Text to Speech Synthesis with Prosody feature: Implementation of Emotion in Speech Output using Forward Parsing

M.B.Chandak chandakmb@gmail.com
Department of Computer Science and Engineering
Shri Ramdeoababa Kamla Nehru Engineering College,
Nagpur, INDIA

Dr.R.V.Dharaskar rvdharaskar@rediffmail.com
Department of Computer Science and Engineering
G.H.Raisoni College of Engineering,
Nagpur, INDIA

Dr.V.M.Thakre thakrevm@gmail.com
Department of Computer Science and Engineering
AMRVATI UNIVERSITY,
Amravti, INDIA

Abstract

One of the key components of Text to Speech Synthesizer is prosody generator. There are basically two types of Text to Speech Synthesizer, (i) single tone synthesizer and (ii) multi tone synthesizer. The basic difference between two approaches is the prosody feature. If the output of the synthesizer is required in normal form just like human conversation, then it should be added with prosody feature. The prosody feature allows the synthesizer to vary the pitch of the voice so as to generate the output in the same form as if it is actually spoken or generated by people in conversation.
The paper describes various aspects of the design and implementation of speech synthesizer, which is capable of generating variable pitch output for the text. The concept of forward parsing is used to find out the emotion in the text and generate the output accordingly.

Keywords: Text to speech synthesizer, Forward Parsing, Emotion Generator, Prosody feature.

1. INTRODUCTION

Prosody is one of the key components of Speech Synthesizers, which allows implementing complex weave of physical, phonetic effects that is being employed to express attitude, assumptions, and attention as a parallel channel in our daily speech communication. In general any communication is collection of two phases: *Denotation*, which represents written content or spoken content and *Connotation*, which represent emotional and attentional effects intended by the speaker or inferred by a listener. Prosody plays important role in guiding listener for speaker attitude towards the message, towards the listener and towards the complete communication event. [2,3,4]
From listener point of view, prosody consists of systematic perception and recovery of speaker intentions based on: [3,4]

a) Pauses: To indicate phrases and separate the two words
b) Pitch: Rate of vocal fold cycle as function of time
c) Rate: Phoneme duration and time
d) Loudness: Relative amplitude or volume.

2. ARCHITECTURE FOR PROSODY GENERATION

The Figure 1, shows the basic architecture of prosodic generator and various elements of prosodic generation in TTS, from pragmatic abstraction to phonetic realization. The input of the prosody module in Figure 1; is parsed text with a phoneme string, and the output specifies the duration of each phoneme and the pitch contour. Before providing input to the prosody generator, the input is parsed and is converted into phonemes depending upon the key stokes involved in the characters present in the input. The standard phonetic vocabulary of English language is used in conversion of text to phoneme. The duration and pitch of each phoneme depends upon the content and context of the text [6,7]. For example in the context, the mood of conversation is happy, then pitch of the words is changed accordingly to allow listener to understand "happy" mood of the content. Similarly, if after some time period the mood and emotion in the text are changed, then words pronounced in voice format should be accordingly modified in pitch sense to generate the desired effects. Prosody has an important supporting role in guiding a listener's recovery of the basic messages (denotation).

