International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

Circumstance affecting the Speech Prosody in Speech Synthesis

Mahak*

Department of CSE, Kurukshetra University Kurukshetra, India

Accepted 05 Feb 2016, Available online 27 Feb 2016, **Vol.6, No.1 (Feb 2016)**

Abstract

Speech is the key for communication between human. From the last few decades speech synthesis is becoming the most important for the communication purpose for the visual handicapped people. The primary goals of speech synthesis are to produce speech with high intelligibility and also with high quality of sound and naturalness. At present era the first parameter of goal has been reached, but the naturalness and quality are the still major problem. In this paper the focus is given on the way to produce intelligible speech with implementing proper emotion.

Keywords: *Speech Prosody etc.*

1. Introduction

Synthesized speech can be generated using several different methods, which can be classified by these three groups.

A. *Articulatory synthesis*: It models the human speech production system directly.
B. *Formant synthesis:* It is based on source filter model, and models the pole frequencies of speech signal or transfer function of vocal tract.
C. *Concatenative synthesis:* Use recorder speech sample for speech synthesis.

The formant and concatenative methods are the most commonly used in today's speech synthesis era. But presently concatenative speech synthesis procedure is chosen at most and is becoming more and more popular. The cause is simplicity and fewer complexes.
Every TTS (Text to Speech) system has its own implementation, but the basic steps to implement a TTS synthesizer are somewhat same. At first the text is given as input to the system, the tokenization of text is done. Then every token is identified with a tag and token is converted to word. At this particular step pronunciation generation is done with the help of LTS (Letter to Sound) rules, lexsion etc. After generation the pronunciation the most vital and complex part of TTS system i.e. prosody generation can be done. Prosody is basically a combination of intonation, duration, pitch etc. Then from the generated prosody wave form is rendered with the help of labeled token/databases, and speech output is generated. The pictorial view of TTS is given here.

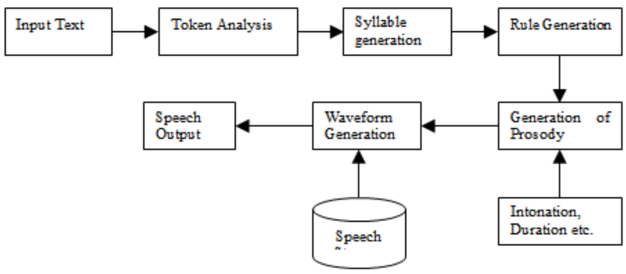

Fig 1: Building blocks of a TTS

As our primary focus is to get highly natural speech, that why we will discuss about prosody later on.

2. Prosody and its effect on speech synthesis

Prosody is the intonation, rhythm, and the lexical stress on the speech. The prosodic features of a unit of speech called Suprasegmental Features. The suprasegmental features affects all the segment of the unit.

Prosody can

1. Reduce the load on listener.
2. Help to disambiguate the meaning of the sentence.

It is very much essential to identify the good models for prosodic phrase. Prosody not only increase the intelligibly but also the naturalness which is the biggest concern for a TTS system, and thus have a major influence on their performance. For example,

1. Pronunciation generation module which gives the phoneme sequence needed to synthesize an utterance

*Corresponding author is a PhD Scholar

introduces *pause* at phrase boundaries and might even introduce some allophonic variation of the same phoneme at the phrase boundaries.
2. The duration module lengthens the segments which occur immediately prior to a phrase boundary.

Various models have been proposed which are ranging from simple deterministic rule like rule based on the punctuation to the complex model that requires full syntax parsing of the sentence to be synthesized. Research is going on the location of phrase boundaries focusing primarily on the relationship between prosodic structure and the naturalness, and it is using some syntactic information to predict prosodic boundaries in the form of some basic rules. But these hand-written rules are very difficult to write, understand, implement and also adapt in the new domain or language.

