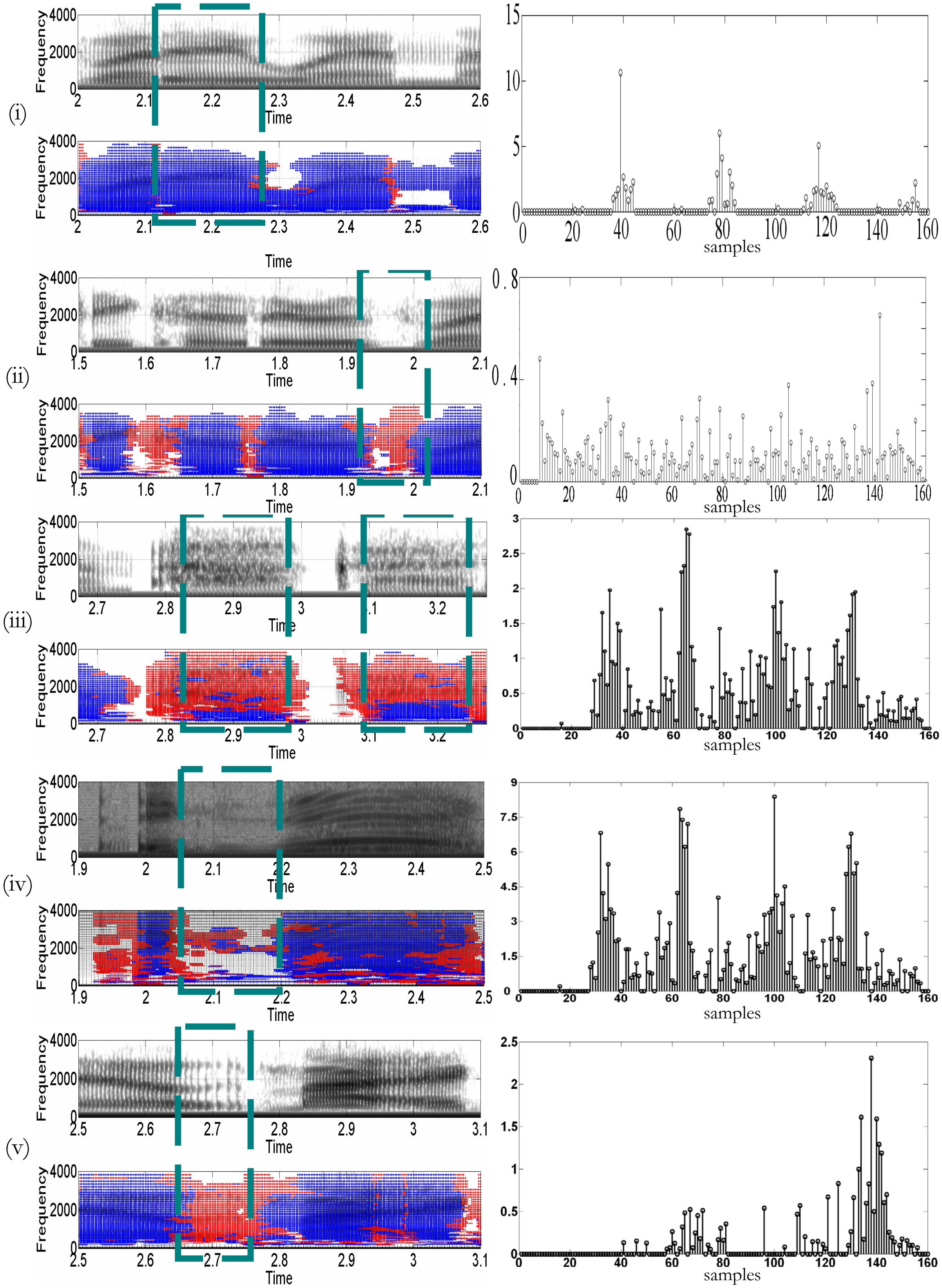

Figure 2.3 : (left) Spectro-temporal profile of periodic energy (blue) and aperiodic energy (red) generated by the APP detector for a (a) periodic vowel (b) aperiodic (c) breathy vowel (d) voiced fricative and (e) creaky vowel region. The relevant regions are highlighted in the dotted box. (right) Corresponding dip-profiles for a sample frame of the respective regions.

The key to detection of regions of irregular phonation, and pitch correction in such regions, is the dip summary or profile in those frames. Figure 2.3 demonstrates the dip profiles for the cases of a single frame of a vowel, fricative, breathy vowel, voiced fricative and a creaky frame. It may be noted that these profiles are obtained by adding together the dip strengths across all the channels for each frame. The vowel, being a periodic signal, shows strong dip clusters at multiples of pitch period. The maximum strength of the dip profile is noteworthy, as well as the distribution of the dips. Typically, this value is greater than 10 for periodic frames, because of the fact that most channels are periodic in nature and thus show the AMDF dip in the same location, giving an added strength of at least 10. Because of the well formed strong clusters, the decision is that the frame is periodic. The channels that contribute to this dip cluster are all called periodic, while the others are called aperiodic. Since most channels in periodic channels also show a periodic AMDF, thus, the APP detector spectro-temporal output is seen to be periodic (blue) in most of the vowel regions. The fricative, on the other hand, displays a random distribution of dips, which have a very low strength even upon summary across channels – this is clearly because the dips do not cluster to add up together. Thus, the typical value of the maximum dip in the summary profile is less than 1.5. Due to the lack of strong clusters, the APP detector decision is that the frame is aperiodic, and since the channels do not show any periodic behavior that contribute to any clusters, the APP spectro-temporal profile shows mostly aperiodic (red) behavior. In the case of the breathy vowel and the voiced fricative, the dips show a mixed behavior – there is clustering of some dips (which is due to the voiced (glottal) source at the low frequencies), and there is also some randomness in distribution of some other dips (which is due to the turbulence caused by the supra-glottal source, dominating at the higher frequencies). Thus, the strength of the dips is less than that of the periodic frame, but greater than that of the aperiodic frame. Typical values of the dip maximum lie between 1 and 10. The corresponding decision made by the APP detector depends on the strength of clustering of dips. If the dips are clustered strongly enough, the decision is that the frame has some periodicity, and the channels contributing to the periodicity are marked as periodic in the spectro-temporal profile. The channels that do not contribute to the periodicity are called aperiodic, and they are identified in red in the spectro-temporal profile. As is expected, most of the periodic energy in the APP output profile is located at the lower frequencies, and most of the aperiodic energy is located at the higher frequencies.

Finally, the case of the creaky frame is midway between that of the periodic frame and the aperiodic frame. It lacks the tightly packed dip structure of the periodic frame, but also lacks the randomness of the dip structure of the aperiodic frame because of an inherent structure in the voicing source. In the case of breathy vowels and voiced fricatives, the clusters are not strongly prominent, but are still tight in the sense that the dips that belong to periodic channels are located in a cluster that is very narrow. This is the reason why the APP detector identifies such regions as being periodic. However, for the creaky vowel, the clusters that are formed by the dips are much broader than would be for the periodic frames (the reason for this is explained shortly), and thus, the APP detector does not identify any "periodic" (or more precisely, phonation) structure in these dip profiles. The lack of tight clusters and the weakness of the maximum dip (which is usually around 1.5 and thus comparable to that of the aperiodic frame) cause the APP detector to make a decision of aperiodic for all creaky frames. As the corresponding frames' channels are also called aperiodic, the spectro-temporal profile in such cases would show primarily aperiodic energy (red), as shown in the figure 2.3.

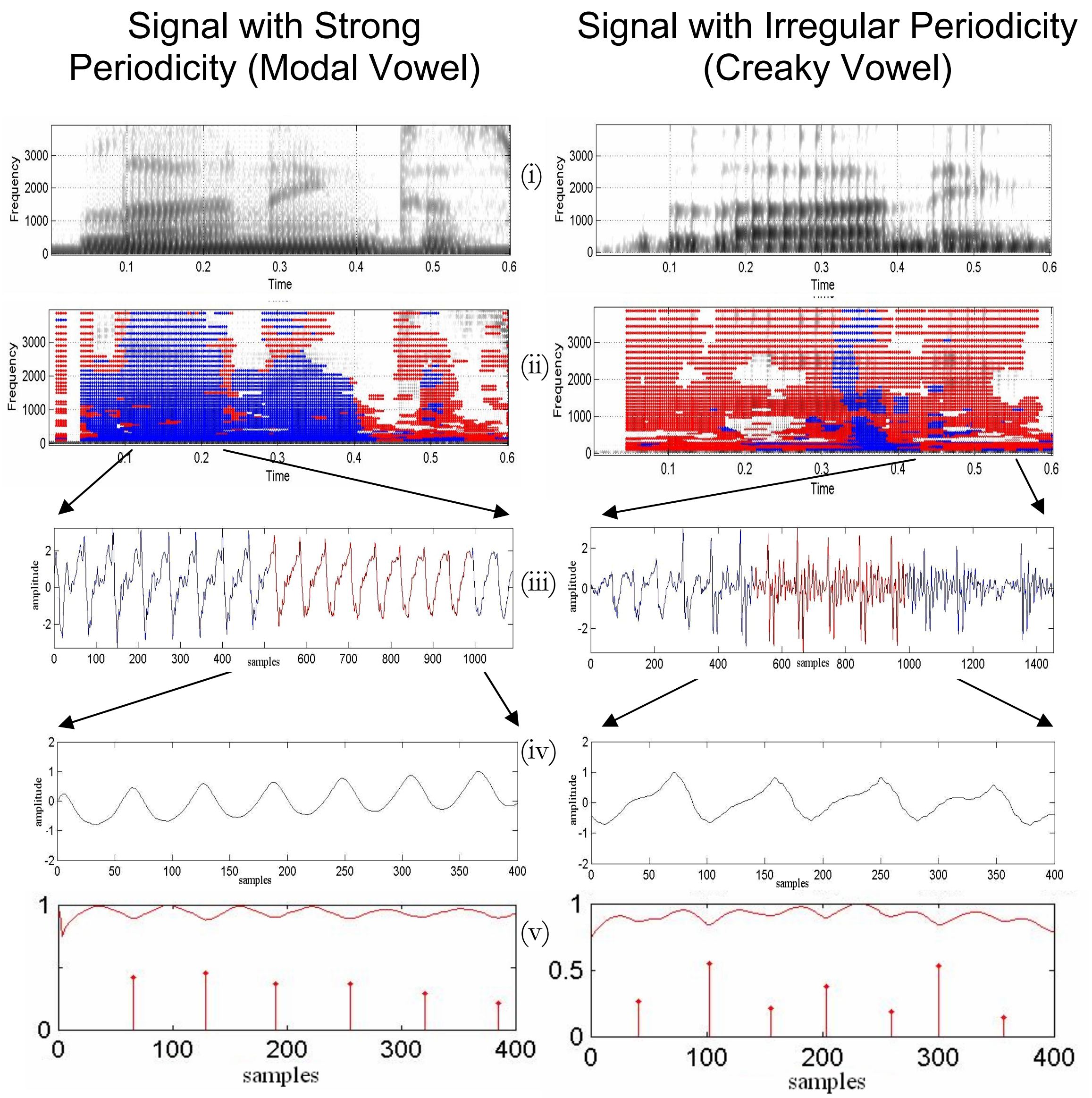

Figure 2.4 : A sample modal vowel and creaky vowel compared in terms of (i) spectrogram, (ii) periodicity (blue) /aperiodicity (red) profile, (iii) time waveform, (iv) source or glottal signal exciting the vocal tract and (v) AMDF of the signal and associated dip structure. The vowels being compared are the same (/e/)

The root of the behavior of the APP Detector for frames with irregular phonation can be traced to the AMDF structure of such frames. Figure 2.4 shows the time waveform, the corresponding glottal waveform obtained by the Inverse Filtering technique [22] and the corresponding AMDF obtained by the APP Detector for one frame of the modal and creaky vowels shown in Figure 2.4. The most striking point here is that the AMDF closely captures the information in the glottal waveform – this is evident in the relative position of the peak of the glottal waveform and dips of the AMDF in the case of the modal vowel. This correspondence is not so exact for the creaky vowel, owing to the fact that the pitch shows some jitter and thus, the AMDF does not have exact alignment of the signal and its delayed version to give dips at exactly the pitch period. This causes the AMDF to have dips that do not cluster properly, and are not in alignment with their counterparts from other channels. Further, these dips are not as strong as in the case of periodic frames, due to the fact that the irregular vibration of the vocal folds results in a glottal pulse that has lower amplitude than modal phonation. The dips may also not be of equal strength. For example, in certain cases like diplophonia, where the glottal pulse consists of a main pulse followed by a delayed pulse reduced in amplitude, the AMDF shows two kinds of dips corresponding to each of the two different pulses – one stronger than the other, and each repeating after periods corresponding to their respective pulse. In all such cases of irregular phonation, the dips are of different strengths and at different locations. A summary measure of the dips across channels exhibits a dip profile significantly different from the modal vowel, due to the lack of periodicity and varying dip strengths at different lags.