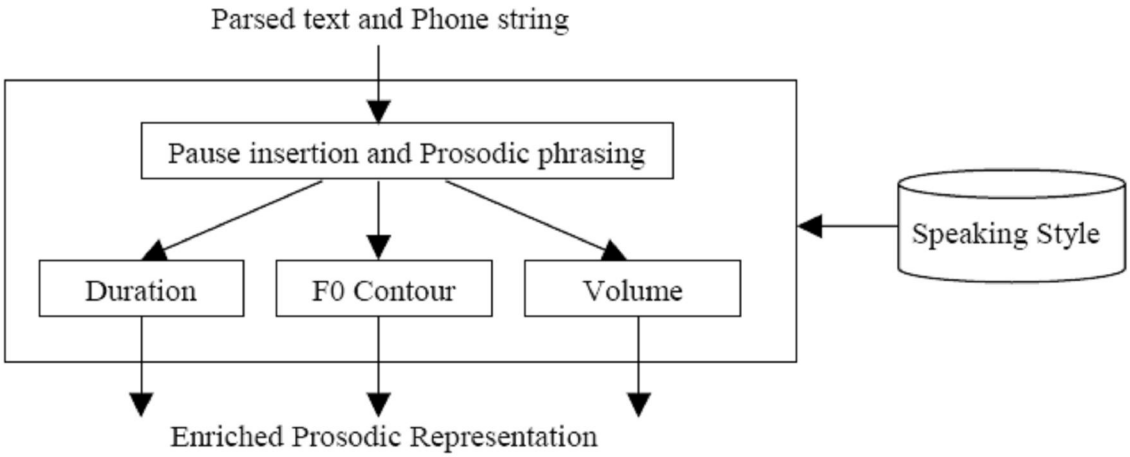

Figure 1: Architecture of Prosody Generator.

The various modules of Prosody Generator are described in detail as follows:

1) **Speaking Style:** Prosody depends not only on the linguistic content of a sentence. Different people generate different prosody for the same sentence. Even the same person generates a different prosody depending on his or her mood. The *speaking style* of the voice in Figure 1, can impart an overall tone to a communication. Examples of such global settings include a low register, voice quality (falsetto, creaky, breathy, etc), narrowed pitch range indicating boredom, depression, or controlled anger, as well as more local effects, such as notable excursion of pitch, higher or lower than surrounding syllables, for a syllable in a word chosen for special emphasis. The various parameter which influence the speaking Style are [8,9]:
   a. **Character:** Character, as a determining element in prosody, refers primarily to long-term, stable, extra-linguistic properties of a speaker, such as membership in a group and individual personality. It also includes socio-syncratic features such as a speaker's region and economic status, to the degree that these influence characteristic speech patterns. In addition, idiosyncratic features such as gender, age, speech defects, etc. affect speech, and physical status may also be a background determiner of prosodic character. Finally, character may sometimes

include temporary conditions such as fatigue, inebriation, talking with mouth full, etc. Since many of these elements have implications for both the prosodic and voice quality of speech output, they can be very challenging to model jointly in a TTS system. The current state of the art is insufficient to convincingly render most combinations of the character features listed above.[5,7]

b. Emotion: Temporary emotional conditions such as amusement, anger, contempt, grief, sympathy, suspicion, etc. have an effect on prosody. Just as a film director explains the emotional context of a scene to her actors to motivate their most convincing performance, so TTS systems need to provide information on the simulated speaker's state of mind [11,12]. These are relatively unstable properties, somewhat independent of character as defined above. That is, one could imagine a speaker with any combination of social/dialect/gender/age characteristics being in any of a number of emotional states that have been found to have prosodic correlates, such as anger, grief, happiness, etc. Emotion in speech is actually an important area for future research. A large number of high-level factors go into determining emotional effects in speech. Among these are point of view (can the listener interpret what the speaker is really spontaneous vs. symbolic (e.g., acted emotion vs. real feeling); culture-specific vs. universal; basic emotions and compositional emotions that combine basic feelings and effects; and strength or intensity of emotion. We can draw a few preliminary conclusions from existing research on emotion in speech.

Some basic emotions that have been studied in speech include:

a) **Anger**, though well studied in the literature, may be too broad a category for coherent analysis. One could imagine a threatening kind of anger with a tightly controlled F0, low in the range and near monotone; while a more overtly expressive type of tantrum could be correlated with a wide, raised pitch range [7].
b) **Joy** is generally correlated with increase in pitch and pitch range, with increase in speech rate. Smiling generally raises F0 and formant frequencies and can be well identified by untrained listeners.
c) **Sadness** generally has normal or lower than normal pitch realized in a narrow range, with a slow rate and tempo. It may also be characterized by slurred pronunciation and irregular rhythm.
d) **Fear** is characterized by high pitch in a wide range, variable rate, precise pronunciation, and irregular voicing (perhaps due to disturbed respiratory pattern).

2) SYMBOLIC PROSODY

It deals with two major factors:

a) Breaking the sentence into prosodic phrases, possibly separated by pauses, and
b) Assigning labels, such as emphasis, to different syllables or words within each prosodic phrase [2,3].