To avoid these problems, recently experts are focusing on the design techniques on acquiring phrasing rules automatically by using samples of data that requires large prosodically labeled segments. The techniques they are using are CART (Classification and Regression tree) model, Dependency Graph, ToBI (Tone and Break Indices) model etc. Among all those techniques , as TOBI (Tone and Break Indices) is now leading the prosodic world for its less complex design, and better natural output , so TOBI will be the main topic of discussion from the latter on.

3. TOBI

One of the most widely used linguistic models of prosody is TOBI (Tone and Break Indices) model (Silverman et al., 1992; Beckman and Hirschberg, 1994; Pierrehumbert, 1980; Pitrelli et al., 1994).

TOBI is a phologic theory of intonation which models prominence, tune, and boundaries. ToBI's model of prominence and tunes is based on the 5 pitch accents and 4 boundary tones shown in Figure below.

Table 1: ToBI classification

Pitch Accent		Boundary Tones	
H*	Peak Accent	L-L%	Final Fall
L*	Low Accent	L-H%	Continuous Rise
L* + H	Scooped Accent	H-L%	Question rise
L + H*	Rising Peak Accent	H-H%	Final Level Platue
H + !H*	Step Down		

How ever their effect on speech can be studied in terms of

1. Acoustic segment pattern. - Break Indices (BI)
2. Tone labeling for the phrase ending.-Tone (To)
3. The stress pattern on vocal.-Prominence

In addition to accent and boundary tones, TOBI distinguishes 5 level of phrasing which are labeled on a separate BI tier.

Labeling is demonstrated in the following table.
In addition of pitch accent and boundary, ToBi also explains 5 level of phrasing which are labeled on a separate BI tier.
Labeling is demonstrated in the following table.

Table 2: Phrasing level of BI tier

Label No	Pattern
0	Tight Juncture
1	Normal Phrase Medial Word Bound Boundaries
2	Pause betn. Words
3	Intermediate Phrase
4	Intonational Phrase(Most Disjoint)

The uses of this three tier (BI, To, Prominence) can be explained by using this example (from analysis).

Consider the utterance

W: We are from the campus of GGSIP University
BI: 1 1 3 1 3 1p 3p 4
To: L-L% H-H% L-L%
P: L* L+H* H* H +!H*

Additionally, diffluent word boundaries are marked with 'p'

a) 1p: Words that are sensed to pronounce in short.
b) 2p: Hesitation when speaker is searching for word.
c) 3p: Hesitation accompanying for infrequent accent.

At intermediate intonational phrase, the possible tone includes:

a) L-: Ends in a low value relative to the rest of the phrase.
b) H-: Ends in a high value relative to the last pitch accent.
c) !H-: Ends at a mid range point.

Here is a pictorial example which relates ToBI and wave form.

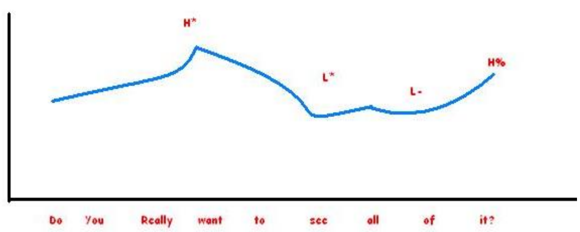

Fig 2: ToBI and wave form

4. Computing Duration from prosodic labeling

The most well-known of the rule-based methods is the method of Klatt (1979), which uses rules to model how the average or 'context-neutral' duration of a phone. Each rule is associated with a duration multiplicative factor; some examples:

a) *Prepausal Lengthening:* The vowel or syllabic consonant before a pause is lengthened by 1.4.
b) *Non-phrase-final shortening:* Segments which are not phrase-final are shortened by 0.6.
c) Phrase-final postvocalic liquids and nasals are lengthened by 1.4.
d) *Unstressed Shortening:* Unstressed segments are more compressible, so their minimum duration is halved, and are shortened by .7 for most phone types.
e) *Lengthening for Accent:* A vowel which bears accent is lengthened by 1.4
f) *Shortening in Clusters:* A consonants followed by a consonant is shortened by 0.5.
g) *Pre-voiceless shortening:* Vowels are shortened before a voiceless plosive by 0.7.