Owing to these differences in the glottal pulse of modal and irregular phonation, the summarized dip profile of irregular phonation exhibits dip clusters that are existent but not very well-defined, and at locations not equal to integral multiples of the pitch. It is not surprising that the APP Detector calls such frames aperiodic – these frames do indeed exhibit a profile that does not show tight clustering and has an overall dip strength comparable to the dip-profiles of aperiodic frames. The goal of this thesis is to modify the decision process of the APP detector in such a way that it accommodates the loose clusters of the creaky vowel, but is still able to identify the fact that tight clusters imply presence of periodic (modal phonation) energy and lack of clusters implies aperiodic energy. Thus, the spectro-temporal profile of the APP detector must change from one that looks in 2.5 (b), to one that looks in 2.5 (c), where the green region represents a region of irregular phonation, as opposed to a blue region that represents periodic energy due to modal phonation and a red region that represents aperiodic energy due to turbulence.

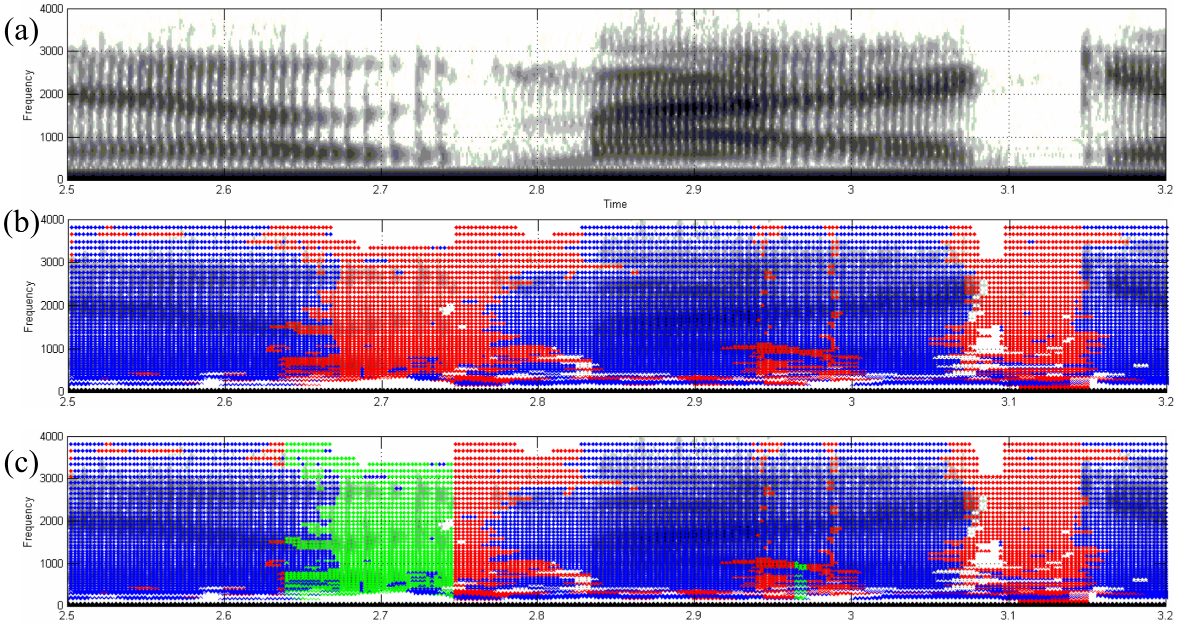

Figure 2.5 : (a) Spectrogram of a segment of speech containing a creaky vowel, (b) Spectro-temporal profile of the old APP detector, which identifies the creaky vowel as being aperiodic, (c) Spectro-temporal profile of the modified APP detector, which identifies the creaky vowel as having irregular phonation and classifies it as being separate from aperiodicity due to turbulence

2.3 Recommended Specifications

The APP detector has been designed to operate at various frame rates and window sizes as desired in potential applications. However, in order to detect creakiness, the set of specifications used should satisfy certain criteria, because of the inherent acoustic characteristics of this phenomenon. In particular, since it is known that the pitch falls relatively low during creaky regions, sometimes as low as 60Hz, it must be ensured that the window size is large enough to accommodate at least one pitch period – this will ensure formation of a loose cluster that can then be used to detect creakiness. When the sampling frequency is high, the window size must be large; when low, the window must be smaller. Typically, it is useful to have a window size 1.5 times that of the minimum pitch frequency expected – this will ensure proper formation of a cluster.

Creakiness usually manifests itself for durations of less than 20 msec. Typical duration of creakiness in the TIMIT corpus is about 10 msec. Thus, if it is desired to properly identify the *region* of creakiness and not just an instance, it is necessary to have a high frame rate. If the frame is displaced by 2.5 msec, then for a 10 msec creaky region, four frames will be called creaky and thus identify a region. If the frame were displaced by 10 msec (as is the typical frame rate for most speech processing applications), there would only be one frame which would be called creaky. This is not a very good idea, because finding just one frame as creaky does not give a method to eliminate any false alarms. If the frame rate were high, then any spurious locations of detection could be median-filtered and removed. In the approaches mentioned in the literature survey, features are extracted by moving the window over one sample, thus giving an extremely high frame rate. In the case of the APP system, a frame rate of 2.5 msec is recommended in order to reduce computational load and at the same time, ensure frames close enough to detect regions of creakiness and remove spurious false alarms.

Chapter 3

ACOUSTIC CHARACTERISTICS FOR DETECTION OF IRREGULAR PHONATION

Detection of regions of irregular phonation is not a trivial problem, and when potential applications include processing speech from real-time corpuses, provision must be made to either counter the influences of corrupting factors, or render the algorithm invariant to such factors. In this thesis, the algorithm has been designed to work on speech from telephone data and no prior assumptions are made about knowledge of regions of voicing or characteristics of channel. The general approach to the problem has been to first detect suspect regions of irregular phonation, and then to identify if such regions do indeed exhibit voicing, and if so, if the voicing is due to creakiness (voicing might, otherwise, be in the case of voiced fricatives or breathy vowels, and show a more periodic structure). This is done by taking into account a number of acoustic characteristics that could deliver hints about voicing and creakiness. Conventional approaches, namely the zero crossing rate and energy threshold, are used to separate the voiced regions from unvoiced. These prove inadequate to make a clear distinction between the two, as demonstrated in Figure 1.7. Therefore, the dip profile is also used to make the voiced/unvoiced decision. At this stage, the dip profile is tested for whether it is due to creakiness or due to voicing in breathiness or voiced fricatives. Once creakiness is suspected, an estimate of the pitch is made from the dip profile. Following this, creakiness is reconfirmed by using the pitch estimate to identify the confidence of creakiness in that dip profile. At the end of this stage, the decision of whether the frame is creaky or not is made. This entire procedure is shown in the following flowchart:

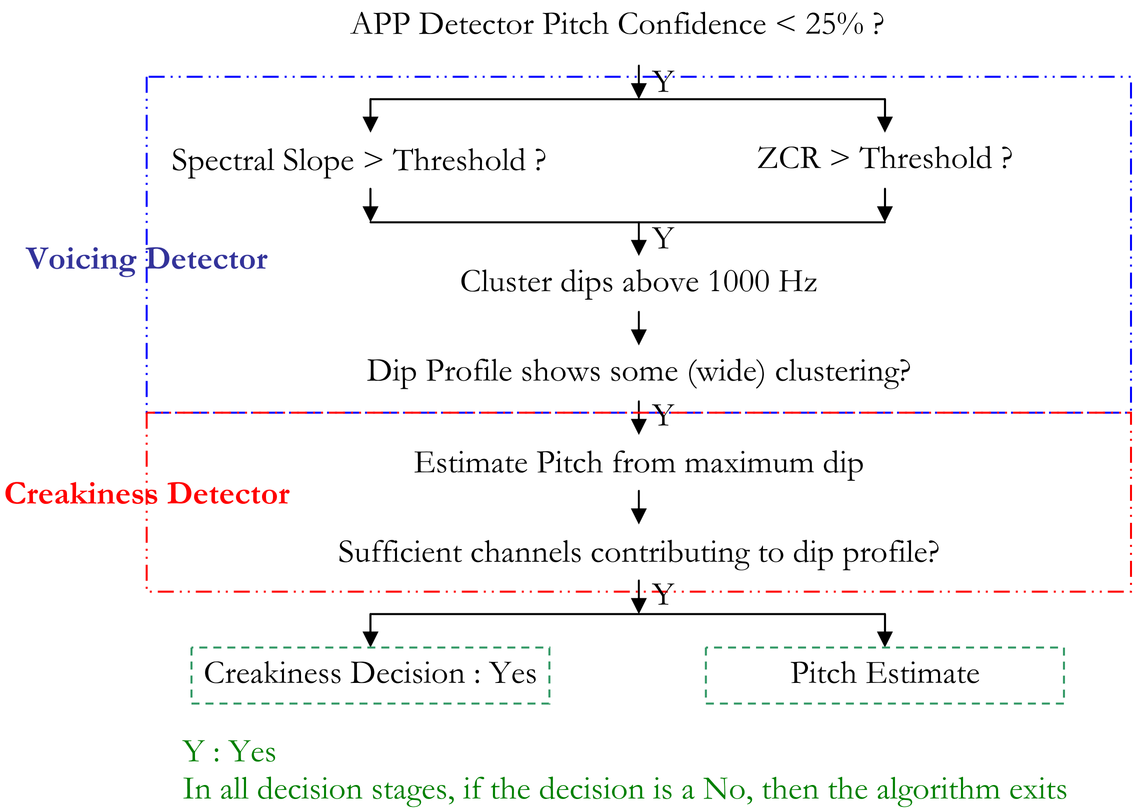
APP Detector Pitch Confidence < 25% ?
Y
Spectral Slope > Threshold ?
ZCR > Threshold ?
Y
Voicing Detector
Cluster dips above 1000 Hz
Dip Profile shows some (wide) clustering?
Y
Estimate Pitch from maximum dip
Creakiness Detector
Sufficient channels contributing to dip profile?
Y
Creakiness Decision : Yes
Pitch Estimate
Y : Yes
In all decision stages, if the decision is a No, then the algorithm exits

Figure 3.1 : Flowchart showing the main stages of the algorithm for detection of irregular phonation

Each of the stages in this flowchart, and the motivation leading to their inclusion in the algorithm, is elaborated in this chapter.

3.1 Detecting Suspect Regions of Irregular Phonation

The APP detector consists of a pitch detector that looks at the dip profiles, and then gives voiced/unvoiced decisions and pitch estimates depending on the clustering of dips in that particular frame. If the frame is judged as voiced, then the main cluster is said to be that cluster which has a maximum, and which was used to estimate pitch. The channels which contribute to the main cluster are then identified, and the (normalized) energy of all these channels is added together to obtain what is called the pitch confidence. Thus, the maximum possible pitch confidence would be equal to the number of channels present. In periodic frames, the pitch confidence is typically greater than 50% of the maximum possible. In regions of creakiness, voiced fricatives and breathy vowels, where there is some voicing but clustering is not as strong as that of the periodic frames, the pitch confidence lies between 10% and 40% of the maximum possible. In creaky frames specifically, the pitch confidence is seen to be rarely greater than 25% of the maximum. For purely aperiodic frames, the pitch confidence is usually near zero and does not exceed 10% of the maximum. Thus, regions having pitch confidence greater than zero and less than 25% of the maximum possible are identified as suspect regions where creakiness could be possible.

3.2 Energy Threshold

Using an energy threshold is a common strategy applied for speech detection in many speech processing applications. In real and spontaneous speech data as in the case of telephone speech corpuses, the presence of background noise and instrument noise due to telephone microphone and recording instruments cause the existence of a non-zero random signal in the recorded speech signal. This random signal component can sometimes demonstrate a dip profile that appears very similar to that of irregular phonation – this is purely a coincidence and not due to voicing. Such instances of dip profiles must be eliminated from further analysis since they may be called creaky if subject to the dip profile analysis. A simple solution would be to apply an energy threshold that would suffice to separate speech from non-speech. In the original APP detector algorithm, a threshold of 0.015 was empirically found to be apt for such separation, and it seems to suffice as a threshold for creakiness as well. It was also seen that this threshold does an appreciably good job in separating some weak fricatives, glides and consonants from voiced regions.

3.3 Zero Crossing Rate

Another common strategy to separate fricatives from regions of voicing is to exploit the zero crossing rate (ZCR). Fricatives, having a more noise-like random nature, have a higher ZCR than voiced regions. Thus, in typical clean speech processing applications, the ZCR proves to be a good candidate for detection of voicing. However, in case of telephone speech, ZCR is not a good candidate. It is observed that fricatives can sometimes have the same ZCR as voiced regions. This is due to the channel and device noise that rides the voiced regions, causing the ZCR to increase and approach that of the fricatives. Figure 3.2 demonstrates one such case, where the ZCR for a fricative is different from that of one voiced region, but is about the same as another voiced region. Thus, one needs a more complex analysis for using the ZCR than merely using a fixed threshold.