Words are normally spoken continuously, unless there are specific linguistic reasons to signal a discontinuity. The term *juncture* refers to prosodic phrasing—that is, where do words cohere, and where do prosodic breaks (pauses and/or special pitch movements) occur.

The primary phonetic means of signaling juncture are:

i. Silence insertion.
ii. Characteristic pitch movements in the phrase-final syllable.
iii. Lengthening of a few phones in the phrase-final syllable.
iv. Irregular voice quality such as vocal fry

The block diagram of the pitch generator decomposed in Symbolic and phonetic prosody is as shown in the Figure 2.

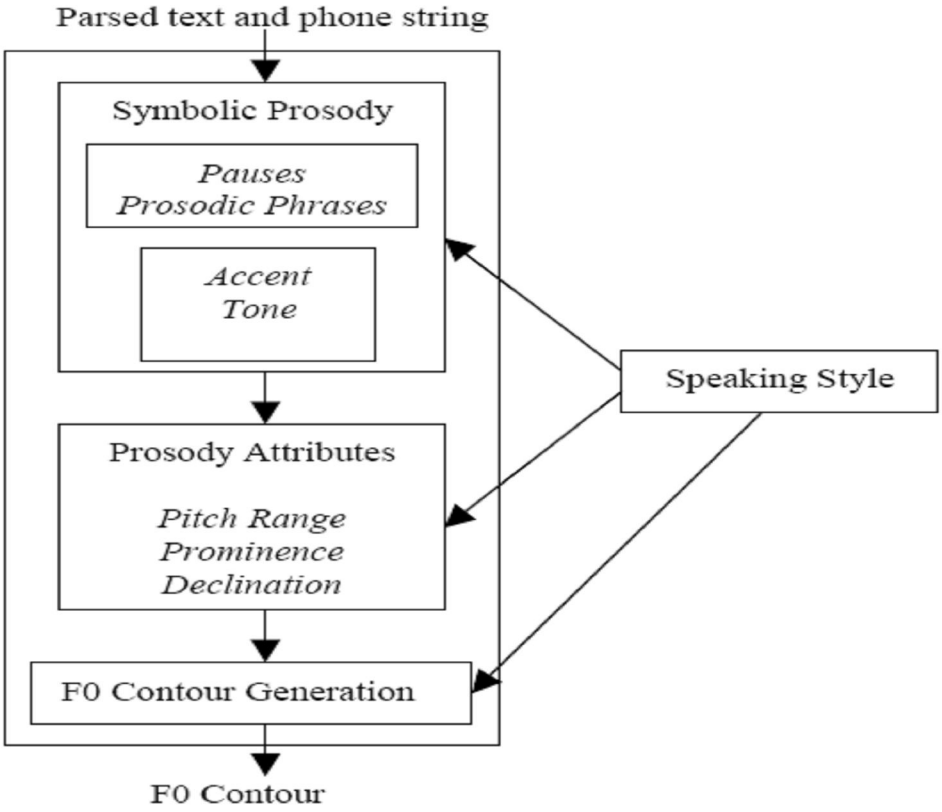

Figure 2: Pitch generator decomposed into Symbolic and phonetic prosody.

The various components are described in detailed in the following discussion.

1. **Pause:**
   The main aim to insert pause in running text is to structure the information which is generated in the form of voice output. In typical systems, the reliable location which indicates the insertion of pause is pronunciation symbols [5].
   In predicting pauses it is necessary to consider their occurrence and their duration, the simple presence or absence of a silence (of greater than 30 ms) is the most significant decision, and its exact duration is secondary, based partially on the current rate setting and other extraneous factors.
   The goal of a TTS system should be to avoid placing pauses anywhere that might lead to ambiguity, misinterpretation, or complete breakdown of understanding. Fortunately, most decent writing (apart from email) incorporates punctuation according to exactly this metric: no need to punctuate after every word, just where it aids interpretation
2. Prosodic Phrases:
   Based on punctuation symbols present in the text, commercial TTS systems are using the simple rules to vary the pitch of text depending on the prosodic phrases, for example if in the text comma symbol appears the next word will be in the slightly higher pitch than the current pitch [11].
   The tone of particular utterance is set by using standard indices called as ToBI (Tone and Break Indices). These are standard for transcribing symbolic intonation of American English utterances, and can be adapted to other languages as well.