5. Guiding principles for ToBI systems

a) *Accurate as possible:* based on a rigorous analysis of the intonational phonology
b) *Does **not** replace a permanent record of the speech signal:* tagging, not encoding of signal
c) *Efficient:* only transcribe phenomena not automatically retrievable from signal
d) *Use not limited to a few experts:* easy to teach with freely available manual
e) *Consistent:* inter-transcriber consistency testing across sites

6. Labeling utterance using pitch pattern

Using PRAAT program, we can generate the intensity pattern, pitch patter, spectrogram of speech. From the pitch pattern the labeling can be done. Fig. 3 describing the example.

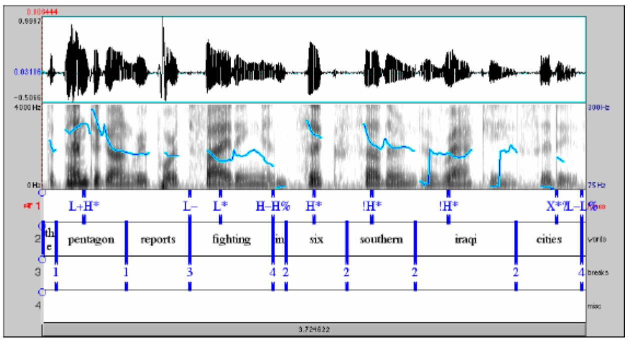

Fig.3 Labeling utterance using pitch pattern

In this graph top most wave from shows the wave pattern of utterance, and the second is showing the pitch pattern. According to this pattern the labeling is done automatically. To do the labeling automatically, the labeling engine should be trained using some intelligent rules. The pitch range model for English includes a upline and a bottomline. The upline is determined at every max. fundamental frequency, and the bottomline shows the minimum pitch value.

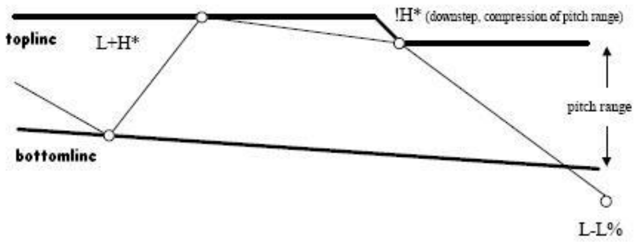

Fig. 4 Pitch range for American English

The following examples shows how from rhis pitch range proper prosody can be generated for different emotions.

Utterance:
Marianna made the marmalade.
Default
1. T: H* H* L-L
B: 1 1 1 4
Emphasis on Marianna
2. H* L-L%
Contrastive
3. L+H* L-L%
Incredulous
4. L* H-H%
DoublyIncredulous
5. L* L* H-H%
(2_intonation_phrases)

6. L+H*L-H% L* H* L-L% 4 1 1 4

He's only a millionaire.
comment
H* H* L-H%
exaggerated surprise
L+H* L+H* L-H%

ToBI labelling is fun.
default
H* H*L-L%
skeptical re fun
H*+L H*L-H%
Insisting
H* H* L-L%

These example have been taken from

7. A Reality

Now apart from this labeling, now we can concentrate on the factors affecting prosody. The emotion is the main factor that have to be handled in ToBI. For example , for every emotional utterance like anger, question, normal, the pitch pattern, intensity pattern of the wave form is changed. A details look is described in the following figures, from which a neat idea we can get about how the emotion can be generated using ToBI.

Fig. 5 is showing the comparison of intensity between normal, question and anger utterance.