It has been observed that in case of laughter and coughing, unlike the case of fricatives, noise is spread over all the frequencies and has an approximately flat spectrum, except at the really low frequencies (< 300Hz). Therefore, if the ZCR should be used, there should be a provision for accounting for the spectral spread of various phenomena – the noise due to fricatives (which occurs at frequencies above 2000 Hz) should be treated the same as that due to laughter (which occurs at almost all frequencies). However, if the ZCR were merely used taking all frequencies into account, the fricatives would show a lower ZCR than laughter, because laughter would consist of high frequency noise (equivalent to fricative noise) riding on a low frequency noise. In order to treat these phenomena at the same level

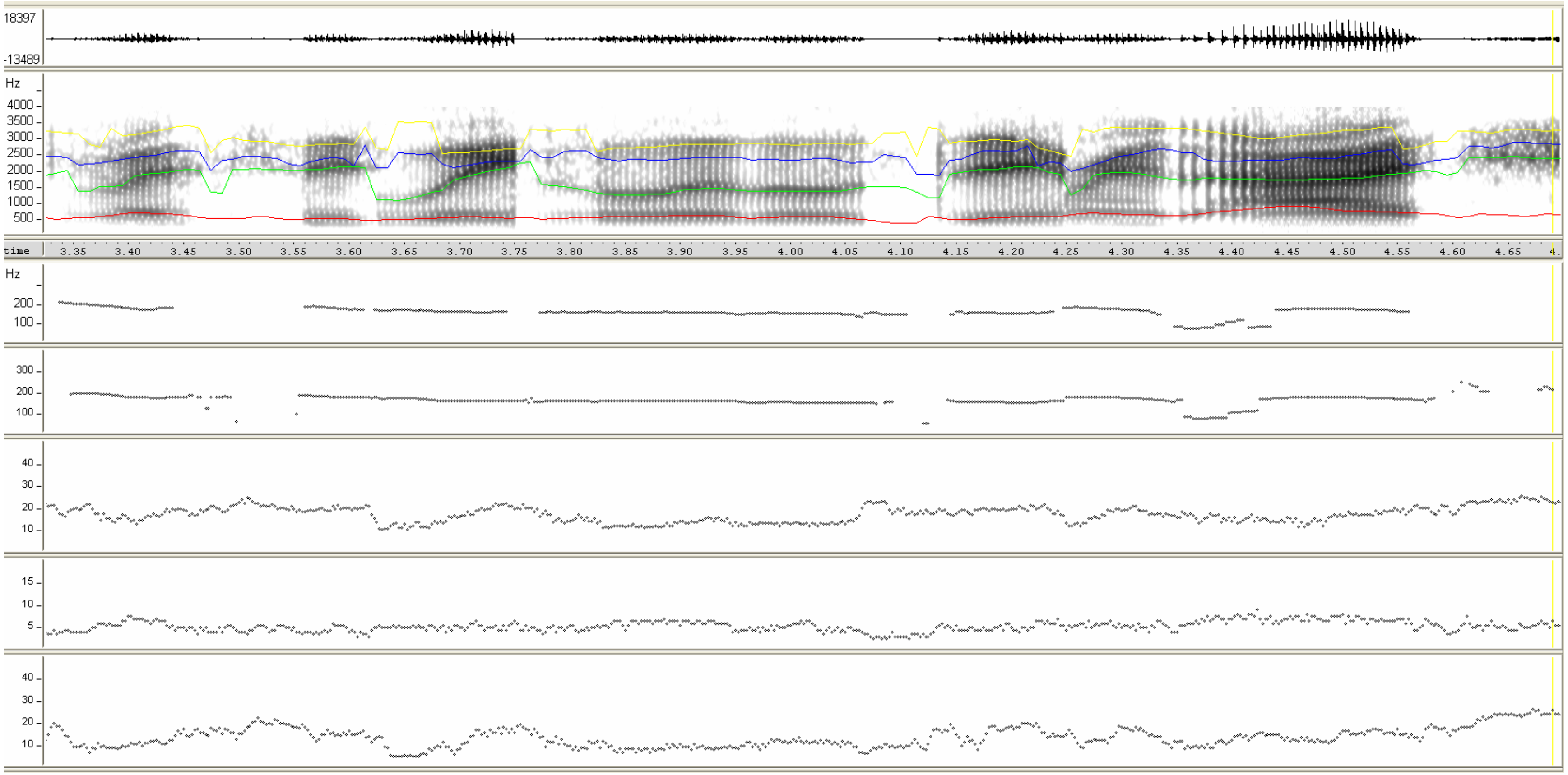

Figure 3.2 : Illustration of the zero crossing rate in case of the voiced and unvoiced regions. Here, the 3rd pane is the zero crossing rate (ZCR) for the signal components > 2000Hz, and the 4th for the signal components < 2000 Hz. The last pane shows the ZCR of the speech signal, with all its frequency components. It may be seen that the ZCR for the fricative in region 3.45 to 3.55 sec is the same as that of the vowel in region 3.55 to 3.6, and 4.15 to 4.25 sec. These, in turn, are significantly different from the ZCR in the vowel region between 3.85 to 3.95 secs.

for ZCR processing, it is necessary to account for frequency of extent of these phenomena.

A further problem that is encountered is that the ZCR of creakiness is very close to that due to fricatives. This is explained as follows: due to the significant damping of the vocal tract impulse response in creakiness, the speech signal becomes weaker in amplitude, and thus more susceptible to noise, than the modal voiced signal. Since the pulses exciting the vocal tract are placed far apart in creaky voicing, the vocal tract sometimes damps down completely and thus the signal has a speech component equaling that of the noise component. Thus, the noisy characteristics of the channel and recording device render the creaky signal having a ZCR that is nearer to that of the fricative.

With these factors determining the behavior of the ZCR, a possible solution seems to split the signal into separate frequency regions and compare the ZCR in these spectral bands, rather than across the spectrum. This has a two pronged advantage. The first is that it gives a good way of comparing the ZCR due to frication to that due to laughter – for example, by looking at frequencies above 2000 Hz in both cases, these noise factors are both brought into a common platform for comparison with other phenomena like voicing etc. The second is that if the frequencies below 2000 Hz alone are used, the noise riding on the weak transients of the creaky voicing (which typically have great high frequency content, because of its "white" spectral nature) are filtered out significantly, and the resulting signal is more free of the imposing channel and device noises. Thus, noise due to laughter will now be better discernible from the noise in creaky voicing, the latter having been considerably reduced after filtering. The choice of 2000 Hz is motivated by the fact that strident fricatives show spectral energy at frequencies above 2000 Hz, and thus, the ZCR for fricatives could be attributed to being caused by spectral bands above this frequency. The temporal behavior of the ZCR obtained after filtering the signal for frequencies above 2000 Hz (called herein the high-spectrum ZCR) for a fricative, modal vowel, creaky vowel and laughter, is shown in the figure 3.3. It is seen that the high-spectrum ZCR is different for the vowel regions (both modal and creaky), from that of the fricative and laughter regions. In addition, the fricative and laughter regions have a ZCR that fall in the same range, while the creaky vowel and modal vowel have a ZCR falling in the same range. An additional point to note is that the speech sample from the two figures are taken from two different speakers – the creaky samples from two different speakers are seen to be falling in the same range as well. It might be the case that the high-spectrum ZCR is a speaker-independent parameter that can be used with reasonable success for the task of separation of voiced regions from unvoiced regions.

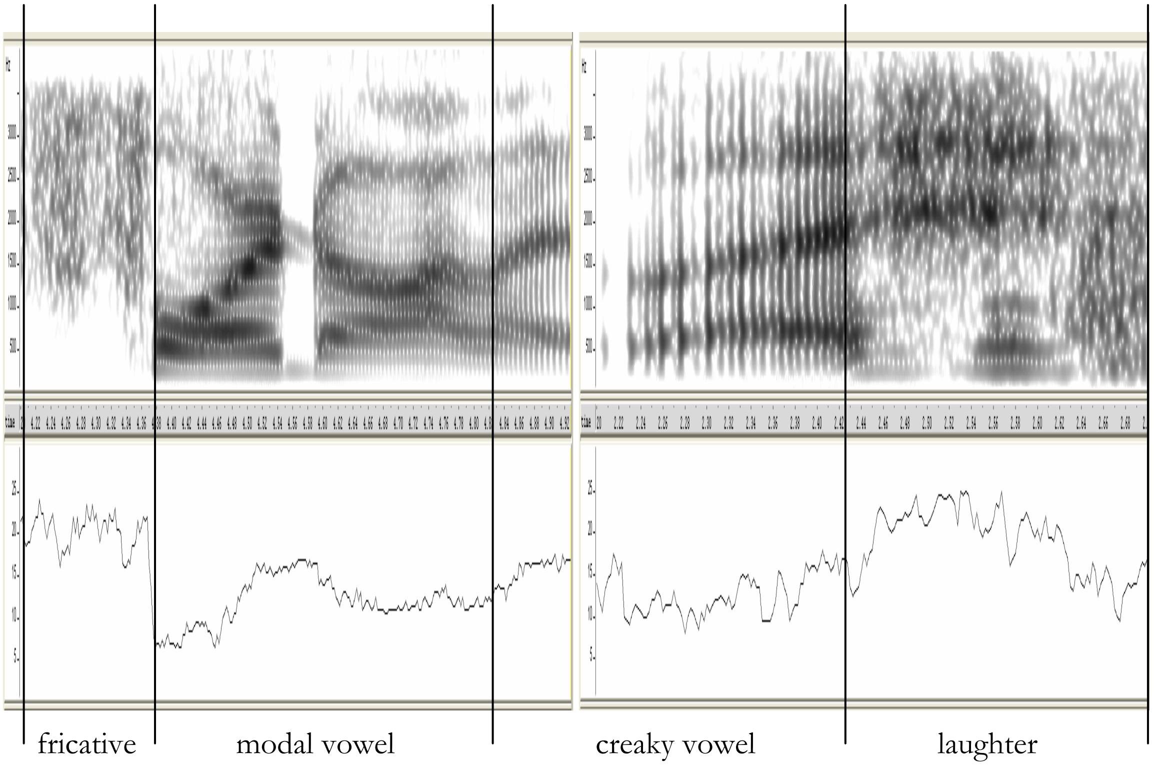

Figure 3.3 : Spectrogram and high-spectrum ZCR of a region of speech, showing a fricative, modal vowel, creaky vowel and laughter. Approximate boundaries of these regions are marked with vertical lines. The speech file on the left and right are from two different speakers.

In the algorithm, motivated by the above reasons, three kinds of ZCR have been used – the full-spectrum ZCR (which gives a gross level separation from some undesirable sounds and hence proves useful sometimes), the ZCR for all frequencies above 2000 Hz (high-channels ZCR) and the ZCR for all frequencies below 2000 Hz (low-channels ZCR). A critical point here is determining the threshold that can be used to separate voiced regions from unvoiced and noisy regions. As has been observed from inspection of various speech files from different databases, such a universally applicable threshold is not valid, because of the variation of the noise (and hence the ZCR) characteristics over different channels, and the susceptibility of speech to such noise (which is a function of the speaker's speaking strategy). Thus, a better idea is to arrive at a proper threshold for each speech signal separately. A useful strategy seems to be to store the three ZCR characteristics for the voiced and unvoiced regions and maintain their running average over frames. For each current frame, the average ZCR characteristic is calculated for the voiced and unvoiced frames that have passed and been identified until then. Then the threshold of ZCR for that specific frame is set to be the mid-point between these representative means. This procedure is followed for each of the suspected frames, using all the voiced and unvoiced regions obtained until then (including the purely periodic and aperiodic frames). If all three ZCR characteristics are classified as belonging to the voiced region, then the frame is classified as being voiced; if not, then unvoiced. It was observed that this apparently conservative decision rule in fact does not miss a significant number of creaky regions, and actually allows a leakage of unvoiced regions into voiced regions. In essence, it could be said that the false alarm rate is high and the false rejection rate is low using this strategy.

3.4 Spectral Slope

While the above steps allow some separation between the voiced and unvoiced frames, they are still not adequate to ensure false alarm rate is low. Indeed, the above characteristics have been conditioned in such a way as to ensure that creaky regions are not rejected in the decision process, even if that allows certain unvoiced regions to the next stages of processing. A further issue is that in spontaneous speech, where co-articulation effects are significantly high on one hand and articulation of phonemes may not be accurate due to speaker tendencies on the other, there can exist other regions of speech where the consonants could take a form that closely matches that of irregular phonation. For example, there exist many instances where stops show a phonation pattern (and therefore, dip profile) very similar to that of a creaky vowel [12], and they would therefore be allowed by the ZCR constraints. Further, in case of some voiced stops, the voicing pattern also seems very similar to that of the creaky voicing, and could easily be mistaken for irregular phonation. (In fact, for certain cases, it is not clear whether to call such stops as having regular or irregular voicing. Such cases are classified as false alarms in this thesis, though there is reason to question if they are regular phonation. Please see chapter 4 for more discussion). There is a need to eliminate the acceptance of such consonants into the following stages of the algorithm.

To achieve this, a useful parameter to use is the Spectral Tilt. Vowels are produced by a source signal whose spectrum has a falling slope, exciting a vocal tract whose spectrum contains several peaks (formants), which causes the speech spectrum to essentially have a falling slope. However, consonants are produced by forming constrictions in the front part of the vocal tract. This either increases the resonant frequencies of the vocal tract and causes the speech spectrum to have a rising slope (like in case of fricatives), or causes the spectrum to be flat like in cases of stops which have an impulse like nature. As an illustration, Figure 3.4 shows the spectral behavior of three different kinds of phonemes – a creaky vowel, a stop and a fricative. It can be seen that in the frequency region of 0 to 2000 Hz, the creaky vowel has a falling slope, while the stop and the fricative show a relatively flatter slope. At frequencies above 2000 Hz, the vowel still shows a slope though not as pronounced, and the stop shows a more or less flat slope. The fricative, however, shows a rising slope in this frequency range. We could thus exploit the spectral slope to separate voiced sounds from unvoiced sounds. Indeed, all voice qualities – breathy, creaky or modal – exhibit a falling slope, albeit at different slopes, thus characterizing voicing.

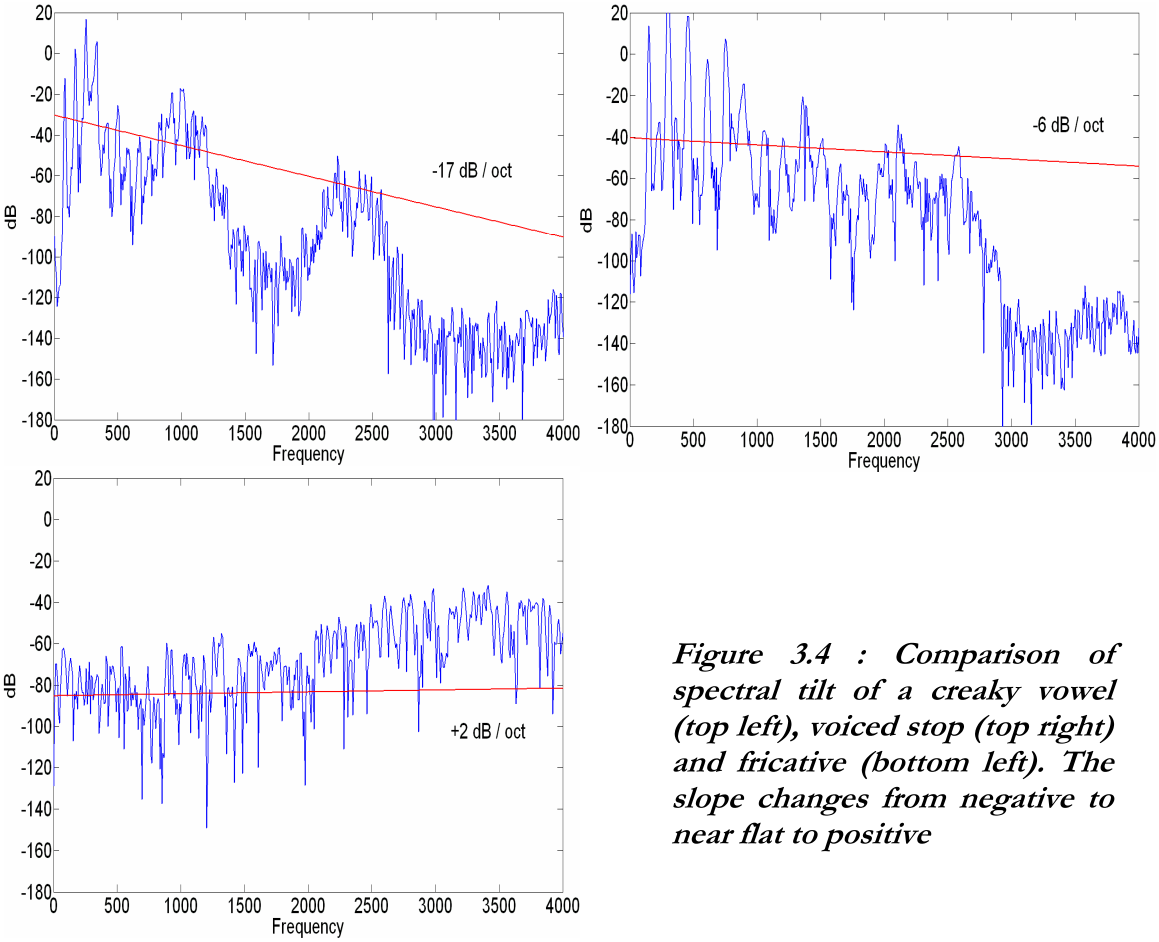

Figure 3.4 : Comparison of spectral tilt of a creaky vowel (top left), voiced stop (top right) and fricative (bottom left). The slope changes from negative to near flat to positive

Important considerations for using this parameter then are – the range of frequencies over which to calculate the spectral slope, and the threshold value that can be used to separate out the voiced and unvoiced regions. One important detail to be considered is that this parameter should be independent of, or at least robust to, the channel being used or the recording device. The spectral characteristics of these devices can significantly influence the spectral tilt and thus render the parameter unusable. In order to make the parameters least susceptible to such effects, the range of frequencies over which the tilt is calculated is restricted to a small range – 100 to 2000 Hz. The motivation is that the range should include voicing information and thus include the lower frequencies, and yet contain a significantly large spectral band to ensure consistency of the parameter across that spectral range. The cutoff of 2000 Hz ensures that the spectral characteristics of various phonemes remain consistent in that range. The operating bandwidth of most devices also includes this range of frequencies, although the lower cutoff frequency may be slightly higher than 100 Hz sometimes. Thus, it can be expected that this parameter will be sufficiently robust to any channel variations. The spectral tilt parameter is extracted by calculating the FFT of the windowed signal, and then calculating the slope of the spectrum between frequencies 100 to 2000 Hz by fitting a line using Minimum Mean-Square Error (MMSE) criterion.

The threshold for using spectral tilt to separate voiced and unvoiced regions was found using a linear support vector machine (LSVM). The spectral tilt parameter was calculated for 1000 frames each of voiced and unvoiced speech, from different databases including both clean and telephone speech, and this parameter was then fed to a LSVM for classification of the unvoiced and voiced. The entire set was used for both training and testing, and the minimum false rejection rate for the voicing class was obtained at an effective threshold of -16 dB/octave for the spectral tilt. Thus, -16 dB/octave is used as the threshold for detection of voiced/unvoiced to eliminate frames that have consonants.

3.5 Dip Profile – Eliminating Voiced Fricatives and Breathy Vowels

Once the suspected creaky frames have been identified and any possible unvoiced frames have been eliminated, the first test for creakiness is performed. In this stage, the dip profile of the frame is analyzed for the pattern that should be characteristic of creaky regions – namely, loosely formed clusters that yield a high pitch period. It was observed that by this stage, the majority of the frames detected had irregular phonation, but some breathy vowels and voiced fricatives were also being detected. This can be explained by the fact that similar to the irregular phonation, the breathy vowels and voiced fricatives have some underlying periodic energy because of voicing, and this causes them to be identified as voiced, which is not an error after all. However, as had been pointed earlier, these frames have a dip profile that looks very similar to that of the creaky voicing, because of the voicing accompanied by the noise. If these frames are directly checked for their dip profile properties, they would also be called creaky. To eliminate the breathy sounds and voiced fricatives, the fact that can be exploited is that the voicing in these two cases will often be in frequencies 0 – 1000 Hz, while irregular phonation is expected to show its characteristic irregularity at all frequencies. This is because as mentioned in Chapter 1, the spectral roll-off of breathy voice is high, and thus, the source signal is not very strong at the higher frequencies. Even for modal voicing, the roll-off is higher than the creaky voicing. For the creaky voicing, which typically shows distinct impulse-like glottal pulses, the glottal spectrum is usually flatter and has strong spectral harmonics even at higher frequencies. This fact is also apparent from the spectrogram of a creaky vowel, where the vertical striations arising from the glottal pulses show a strong presence at the higher frequencies as well. Thus, the dip profiles are obtained by summarizing across only those channels that span frequencies above 1000 Hz.

The next step is to characterize the dip-profile that is unique to irregular phonation. This is done by finding all the local maxima in the dip-profile, and eliminating all those maxima that lie too close to each other – this is done in order to ensure that the corresponding pitch estimate remains below 150 Hz, as is expected for irregular phonation. Local maxima in the dip clusters are then identified and defined as cluster centers, and clusters are formed around these centers that are loose, in the sense that they are wider than expected for periodic clusters. For periodic clusters, a typical width of five lags on either side of the cluster center ensures more than 50% contributing to the cluster. In the case of irregular phonation, the width of the cluster is increased from 11 to 31, by allowing 10 lags on either side. Out of the total number of channels contributing to the dip cluster, some will contribute to the clusters thus formed, while others have dips lying outside the cluster. A score is made identify the channels which have a majority of their dips contributing to the clusters – a score of 1 is given if not more than one dip falls outside the clusters, and 0 otherwise. The cluster center is then recalculated, using only those channels that have been identified above to have confidence 1. The motivation for this is that by eliminating the dips lying outside the clusters and allowing only channels that contribute to clusters, any potential aperiodic (arising from voiced fricatives & breathy vowels) channels are not allowed to contribute to the process.

Redefining the cluster centers using only these channels, the channel scores of all the channels are calculated again. The channel scores are then added across frequency, and if at least 50% of the channels show a score of 1, the frame is declared to have the characteristic dip-profile. It is obvious that this process ensures that aperiodic frames are eliminated – if a frame is aperiodic, it would have a large number of channels with their dips lying outside the clusters, and thus, the required channel threshold of 50% would not be crossed. The possibility of capturing turbulence is further reduced by conditioning the ratio of number of dips present within the cluster, to the number of non-zero dips, to exceed a threshold of 40%. For frames with irregular phonation, the aperiodicity is not random, and therefore the number of non-zero dips lying inside the cluster will be higher compared to the cases of breathiness and voiced fricative, which show random aperiodicity and hence will have a large number of dips outside the cluster.

Thus, at the end of this stage, a frame is classified as being either irregular or aperiodic. If the frame has been classified as having irregular phonation, it is then used for pitch detection in that frame.

3.6 Pitch Estimation

The next step would be to identify regions of voicing and then label these regions as creaky, if certain pitch constraints are satisfied. One of the most obvious solutions for detection of creaky regions in speech is to identify regions with significantly low or irregular pitch. Thus, pitch correction is a very useful step in the detection of creaky regions. The APP detector has a pitch detector algorithm that relies on the AMDF dip clusters to estimate the pitch in the voiced regions. However, since the APP detector does not see creaky regions as being periodic or voiced, it does not give any pitch estimates for such regions. This is not an uncommon problem with pitch detection algorithms – most pitch detection algorithms fail to detect the pitch correctly in regions of irregular phonation, due to various algorithmic constraints, like in-built pitch memory, maximum allowed change in pitch etc. If these constraints would not be placed, then the pitch detection algorithm would not perform a good job in cases of spontaneous speech due to noise and distortion, and would be susceptible to pitch halving and doubling errors. However, enforcing such constraints does not allow detection of creaky regions (since absence of pitch implies absence of voicing, thus detection of creakiness and estimation of pitch must go hand in hand).

If it were possible to detect voicing by some other means, and then declare that frame as being voiced, then it would be possible to either estimate the pitch first and then detect creakiness in that frame, or call that frame creaky because of presence of voicing and absence of pitch, and then proceed to find the pitch estimate. However, such a voicing-detection scheme must be robust to channel distortions, noise due to various sources and other phenomena encountered in spontaneous speech, like coughing, laughter etc., in order to be put to practical use. Typical voicing detection techniques, like pitch estimation, zero crossing rate, zero crossing rate in different channels (frequency bands), energy thresholding, rate of energy change in different channels etc. do not perform up to expectations in case of spontaneous speech. Figure 1.7 illustrated this difficulty in the case of a speech file from the NIST 98 corpus, which consists of telephone speech. The speech signal in question contains instances of modal voicing, laughter, breathy vowels and creaky vowels. It may be seen that there is no possible combination of all of the above mentioned features that could lead to a separation of the voiced region from the unvoiced region. Furthermore, a comparison of the same features for the laughter, and the creaky region, prove the above mentioned parameters to be totally inadequate for separation of voiced and unvoiced regions. In fact, in spite of an extensive literature survey by the author, there have been no methods that proved efficient to separate voiced regions from unvoiced regions in such cases.

It becomes clear thus that pitch estimation is indeed a very difficult problem and can not be solved easily. Indeed, the nature of the problem is in itself rather tricky – what is needed is a pitch detection algorithm for creakiness detection, but such pitch detection can be done only when voicing (i.e., creakiness) is detected. In order to solve this problem, the approach taken in this work is to use the channels of the auditory filter-bank to make a vote about whether the particular frame is voiced or not. This is done by first considering the dip profile of the frame and forming loose clusters of dips and identifying how many channels contribute to the cluster. If the number of channels exceeds a certain threshold (40% of total channels contributing to the dip cluster), an initial guess is made about the frame being creaky. Then, conditioned on the satisfaction of some more acoustic properties detailed in the next chapter, the maximum of the cluster with most dip contribution is estimated to be the dip corresponding to the pitch period. The pitch frequency is calculated accordingly. Thus, an estimate of the pitch is made in suspected creaky frames. Once the pitch estimate is made, the confidence of creakiness is recalculated using this estimate. In this second pass, slightly narrower clusters are formed near this pitch period estimate, and the required threshold for the contributing channels is increased to 60%. If the frames exhibit a dip cluster confidence that exceeds this threshold, then the frame is judged to be creaky and the pitch estimate is retained. If the dip cluster fails to satisfy the required threshold of confidence, the pitch estimate is disposed and the frame is adjudged to be unvoiced.