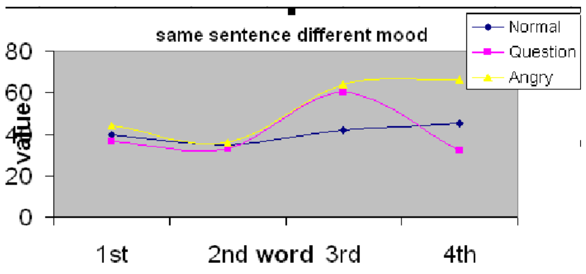

Fig 5 Comparison between different emotional intensity

It shows that the L* and the H* are varying deeply. The highest is for anger utterance, but the sentence is same for all three types.

Fig 6, 7,& 8 is showing comparison about the pitch range wave form and intensity waveform for those emotion defined earlier. The waveform has been generated using PRAAT software. All waveform are for same sentence: You Are Going There.

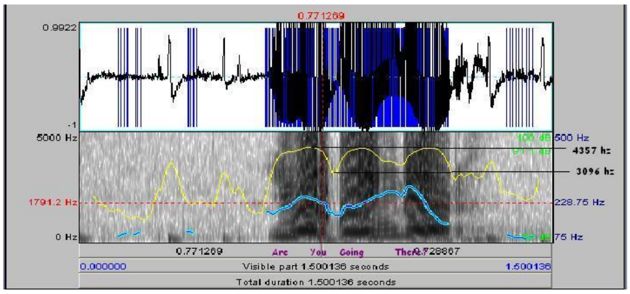

Fig. 7: The waveform for anger emotion

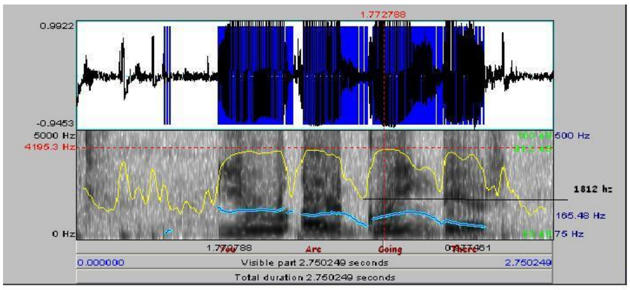

Fig. 8: The waveform for normal emotion

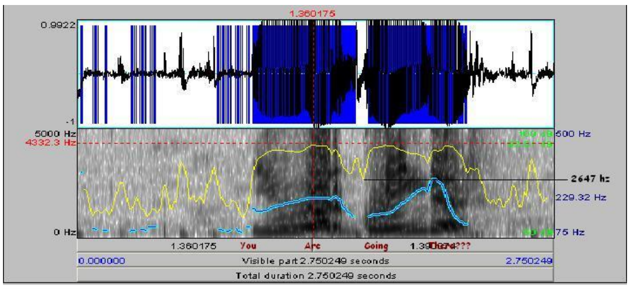

Fig. 9: The waveform for question emotion

From these figures, statistically the following table can be drawn.

Table 3: comparison table for emotion

	H*(Hz)	L*(Hz)	Duration(sec)
Anger	4357	3096	1.5
Normal	4195	1812	2.76
Question	4332	2647	2.76

So it can be said that with the labeling of waveform, the topline, bottomline, their differences and the duration of speech also differs regularly. So during ToBI framework design these matters should be considered.

8. Factors affecting Prosody, in details

From the last section, an idea can be generated about how speech parameters can affect emotion during speech synthesis, but these parameters are not limited within the duration and difference of intensity, but these also depends on max pitch, min pitch, mean pitch, mean energy also. Here a compilation is given with details table, and using those tables, the graphs will show how these parameters affect prosody.

For compilation initially 10 sentences are selected, each of with 4 words and they are recorded in different emotion (Anger, Normal, Happy, Question, Whisper). And then from signal behavior the table 4 is generated.

8. Factors affecting Prosody, in details

From the last section, an idea can be generated about how speech parameters can affect emotion during speech synthesis, but these parameters are not limited within the duration and difference of intensity, but these also depends on max pitch, min pitch, mean pitch, mean energy also. Here a compilation is given with details table, and using those tables, the graphs will show how these parameters affect prosody.

For compilation initially 10 sentences are selected, each of with 4 words and they are recorded in different emotion (Anger, Normal, Happy, Question, Whisper). And then from signal behavior the table 4 is generated.

Table 4: Parameters affecting speech prosody

ANGER	Intensity (Max)(HZ)	Intensity (Min) (HZ)	Intensity (Difference) (HZ)	Pitch (Max) (HZ)	Pitch (Min) (HZ)	Pitch (Difference) (HZ)	Avg. Pitch	Mean Energy (DB)	Duration (Sec)
1	4381	2053	2328	498	127	371	198	89	1.7647
2	4370	2431	1939	442	126	316	197	90	1.0000
3	4388	1623	2715	413	115	298	212	89	0.77
4	4340	1849	2491	233	158	75	211	90	0.67
5	4358	2736	1622	251	121	130	205	73	0.75
6	4376	2033	2343	237	138	99	200	90	1.13
7	4328	2420	1908	229	132	97	191	88	0.72
8	4314	2641	1673	258	103	155	194	92	1.095
9	4365	1520	2845	227	137	90	196	84	1.00
10	4350	1938	2412	271	156	115	217	85	1.56
NORMAL	Intensity (Max)(HZ)	Intensity (Min) (HZ)	Intensity (Difference) (HZ)	Pitch (Max) (HZ)	Pitch (Min) (HZ)	Pitch (Difference) (HZ)	Avg. Pitch	Mean Energy (DB)	Duration (Sec)
1	2711	826	1885	298	141	175	198	70	2.00
2	2631	661	1970	239	164	75	195	74	1.17
3	2595	77	2518	233	158	75	191	64	0.87
4	2995	1037	1958	207	161	46	186	71	1.21
5	3440	713	2727	223	142	81	183	77	0.99
6	2591	135	2456	219	139	80	187	66	1.49
7	2233	229	2004	210	130	80	164	52	1.48
8	2986	957	2029	240	160	80	198	62	1.60
9	2871	603	2268	219	142	77	186	64	1.24
10	3490	474	3016	226	152	74	195	63	1.812

HAPPY	Intensity (Max)(HZ)	Intensity (Min) (HZ)	Intensity (Difference) (HZ)	Pitch (Max) (HZ)	Pitch (Min) (HZ)	Pitch (Differenc e) (HZ)	Avg. Pitch	Mean Energy (DB)	Durat ion (Sec)
1	3872.8	1191.8	2681	488.27	102	386.27	162.62	83.80	2.07
2	4184	1580.8	2603.2	225	96	129	169	87	1.09
3	4113	2580	1533	486	122.29	364.29	193	87	0.82
4	4275	1500	2775	188	102	86	161	84	1.16
5	4250	1752	2498	470	121	349	156	89	0.84
6	4063.9	1543.6	2519.4	171	126.44	44.56	149	85	1.41
7	4169.2	1683	2486.2	207	134	73	168	86	0.95
8	4097	1014	3083	480	91	389	168	86	1.48
9	4159.2	835	3324.2	185	128	57	159	87	1.07
10	4175	1461.3	2713.7	226	103	123	170	85	1.61
QUESTIO N	Intensity (Max)(HZ)	Intensity (Min) (HZ)	Intensity (Difference) (HZ)	Pitch (Max) (HZ)	Pitch (Min) (HZ)	Pitch (Differenc e) (HZ)	Avg. Pitch	Mean Energy (DB)	Durat ion (Sec)
1	4220	1228	2992	333	103	230	202	87	1.87
2	4195	1645	2550	297	157	140	208	88	1.02
3	4127	2827	1300	267	119	148	193	87	0.72
4	4215	1489	2726	422	142	280	186	88	1.11
5	4175	3026	1149	305	156	149	203	88	0.86
6	4346	2088	2258	397	157	240	242	88	1.18
7	4306	1331	2975	277	144	133	183	88	0.88
8	4183	1216	2967	420	99	321	252	86	1.35
9	4199	1765	2354	484	93	391	212	87	1.15
1	4289	1015	3274	272	151	121	218	86	1.61
WHISPER	Intensity (Max)(HZ)	Intensity (Min) (HZ)	Intensity (Difference) (HZ)	Pitch (Max) (HZ)	Pitch (Min) (HZ)	Pitch (Differenc e) (HZ)	Avg. Pitch	Mean Energy (DB)	Durat ion (Sec)
1	4118	759	3359	473	86	387	157	79	3.07
2	4133	916	3217	429	319	110	397	84	1.22
3	4122	1099	3023	480	441	39	467	84	1.26
4	3926	1280	2646	478	108	370	201	82	1.26
5	4020	1461	2559	485	462	23	471	84	1.00
6	4115	877	3238	499	104	395	349	84	1.50
7	4051	1058	2993	161	89	72	109	81	1.21
8	4145	872	3273	491	101	390	329	84	1.72
9	4160	1000	3160	458	103	355	234	85	1.47
10	4279	970	3309	476	84	392	174	82	2.05