Thus, creakiness is used to first estimate pitch, and this pitch is then used to recalculate the confidence of creakiness of the frame. While calculation of creakiness happens to be a deciding factor in pitch estimation, it may be noted that this is not the only criterion, and that there are a significant number of other properly motivated acoustic characteristics that are also used to confirm creakiness, before pitch is estimated. Indeed, pitch estimation is a consequence of those steps and is merely used to reconfirm the confidence of calling the current frame creaky. The following figure shows the pitch estimates obtained in the case of the same speech file as shown in Figure 3.5.

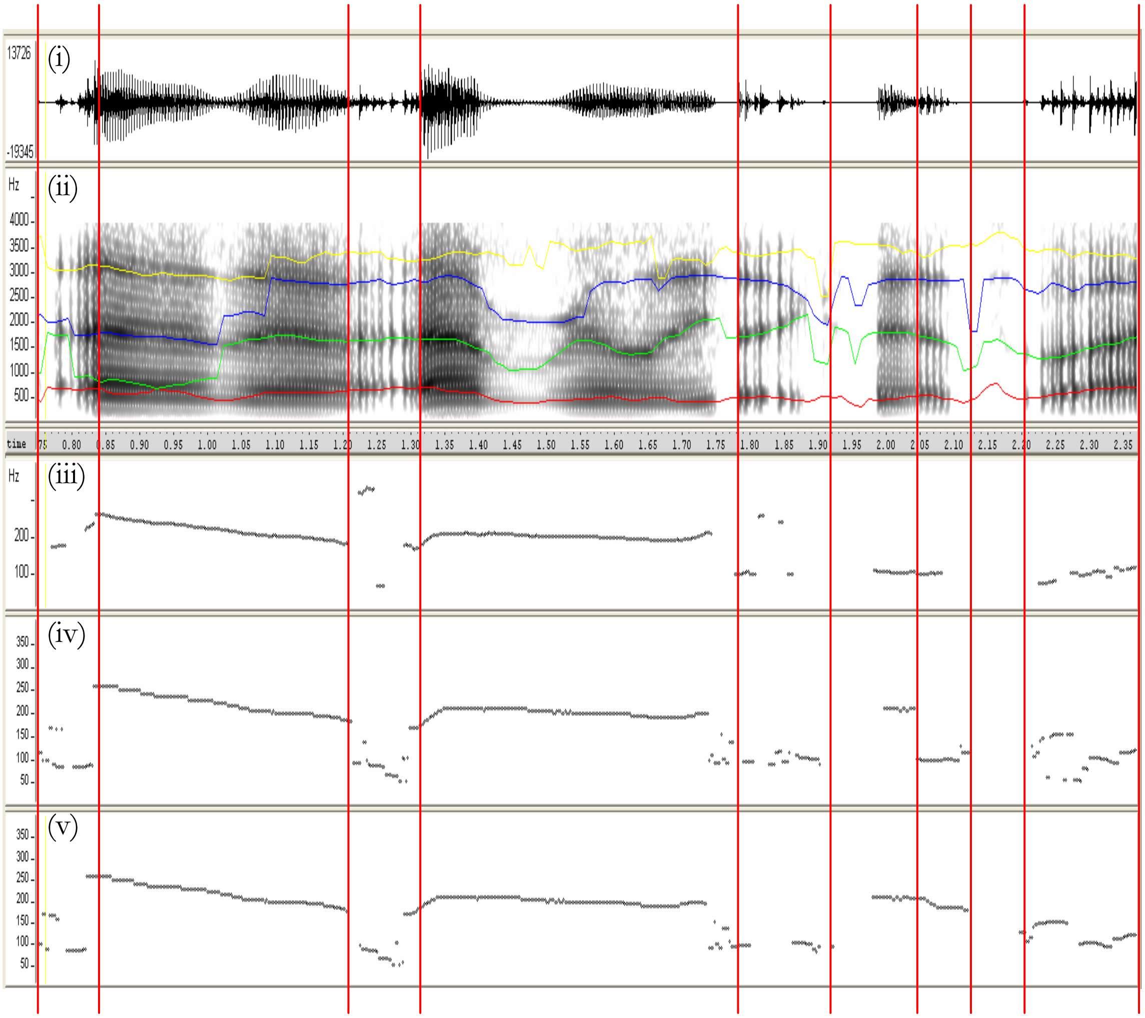

Figure 3.5 : Pitch tracking accuracy of three algorithms in creaky regions (boundaries marked by vertical red lines). (i) waveform, (ii) spectrogram, (iii) pitch tracking using the ESPS software, (i)pitch track from the modified APP detector, (v)pitch track from the original version of the APP detector

It may be seen that the ESPS pitch tracker, as well as the original version of the APP detector, fail to get the correct estimate of the pitch in creaky regions, while the modified APP detector gives the correct estimate. The ESPS pitch tracker fails in most creaky frames, and in one instance confuses the modal voice for creaky voice, yielding a low pitch. A comparison of the number of creaky frames whose pitch has been correctly identified by the original APP detector with that by the modified APP detector, shows that the pitch detection accuracy has increased. Especially in the last creaky region, the original APP detector almost completely loses the pitch track, but the modified APP detector is able to find a pitch track in some of the frames. In addition, the number of frames where pitch

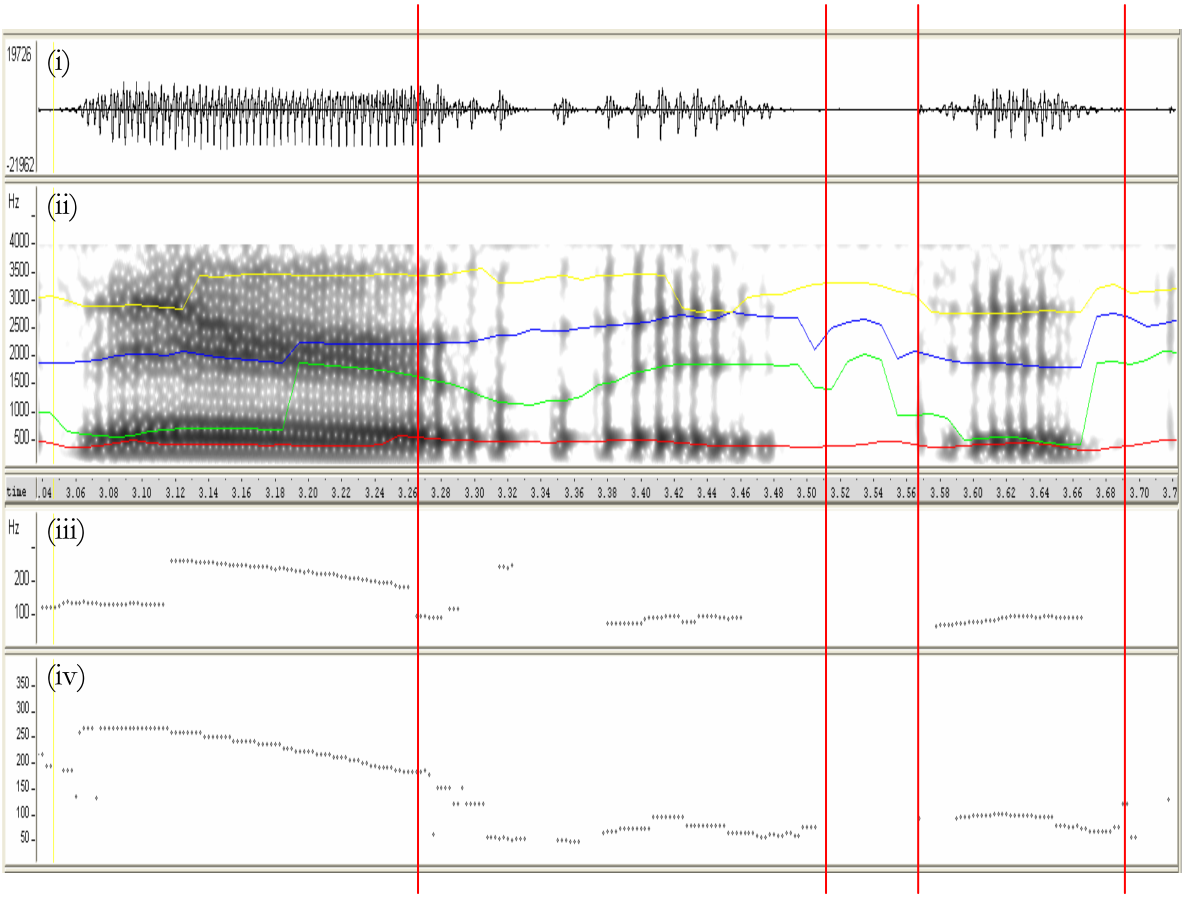

Figure 3.6 : Pitch tracking of the new APP detector algorithm in creaky regions (identified by vertical red lines). (i) waveform, (ii) spectrogram, (iii) pitch tracking using ESPS software, (iv) pitch track from the modified APP detector. The instantaneous pitch tracking accuracy of the algorithm may be noted.

values are given when there actually is no voicing is also very small. Further, it may be seen in the second creaky region of Figure 3.5 (region between 1.2 and 1.3 sec) that the instantaneous changes in the pitch are being tracked very accurately. Figure 3.6 also shows that the pitch tracker is able to track instantaneously the irregular changes in pitch, while at the same time also maintaining a consistent pitch track with no halving or doubling errors in other regions. The algorithm does not necessitate the use of a memory for pitch estimation, as do most pitch detection algorithms – this is what allows pitch tracking even in regions of irregular pitch.

It is also noteworthy that pitch estimates in the voiced regions, and unvoiced regions where there is no reason to suspect creakiness, are left unchanged. Thus, by modifying the pitch estimation algorithm, the pitch estimates can only be expected to get better and not be harmed in any way (in the voiced regions). This improvement is demonstrated in the figures above. However, a consequent problem is that in regions where the APP detector has also made pitch detection errors like other pitch trackers, no corrective measures could be taken since the pitch confidence was sufficiently high not to cause the pitch correction algorithm to be applicable there. An illustrative example is seen in Figure 3.7. Between the region of 0.94 sec to 1.0 sec, it may be seen that the pitch period is 0.004 sec, and therefore the pitch estimate should be around 250 Hz. However, both the ESPS pitch tracker and the APP detector (modified) show a pitch track around 110 Hz. This kind of pitch halving error is probably due to the modulation of the envelope of the speech signal in that region. However, because the pitch confidence is high in that region (the pitch estimate is at twice the actual pitch period, and since there are dips there as well, its strength might override that of the first cluster of dips), the modified version does not try to correct the pitch there.

The next chapter describes the actual creakiness detection algorithm that precedes pitch estimation.

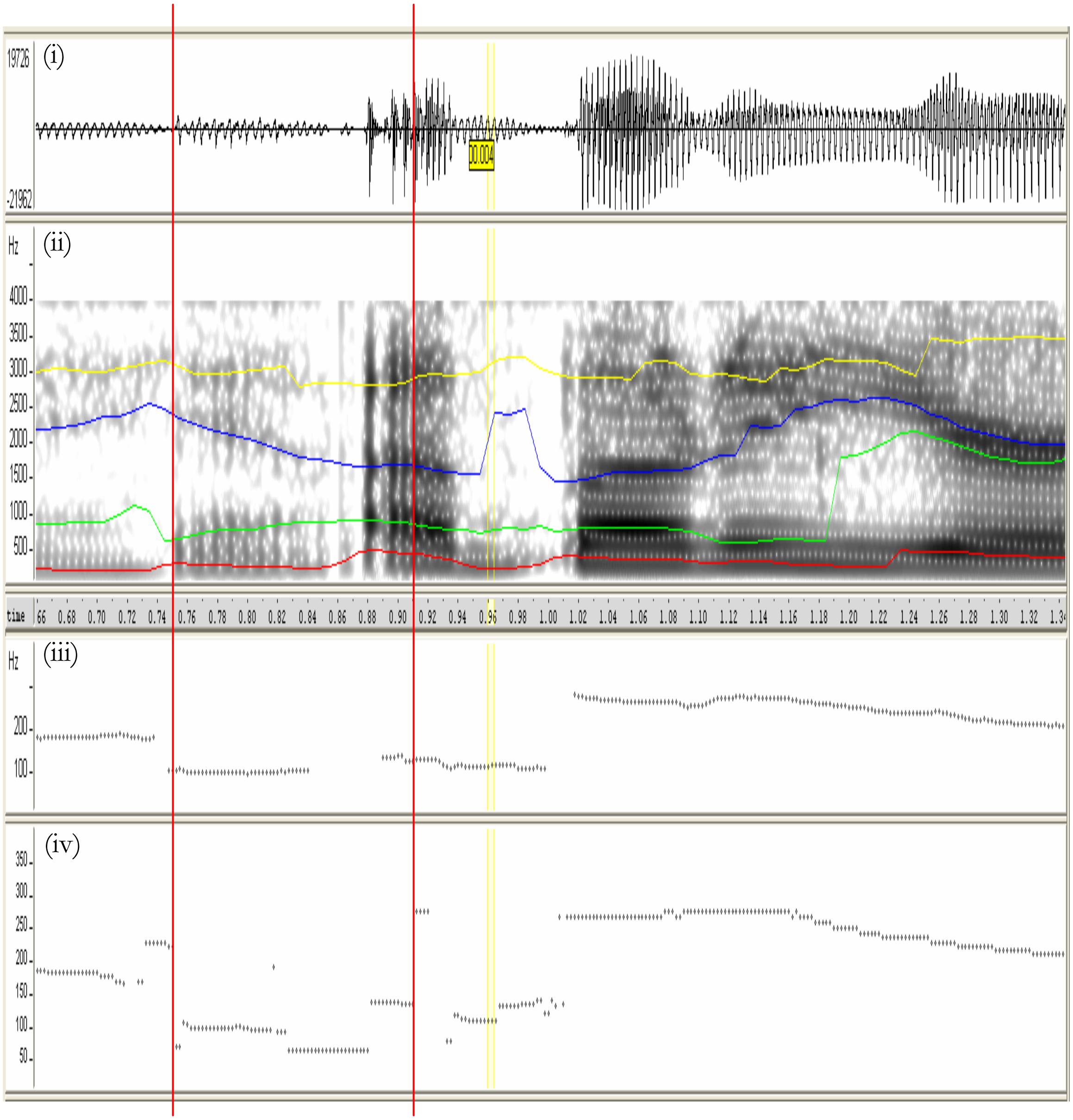

Figure 3.7 : Failure to correct pitch in certain regions, due to high pitch confidence. (i) waveform, (ii) spectrogram, (iii) pitch tracking using ESPS software, (iv) pitch track from the modified APP detector. The pitch period of 0.004 sec (highlighted region in yellow) may be compared to the estimated pitch frequency of 109.59 Hz, which is a pitch halving error. This is because of high pitch confidence at twice the pitch period.