Now from this above table we can generate the following graphs, which are showing how different parameters are affecting the prosodic behavior. The parameters are Intensity Difference (Max –Min), Pitch Difference, Avg. Pitch, Mean Energy, and Duration.

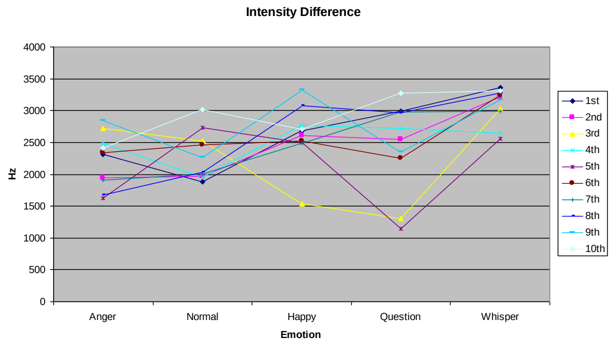

Fig. 10: Intensity Difference affecting Prosody

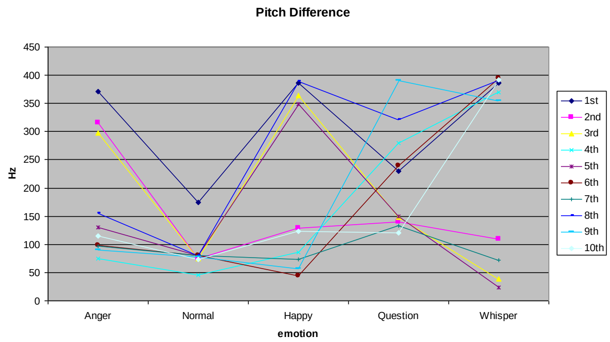

Fig. 11: Pitch Difference affecting prosody

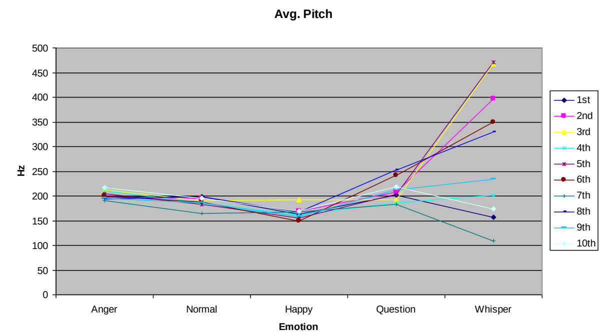

Fig. 12: Avg. Pitch affecting prosody

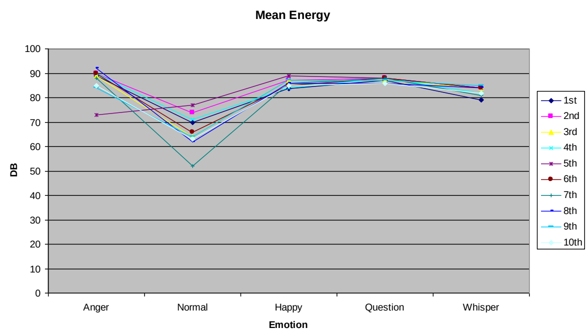

Fig. 13: Mean Energy of Intensity affecting prosody

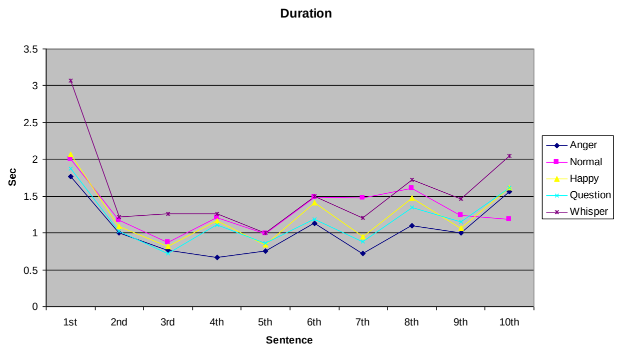

Fig. 14: Relation between Duration and Prosody

Now from table 4, fig. 10-14, we can derive the following relationship between emotion depending on parameters as shown in table 5-9

Table 5: Relationship between emotions on Intensity Difference (HZ)

	Anger	Normal	Happy	Question	Whisper
Anger	------	<	<	<	<
Normal	>	------	<	<	<
Happy	>	>	------	>	<
Question	>	>	<	------	<
Whisper	>	>	>	>	------

Table 6: Relationship between emotions on Pitch Difference (HZ)

	Anger	Normal	Happy	Question	Whisper
Anger	------	>	<	>	<
Normal	<	------	<	<	<
Happy	>	>	------	>	=
Question	<	>	<	------	<
Whisper	>	>	>	>	------

Table 7: Relationship between emotions on Mean Pitch (HZ)

	Anger	Normal	Happy	Question	Whisper
Anger	------	>	>	<	<
Normal	<	------	>	<	<
Happy	<	<	------	<	<
Question	>	>	>	v	<
Whisper	>	>	>	>	------

Table 8: Relationship between emotions on Mean Energy (DB)

	Anger	Normal	Happy	Question	Whisper