3.7 Final Decision Process : Combining Current Pitch Information and Dip Profile Decisions

When a frame has been determined to be creaky, and the pitch has been estimated, it is rechecked for creakiness. This is done by first checking if the pitch estimate obtained is below the "normal" or modal pitch estimate for that speaker. For this, a running average of the pitch values in modal voicing is made for all the frames that have been processed so far. For each frame, all the pitch values for all the regions that have been identified as being voiced but not creaky are stored in memory, and their mean is taken as a threshold. If the current pitch estimate is less than some percentage (empirically set to be 75%) of this "normal pitch", then the current frame could possibly be a creaky frame, as are characteristic of creakiness in particular and irregular phonation in general. Further, in some cases of irregular phonation, the pitch may not fall low, but the pitch track may be irregular, and show a big difference (as high as 30Hz) from its immediate previous value. Such changes are also tracked, by maintaining a running memory that stores the past 10 pitch values, and detects regions where the pitch has deviated significantly from the stored values. If the current pitch estimate exhibits such a behavior, then too the frame is classified as being possibly creaky.

Only if either one of the two above conditions related to pitch is satisfied, the next creakiness check is made. The dip profile is again checked to verify if it shows the dip-profile behavior with the current pitch estimate. In this second pass, tighter clusters are formed (8 lags on either side of the pitch period estimate) and the channel scores for this cluster are calculated. If the total number of channels contributing to the cluster now exceeds 60% of the total number of channels, then the frame is finally adjudged to be creaky, and the pitch estimate obtained above is declared to be the pitch for that frame. Once the frame is called creaky, the energy of all the channels that have shown a score of 1, i.e., which contribute to the creakiness profile, are added together to give what is called the creakiness energy in the frame. This is the measure used to numerically characterize creakiness in that frame.

Thus, the creakiness detection algorithm is a combination of various knowledge-based acoustic features built upon an Aperiodicity, Periodicity and Pitch detector, rendering the APP system now capable of distinguishing aperiodicity due to turbulence from that due to irregular phonation.

The next chapter discusses the performance of the Irregular Phonation detector, some failure modes, and its application to speaker ID experiments.

Chapter 4

EXPERIMENTS & RESULTS

The algorithm for detection of irregular phonation has been described in the preceding chapters. This chapter details the experiments that were performed to validate the performance of the algorithm, and an application of the creakiness parameter in speaker ID experiments. The first set of experiments is to verify the pitch detection accuracy of the modified APP detector. This is a first step validation of creakiness detection. The next set of experiments deals with finding the detection rate and false alarm rate of the creakiness detection algorithm. The final set of experiments demonstrate the improvement in the recognition rates in speaker ID experiments upon inclusion of the creakiness parameter

4.1 Database Used For The Experiments

In order to find the pitch estimation accuracy, a reference can be obtained from the standard pitch detection software available. In this thesis, the reference pitch has been obtained using the ESPS Wavesurfer software using the autocorrelation method. To find the creakiness detection accuracy, however, there is a need for a reference or ground truth which can be used to evaluate the performance. However, there is no available standard database that can be used for describing voice quality and testing the performance of the algorithm. Thus, ground truth needs to be marked by hand in one of the standard speech databases. In this thesis, the TIMIT database has been used for the first two experiments, as hand-marked transcription for irregular phonation in those speech files was available from Dr. Slifka at MIT. Since there is no clearly defined reference that marks irregular phonation, and since the boundaries and occurrences of such phonation are difficult to define, we have called all those locations marked in the reference, as well as those identified by our algorithm but missed in the reference (confirmed by visual inspection by the author), as the total number of instances of irregular phonation. Specifically, the "test" subset of the TIMIT database was used. The number of files processed was 895, and it included 65 male and 45 female speakers from eight dialect regions (dr1 through dr8). The total actual number of instances of irregular phonation was 1,400.

Since the long term goal of this work is to use the algorithm in real world speech applications, the algorithm has also been tested on spontaneous speech from telephone data. The NIST 98 database was used for performing the speaker ID experiments. The first two experiments were also performed on the NIST 98 database, to ensure that the algorithm works well in spontaneous speech. The "test" subset was used for the first two experiments, and 100 files of the NIST 98 database were hand marked for irregular phonation by the author, yielding 200 instances of irregular phonation, 100 for each gender. For both these tasks, the test session of the NIST 98 data was used, and the first 5 sec were extracted for each file and treated as one speech file. For the task of speaker identification, a subset of the database containing 250 male speakers and 250 female speakers was chosen. The training utterances are taken from the train/s1a/ directory and the testing utterances are taken from the test/30/ directory of the database. There is no handset variation between the training and the test utterances. The length of each training utterance is approximately 1 minute and the testing utterances are about 30 seconds in duration. An energy threshold was used to remove the silence portion (which sometimes has low amplitude background noise) from the speech. This resulted in training utterances of about 30 to 40 seconds and testing utterances of about 10 to 20 seconds.

A short comparison of the two databases would be in place here. The TIMIT database is clean speech sampled at 16 kHz, and contains speech from the same speaker. The speech is controlled, in that the phonemes are uttered clearly and except silence, there is little other non-speech in this database. On the other hand, the NIST 98 database is spontaneous speech sampled at 8 kHz. The speech is subject to telephone channel distortion that corrupts the amplitudes of the lower harmonics (up to around 300 Hz) and contains background noise. Further, effects like laughter, coughing etc are also seen in some files. Thus, this latter case presents a tougher challenge than the former.

4.2 Experiment 1 – Accuracy of Pitch Estimation

This experiment was performed on both the databases. As mentioned, the reference pitch is obtained from the ESPS Wavesurfer software using the autocorrelation method. The frame rate of both pitch tracks is the same, namely 2.5 msec. The window size for Wavesurfer was set to be the default (optimal). Pitch detection and all experiments with the APP detector are performed with a window size of 20 msec.

As mentioned in section 2.3, the APP detector makes some errors and detects pitch in the unvoiced regions (even before the incorporation of the creakiness detection algorithm). Thus, the voiced regions and unvoiced regions have been compared separately.

Comparison in the Voiced Regions

The mode of comparison is as follows: in case of voiced regions, only those regions where the reference pitch detector has non-zero pitch estimates have been considered. The voiced regions include regions of both regular and irregular phonation. If, for a frame, the pitch values of the APP detector and the reference vary by more than 15 Hz or 10% of the current pitch, whichever is minimum, then the pitch track of the APP detector for that frame is declared to be in error. It may be noted that the reference pitch detection is not always very accurate. It has been shown in Figure 2.5 that the pitch tracker fails to track the correct pitch in regions of creakiness. However, such regions have not been separated out in this task, since there can be no automatic way of determining if the reference pitch track is indeed correct, and it is manually very difficult to visually verify this for each of the errors. Thus, the numbers given below give a representative idea of the general pitch tracking performance. The following table shows the results for this test:

# of frames	# of frames where pitch was detected correctly	
	Before modification	After modification
20,000 (NIST)	19,080 (95.4%)	19,660 (98.3%)
8,000 (TIMIT)	7,582 (94.8%)	7,795 (97.4%)

Table 4.1 : Pitch Estimation Accuracy of the modified APP detector, as compared to the earlier version

Approximately 50 sec of voiced speech, from both male and female speakers, was taken from the NIST database. In the case of the TIMIT database, about 20 sec of voiced speech was taken, from speakers of both genders. It may be seen that the pitch detection accuracy has increased by about 3% in case of both databases. Recall that the reference tracker sometimes fails in tracking the correct pitch in creaky regions. This might mean that some mismatch can happen when the APP detector tracks the pitch correctly, but the reference is wrong. It could also mean that the APP detector pitch track is not accurate.

Assuming that the reference pitch track is always accurate and correct, the performance of the modified APP detector is then increased by about 3% for both databases. This implies in turn that there is at least a 3% improvement (and possibly more) in the actual pitch tracking accuracy.

Comparison in the Unvoiced Regions

The mode of comparison is as follows: in case of unvoiced regions, it is required to find out how many times the modified APP detector has detected a region as voiced and given a pitch track, in addition to what the earlier version of the APP detector used to do.

Using the reference pitch tracker to identify regions of unvoiced speech, about 50 sec of unvoiced speech in NIST and 20 sec in TIMIT was used for testing. The following table shows the results:

# of frames	# of unvoiced frames where pitch value was given	
	Before modification	After modification
20,000 (NIST)	1,360 (6.8%)	1,440 (7.2%)
8,000 (TIMIT)	256 (3.2%)	272 (3.4%)

Table 4.2 : Insertion of voiced decisions for unvoiced speech

It is seen that the number of insertion errors has increased by a very small amount (<1%) for a large number of frames. This proves that the voiced / unvoiced decision made by the modified APP detector is almost as efficient as the original, and the number of frames which are judged as being voiced and for which pitch estimates are given, does not increase significantly. Once again, it is possible that the region marked as unvoiced by the reference pitch tracker is in fact creaky, but has been marked unvoiced due to the tracker's failure to give a pitch reading in that case. If that were the case for any frame, then the region is actually voiced and thus, the insertion error is less than what is reported above. Thus, it could be said that the maximum possible increase in the insertion error rate is about 0.4%.

These experiments indicate that the pitch tracking algorithm of the APP detector has been improved by the modifications described in chapters 2 and 3, at the expense of little debilitating effects. The next set of experiments details the creakiness detection accuracy.

4.3 Experiment 2 – Detection of Irregular Phonation

The following Figure 4.1 shows a sample run of the creakiness detector on a speech file. The areas identified as being creaky are marked in black on the spectrogram. It may be seen the creaky areas have all been successfully detected.

The creakiness detection performance is evaluated using two metrics: the detection accuracy and the false alarm rate. Each of these will be presented below.

Detection Accuracy of Irregular Phonation

Two kinds of measures are used to measure the performance of the creakiness detector: the percentage of instances of creakiness detected, and the relative number of frames that have been correctly identified. These are defined below:

$$\text{Instance detection accuracy} = \frac{\text{Total number of creakiness instances detected}}{\text{Total number of hand-marked instances}} \times 100\%$$

$$\text{Frame detection accuracy} = \frac{\text{Total number of frames of creakiness detected}}{\text{Number of frames in the hand-marked instances}} \times 100\%$$

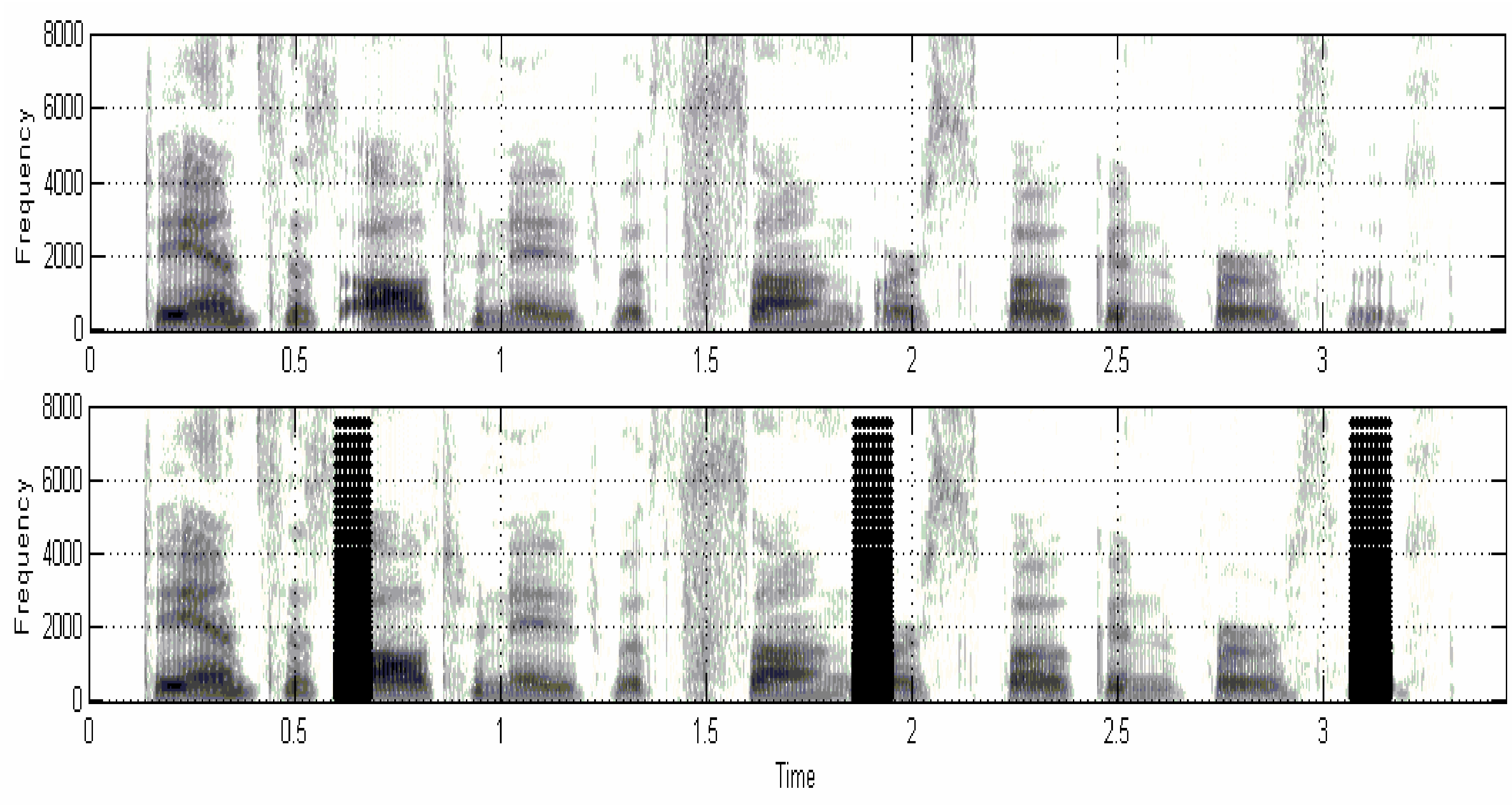

Figure 4.1 : Creakiness detection results. (top pane) Spectrogram showing the regions of creakiness in the original signal; (bottom pane) Same spectrogram as above, with the creaky regions identified by the APP detector overlaid in black

The following tables give a statistical summary of the performance of the creakiness detection algorithm, for the TIMIT database:

	Total # of instances	# Identified	Percentage Identified
Male + Female	1,400	1285	91.8%
Female	584	543	93.0%
Male	816	742	90.9%

Table 4.3a : Accuracy of detection of creakiness instances in the TIMIT database

	Total # of frames	# Identified	Percentage Identified
Male + Female	26,408	23,564	89.2%
Female	13,454	12,243	91.0%
Male	12,954	11,321	87.4%

Table 4.3b : Accuracy of detection of creakiness frames in the TIMIT database

The following tables give a statistical summary of the performance of the creakiness detection algorithm, for the NIST 98 database:

	Total # of instances	# Identified	Percentage Identified
Male + Female	200	183	91.5%
Female	100	94	94.0%
Male	100	89	89.0%

Table 4.4a : Accuracy of detection of creakiness instances in the NIST 98 database

	Total # of frames	# Identified	Percentage Identified
Male + Female	4,291	3,746	87.3%
Female	2,337	2106	90.1%
Male	1,954	1,649	84.4%

Table 4.4b : Accuracy of detection of creakiness frames in the NIST 98 database

The creakiness instances detection accuracy is seen to be slightly above 90% for each of the two datasets, and is about the same for both genders. Both female and male samples are handled equally well by the algorithm, which confirms that irregular phonation possesses acoustic features that do not depend on gender. Further, the algorithm performs well even with the NIST 98 database, in spite of debilitating conditions. Errors were more for male speakers than female speakers, and this is attributed to the low pitch of male speakers, which may often cause the dips to scatter from clusters and thus manifest the sound as being similar to irregular phonation. Also, the inherently low pitch in males causes some of the creaky regions to be missed because creakiness detection is conditioned on the pitch falling below a certain threshold. In terms of identification of number of frames, the performance goes down slightly owing to the possibility that not all frames may be detected by the algorithm, as marked in the reference. Considering the possibility that some of the instances identified as creaky may be marked so for a larger number of frames than in the reference, this implies that the actual frame detection accuracy may be even lesser than the above reported numbers.

False Alarm Rate

The false alarm rate can also measured in terms of two measures: percentage of instances inserted, and percentage of frames inserted. Each of these terms is defined below:

$$\text{False Instances Rate} = \frac{\text{Total number of instances of false alarms caused}}{\text{Total number of actual creaky instances}} \times 100\%$$

$$\text{False Frames Rate} = \frac{\text{Total number of frames of false alarms}}{\text{Total number of frames - Total number of creaky frames}} \times 100\%$$

In reality, it is the second term that actually defines the false alarm rate correctly. However, for this set of experiments, because the total number of frames is too large a number compared to the number of false alarm frames, this second measure becomes a number too small to be of significance. Thus, the second measure is redefined in the following way:

$$\text{False Frames Rate} = \frac{\text{Total number of frames of false alarms}}{\text{Total number of creaky frames}} \times 100\%$$

The first and redefined second measures have been used to describe the false alarms. The false instances rate has been found to be 12.8% for the NIST 98 database, and 17.4% for the TIMIT database. Surprisingly, the false alarms are more for the cleaner database (TIMIT) than the one which has noise and distortions. Of all these false triggers, 35% were due to voiced fricative /sh/. About 40% of the false detections were due to stops being called irregular, while the remaining 25% of the false triggers were due to stops wherein the vowel preceding the stop was identified as creaky. Though we have currently included the detections in the latter category as false detections, studies have shown [11, 12] that there do exist cases of stops in both American English and other languages, where both voiced and unvoiced stops may be accompanied by irregular phonation. Further, it is possible that the end of vowels preceding such stops may also exhibit irregular phonation. Thus, it may actually be the fact that the algorithm is capturing such instances of irregular phonation too, in which case the false triggering rate will actually be considerably low. In addition, it has also observed that we are identifying what we suspect are some glottal squeaks (4 instances). However, due to the lack of a standard reference, it is not possible to confirm these speculations at this juncture, and this count has been added to false triggers.

Typically, each false alarm instance is about 3 to 4 frames long, as compared to actual creaky regions which are typically more than 10 frames long. However, a significant percentage (about 10%) of creaky regions are also short (of the same duration as the above false alarm instances) and thus, it is not possible to set a duration threshold for elimination of false alarms. The false instances rate of 12.8% translates to a false frame rate of 3.4% for the NIST 98 database, and the TIMIT false instances rate of 17.4% translates to a frame rate of 5.6% (which confirms that the false alarm instances are about one fourth of the duration of a creaky instance).

4.4 Speaker Identification using Acoustic Parameters

One of the potential applications of this parameter is to use it in speaker identification tasks. The creakiness parameter is a parameter that characterizes the source information of the speaker, and should hence be able to help in speaker identification. To verify if this parameter can help in speaker identification and if so, how it affects performance, this parameter has been used in combination with seven other acoustic parameters described in section 1.6. These parameters are the four formants, the amount periodic and aperiodic energies, and the spectral tilt. The speaker ID performance for the seven acoustic parameters was seen to be comparable with that of the standard MFCCs in

[6]. The creakiness parameter is added as an eighth parameter to this set, and the speaker ID performance of the seven APs is compared with that of the eight APs.

The speaker models were constructed using the Gaussian Mixture Models (GMM) and were trained using maximum-likelihood parameter estimation [5]. The MIT-LL GMM system was used for constructing the speaker models. Various model orders were tested and it was empirically determined that the 32-mixture GMM gave the best performance. The test utterance was identified with the speaker whose model yields the highest likelihood for the test utterance. The accuracy of the system was computed using the identification errors made by the system. To obtain the accuracies for different population sizes, the 250 speakers of each gender were divided into groups where the number of speakers in each group is the population size. The accuracy for the particular population size is the average of the accuracies over all the groups. The following table gives the error rates of the seven APs versus the eight APs when used for speaker identification:

Pop. Size	**Gender (# of test utt)**	**Seven APs**	**Eight APs**
50	Female (1379)	29.0	28.0
	Male (1308)	27.8	27.6
	Average	28.4	27.8
100	Female (1104)	33.1	31.7
	Male (1093)	29.1	29.1
	Average	31.1	30.4
125	Female (1379)	34.6	32.6
	Male (1308)	30.0	30.0

	Average	32.3	31.3
250	Female (1379)	38.2	36.7
	Male (1308)	34.3	33.9
	Average	36.3	35.3

Table 4.5 : Error Rates for Speaker ID without and with the creakiness parameter

The results in this table are shown in the following plot in Figure 4.2:

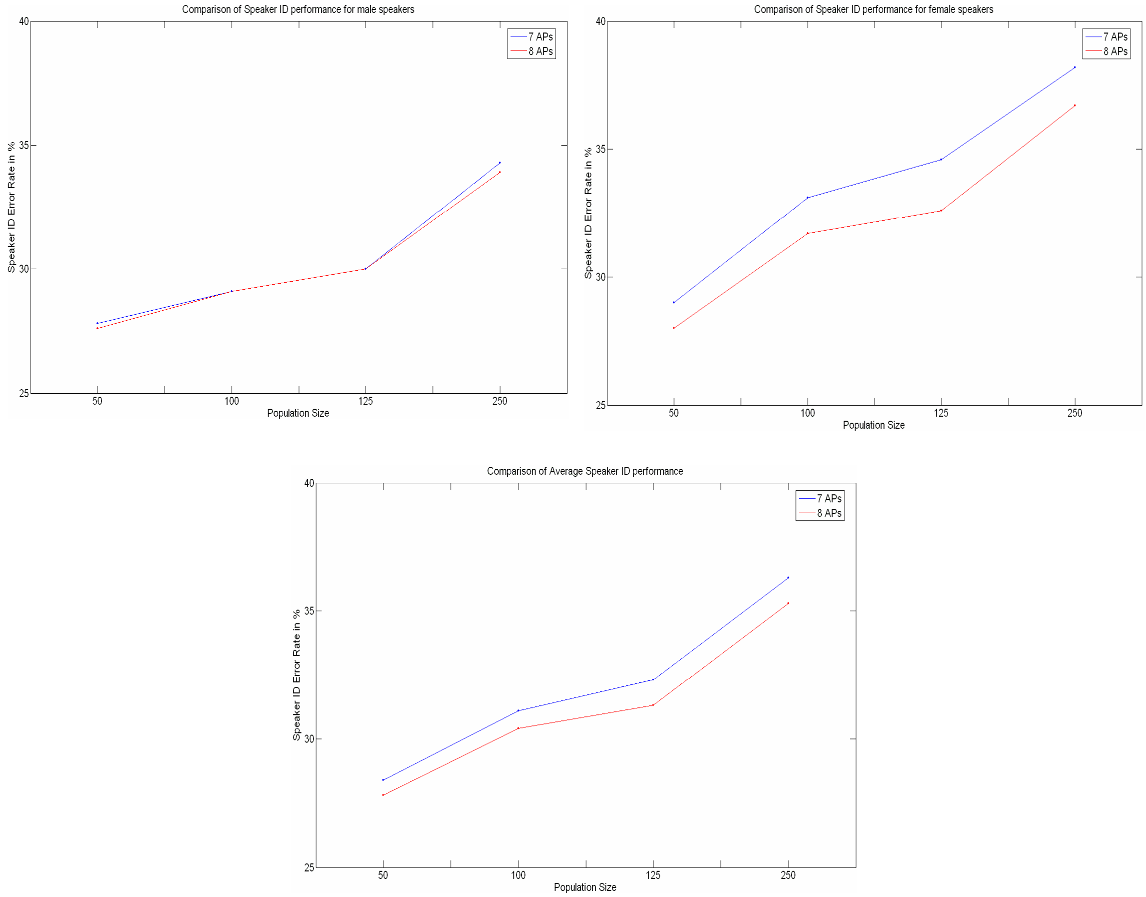

Figure 4.2 : Improvement in speaker ID performance upon introduction of creakiness as the eighth parameter – the error rate has gone down or remained the same in all cases, for all population sizes and both genders

It is seen that for various populations, the speaker identification error rate goes down slightly, upon introduction of the creakiness parameter. Though the improvement is typically about 1% for female speakers and less than 1% for male speakers, this is because of the fact that creakiness occurs very sparsely and hence offers very less training and testing data. Indeed, the creakiness parameter consists of zeros in most regions and hence, when modeled using the GMM, should show a prominent Gaussian at zero for models of all speakers, thus rendering the parameter not significantly useful for speaker ID. However, if there existed a supervised approach that could exploit the creakiness parameter more appropriately for speaker ID, the creakiness parameter would hold some promise as a good acoustic parameter. This is demonstrated by the fact that speaker ID performance improves more for female speakers than for male speakers, though the creakiness detection accuracy does not differ as significantly. This might be explained by the fact that females, having a higher pitch, are thus being more distinctive of their identity whenever they exhibit creakiness. This information is probably captured by the creakiness parameter, which thus helps to characterize effectively some of the female speakers. Therefore, using a supervised approach to exploit the creakiness information could be one possible direction of future research in applications of this parameter.