Anger	------	>	>	>	>
Normal	<	------	<	<	<
Happy	<	>	------	>	>
Question	<	>	<	------	>
Whisper	<	>	<	<	------

Table 9: Relationship between emotions on Duration (Sec)

	Anger	Normal	Happy	Question	Whisper
Anger	------	<	<	<	<
Normal	>	------	>	>	<
Happy	>	<	------	>	<
Question	>	<	<	------	<
Whisper	>	>	>	>	------

So from these tables, figures, and compilations, it is cleared that these parameters (Difference in Intensity, Pitch Difference, Mean Pitch, Mean Energy, and Duration) have significant affects on prosody, which must have to be taken in consideration during synthesizing a signal for speech.

As prosody is quite language specific, different version of ToBI exist for individual languages (like MAE_ToBI for English, KToBI for Korea). But unfortunately still it has not been developed for Indian Languages. So our future aim is to model a ToBI standard for Hindi Language such that it can help in the prosodic implementation of speech synthesis for Indian Languages.

Reference

N. Sridhar Krishna, Hema A. Murthy A New Prosodic Phrasing Model for Indian Language Telugu

Daniel Jurafsky & James H. Martin. November 27, 2006 Speech and Language Processing: An introduction to natural language processing, computational linguistics, and speech recognition.

Izhak Shafran, Nov 29, 2005 Prosody in Spoken Language Processing.

Martine Grice, The ToBI Framework, Institute of Phonetics.

OpenCourseWare, August 09,2006.

Matthias Jilka, Gregor Möhler, Grzegorz Dogil, Rules for the Generation of ToBI-based American English Intonation