4.5 Error Analysis and Discussion

Missed Detections

The creakiness detection algorithm does not give perfect detection accuracy. Almost 10% of creaky instances are not identified, and therefore there is still scope for improvement of the algorithm. The failure of the irregular phonation detector in such cases could be explained by several possible reasons. The first is when the location of irregular phonation is at the very beginning of the file – in this case, the algorithm fails to detect creakiness because the AMDF can be computed only after a few initial frames. Also, even when creakiness is not at the very beginning but is in the first few seconds, it might still be missed on some occasions. This is because the voiced/unvoiced thresholds, which are determined adaptively using the values stored in memory are not yet well adjusted to the proper values. If creakiness occurs before adequate voiced and unvoiced regions are seen, the representative values for both these regions are not yet obtained and thus, the threshold calculated is not accurate to make a correct decision. Thus, some false insertions or false rejections could occur. In one case, if creakiness occurs after voiced regions, with no unvoiced regions until that point, then the unvoiced region threshold is biased towards the voiced threshold, and thus, the creaky region is called voiced. This might also sometimes result in some weak fricatives or voiced fricatives being called voiced – they need not be called creaky at the subsequent stages but they might be called voiced at this stage. In the contrast case, if creakiness follows an unvoiced region, the voiced threshold is not set appropriately. This causes the creaky region to be called unvoiced and causes a missed detection error. The next possible reason of missed detection is when the pitch falls so low that the analysis window used cannot capture even one full cycle – when that happens, the AMDF structure cannot capture the characteristic dip profile. A solution for this is to adaptively change the analysis window size. A fourth reason for false rejection could be due to some instances of creakiness which are not significantly different from modal voicing, or which have most of their voicing energy in the lower frequencies. Because only the higher frequency channels are used to cluster the dips, such creaky regions may be rejected since they do not have much voicing energy in the higher regions.

False Alarms

False alarms are mainly due to stops, and about one third of them are due to fricatives. Typically, when these kinds of consonants are adjacent to vowel regions, their characteristics get influenced by their neighboring phonemes due to co-articulation. Thus, a possible reason for these false alarms could be that the values of spectral tilt, ZCR etc. of these regions are near what would be expected for periodic regions, and this is what allows such regions past the voiced/unvoiced decision stage. The effect of voicing in the co-articulated phonemes could exist in such frames, which possibly gives them a dip profile looking similar to that of the creaky frame. Another possible reason for false alarms could be as mentioned above – the voiced/unvoiced detector has not had enough number of samples to make representative thresholds for the voiced and unvoiced regions and thus, some unvoiced sounds are called voiced.

While reasons for missed detections have been elaborated upon, it is hard to find appropriate solutions for some of these causes. For example, the voiced/unvoiced decisions are always hard in case of spontaneous speech, much more so in channels causing distortions. There is always an error in even the current state-of-the-art systems, and voicing detection accuracy is never 100% accurate. In such cases, the best solution is to rely on the optimal trade-off between false alarms and missed detections. In this work, though a ROC curve has not been used explicitly to determine the most optimal operating point, the parameters of the system were varied around the currently existing parameters, and the performance did not vary significantly in terms of missed detection and false alarm performances.

A matter of consideration is also the performance of the algorithm when multiple speakers are present. It may be recalled that the creakiness detection algorithm relies on a memory track of the pitch in voiced regions, in order to decide if a given region is creaky or not. When more than one speaker is present in a speech file, and their pitches differ significantly, then the average pitch threshold that is used for creakiness detection may become unusable to detect creakiness. For example, if a conversation has a male speaker followed by a female speaker, then the pitch memory would have significant number of low (around 100 Hz) and high (around 200 Hz) pitch values. In such a case, the pitch detection threshold would be lower than what it should be if the speech file had a single female speaker. In such cases, when the female speaker goes creaky, it might not be identified by the algorithm, because the threshold has been reduced by the presence of low pitch values due to the male voice. Similarly, a female speaker preceding a male speaker could cause false alarms during the male speaker's turn for the same reason. Although this problem is not handled in this work, it is of noteworthy importance and will be an improvement in a future version of the system.

Chapter 5

SUMMARY AND FUTURE DIRECTIONS

This thesis has focused on the development of an algorithm that will automatically detect and numerically parameterize regions of irregular phonation (creakiness) in speech. This work is significantly different from other research efforts in this direction, in that it processes all regions of speech and not just the voiced regions. Further, it is meant to handle spontaneous speech with effects like channel distortions, non-speech like laughter etc. Thus, the scope and applications of this work are more far-reaching than other parallel efforts.

The algorithm is an extension of the Aperiodicity, Periodicity and Pitch (APP) detector. The APP detector is now incorporated with the irregular phonation detector, and gives a periodicity, aperiodicity, creakiness profile that identifies the three regions respectively in speech. The pitch estimation algorithm in the APP detector has also been improved to correctly identify the pitch frequency in regions of creakiness. The creakiness detection performance is high and matches that of the best performance of other algorithms, and the pitch detection accuracy has also shown improvement. The number of false insertions, both in creakiness detection and in pitch estimation, is low, and counted frame-wise, is less than 5%. This is a set of encouraging results and proves that the creakiness detection algorithm can be used for speech technology on real speech data.

5.1 Future Research Directions

Improving performance of the algorithm

The creakiness detection algorithm shows about 10% missed detection rate, and this implies that there is still some work to be done to improve the detection rate. However, the problem happens to be very difficult, as creakiness and irregular phonation are not discrete steps over modal phonation, but are actually a part of a continuous spectrum with breathy phonation on the other end. As such, it is difficult to define the occurrences of irregular phonation and associated acoustic properties in various cases. Further, because of various effects like co-articulation and channel distortions, some of the acoustic features that characterize creakiness do not manifest themselves in some instances. In order to improve the detection accuracy, it is thus necessary to investigate for further acoustic cues that would help identify creakiness, and incorporate them into the APP detector. Further, some of the signal processing strategies of the APP detector should also be modified in order to catch instances that have been missed. This should, however, be done in a way that the current performance of the APP detector is not deteriorated, which is not an easy task considering the fact that pitch values from 60Hz to 350 Hz should all be detected and recognized. Thus, an adaptive signal processing scheme should be incorporated into the APP detector, which toggles between various sets of signal processing parameters (like window size, frame rate etc) to find the optimal set to process a particular region of speech. In addition, the voiced/unvoiced detector can also be improved by incorporating additional conditions, to reduce the false insertion rate. Finally, the degradation in performance that is expected when multiple speakers are present in the speech file should also be handled. One possible way to do this is to rely on creakiness to identify turn taking in the speech signal (or rely on other methods) and then set separate threshold for each speaker, to identify creakiness in the following regions.

Application to speaker ID

In terms of applications, the creakiness parameter has shown to be of use for speaker identification. The improvement in performance may not be marked, but it proves that the parameter is helping in speaker recognition by modeling some of the speakers' inherent voice qualities and speaking strategies. However, the creakiness parameter is zero for most of the frames, since creakiness is not a very common occurrence. This, when modeled in usual statistical frameworks like the Gaussian Mixture Models (GMM)(, gives a large Gaussian centered at zero, which might override the effects of other peaks nearby. However, the GMM does not have the provision to supply non-applicable data, and therefore, the performance of speaker ID is not optimal when creakiness is used as described above. Thus, there is a need to find a better scheme of modeling the data, and to use a semi-supervised approach, wherein the creakiness parameter is used only when applicable (i.e., non zero). A possible modeling scheme could be the Hidden Markov Model (HMM), or a more general approach like the Dynamic Bayesian Networks (DBN) [23]. In either case, it remains true that modeling this parameter should be done in a more useful and applicable way.

Application to ASR

The application of creakiness to speaker ID motivates its use for ASR as well. In landmark-based ASR [ref], the features that are used are acoustically motivated, as the ones that have been used in this thesis for the speaker identification experiments. During the extraction of such parameters, creakiness might cause some hindrance because of the difference in phonation that generates slightly different acoustic cues than expected. Detection of creakiness before extraction of such features could help reduce such confusions, by provision of alternative features for such regions, or compensating the effect that creakiness has. Further, identifying regions of creakiness can also help in identifying turn-taking in spontaneous speech, as speakers often tend to get creaky when they pass the turn to the next speaker.

Application to diagnosis of voice disorders

An important application of the irregular phonation detector is to help identify voice disorders by non-intrusive methods. The algorithm runs purely on the speech signal produced by the speaker, by looking at specific acoustic cues that would manifest themselves because of the activity at the vocal folds. This motivates the idea that the activity at vocal folds can be understood in an implicit way, and the specific cues found in the speech produced might give an idea about the specific physiological disorder that the speaker has, at the vocal folds.

BIBLIOGRAPHY

[1] R P Lippman, "Speech recognition by machines and humans", Speech Communication 22, pp. 1–15, 1997.

[2] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", IEEE Trans. Acoustics, Speech, and Signal Processing, vol. 28, pp. 357-366, 1980.

[3] H. Hermansky and N. Morgan, "Rasta processing of speech", IEEE Transactions on Speech and Audio Processing, 2(4), pp. 578–589, 1994.

[4] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", in Proceedings of IEEE, volume 77, pages 258–286, 1989.

[5] D.A. Reynolds and R. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing, 3, 1995.

[6] C.Y. Espy-Wilson, S. Manocha, and S. Vishnubhotla, "A new set of features for text-independent speaker identification", in Proceedings of 9th International Conference on Spoken Language Processing (ICSLP - INTERSPEECH), 2006.

[7] K. N. Stevens, "Acoustic Phonetics", MIT Press, Cambridge, Massachusetts, 1998.

[8] S. Vishnubhotla & C.Y. Espy-Wilson, "Automatic detection of irregular phonation in continuous speech", in Proceedings of 9th International Conference on Spoken Language Processing (ICSLP - INTERSPEECH), 2006.

[9] D. Klatt and L. Klatt, "Analysis, synthesis and perception of voice quality variations among female and male talkers," J. Acoust. Soc. Am. 87, 820-857, 1990.

[10] Bruce R. Gerratt and Jody Kreiman, "Toward a taxonomy of nonmodal phonation", Journal of Phonetics, 29, 365-381, 2001.

[11] Matthew Gordon & Peter Ladefoged, "Phonation types: a cross-linguistic overview", Journal of Phonetics, 29, 383-406, 2001.

[12] L Redi & S. Shattuck-Hufnagel, "Variation in the realization of glottalization in normal speakers", Journal of Phonetics, 29, 407-429, 2001.

[13] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks", Ph.D. thesis, University of Maryland, College Park, December 2004.

[14] C. T. Ishi, H. Ishiguro, N. Hagita, "Proposal of. acoustic measures for automatic detection of vocal fry", Proc. Eurospeech 2005, 481-484, 2005.

[15] K. Surana and J. Slifka, "Acoustic cues for the classification of regular and irregular phonation", in Proceedings of 9th International Conference on Spoken Language Processing (ICSLP - INTERSPEECH), 2006.

[16] T. Yoon, J. Cole, M Hasegawa-Johnson, and C. Shih. "Detecting Non-modal Phonation in Telephone Speech", UIUC ms.

[17] P., Taylor, "Analysis and synthesis of intonation using the tilt model.", J. Acoust. Soc. Am. 107(3): 1697-1714, 2000.

[18] O. Deshmukh, C. Y. Espy-Wilson, A. Salomon, and J. Singh, “Use of Temporal Information: Detection of Periodicity, Aperiodicity, and Pitch in Speech”, IEEE Transactions on Speech And Audio Processing, 13(5), 776-786, 2005.

[19] M. Plumpe, T. Quatieri, & D. Reynolds, “Modeling of the glottal flow derivative waveform with application to speaker identification”, IEEE Trans. Speech and Audio Proc., vol. 1, no. 5, pp. 569-586, 1999.

[20] D. Lewis, “Vocal resonance,” J. Acoust Soc. Am. 8, 91-99, 1936.

[21] J. Sundberg, “Articulatory interpretation of the ‘singing formant’,” J. Acoust. Soc. Am. 55, 838-844, 1974.

[22] E, Moore and M. Clements "Algorithm for Automatic Glottal Waveform Estimation Without the Reliance on Precise Glottal Information”, Proc. IEEE ICASSP 2004, pp. 101-104, 2004.

[23] K. Murphy, “Dynamic Bayesian Networks”, (Draft) To appear in “Probabilistic Graphical Models”, ed. M. Jordan, 2002.