   The *Break Indices* part of ToBI specifies an inventory of numbers expressing the strength of a prosodic juncture. The Break Indices are marked for any utterance on their own discrete *break index tier* (or layer of information), with the BI notations aligned in time with a representation of the speech phonetics and pitch track. On the break index tier, the prosodic association of words in an utterance is shown by labeling the end of each word

for the subjective strength of its association with the next word, on a scale from 0 (strongest perceived conjoining) to 4 (most disjoint), defined as follows: [5]

3. PROSODIC TRANSCRIPTION SYSTEM

This system is used to introduce the prosodic parameters to the tones used to generate the voice output. The system is so designed that it is capable of handling both qualitative and quantitative aspect of tones by generating necessary "curve" structure. The curve represents the final pitch used to tone the particular word. The tone is determined by calculating "TILT". Following parameters are used to calculate "TILT" [11,12]

- starting f0 value (Hz)
- duration
- amplitude of rise (*Arise*, in Hz)
- amplitude of fall (*Afall*, in Hz)
- starting point, time aligned with the signal and with the vowel onset

The tone shape, mathematically represented by its *tilt*, is a value computed directly from the f0 curve by the following formula:

$$tilt = \frac{\left|A_{rise}\right| - \left|A_{fall}\right|}{\left|A_{rise}\right| + \left|A_{fall}\right|}$$

The label schemes for the syllable to calculate the TILT is as shown in the table. These labels identify the specific syllable and alter the tone based on the presence of the syllable.

Sil	Silence / Pause
C	Connection
A	Major Pitch accent
Fb	Falling boundary
Rb	Rising boundary
Aft	After falling boundary
Arb	Accent + Rising boundary
M	Minor accent
Mfb	Minor accent + Falling boundary
Mrb	Minor accent + Rising boundary
L	Level accent
Lrb	Level accent + Rising boundary
Lfb	Level accent + Falling boundary

The likely syllable for "TILT" analysis in the contour can be automatically detected based on high energy and relatively extreme F0 values or movements.

4. DURATION ASSIGNMENT

There are various factors which influence the phoneme durations. The common factors are

a. Semantic and Pragmatic Conditions
b. Speech rate relative to speaker intent, mood and emotion
c. The use of duration or rhythm to possibly signal document structure above the level of phrase or sentence [5]
d. The lack of a consistent and coherent practical definition of the phone such that boundaries can be clearly located for measurement

One of the commonly used methods for Duration Assignment is called as Rule based method. This method uses table lookup for minimum and inherent duration for every phone type. The duration is rate dependent, so all phones can be globally scaled in their minimum duration for faster or slower rates. The inherent duration is raw material and using the specified rules, it may be stretched or contracted by pre-specified percentage attached to each rule type as specified and then it is finally added back to the minimum duration to yield a millisecond time for a given phone.

The duration of phone is expressed as

$$d = d_{min} + r(\overline{d} - d_{min})$$

Where d_{min} is the minimum duration of the phoneme, d is average duration of the phoneme and correction "r" is given by:

$$r = \prod_{i=1}^{N} r_i$$

For the case of N rules applied, where each rule has correction r_i. At the very end, a rule may apply that lengthens vowels when they are preceded by voiceless plosives.
The list of rules used for calculating duration as follows:

Lengthening of final vowel and following consonant in prepausal syllables
Shortening of all syllabic segments in non-prepausal positions
Shortening of syllabic segments if not in a word final syllable
Consonant in non word initial positions are shortened
Un-stressed and secondary stressed phones are shortened
Emphasized vowels are lengthened
Vowels may be shortened or lengthened according to phonetic features of their context.
Consonants may be shortened in cluster

5. PITCH GENERATION

Since generating pitch contours is an incredibly complicated problem, pitch generation is often divided into two levels, with the first level computing the so-called symbolic prosody described in Section 2 and the second level generating pitch contours from this symbolic prosody. This division is somewhat arbitrary since, as we shall see below, a number of important prosodic phenomena do not fall cleanly on one side or the other but seem to involve aspects of both. Often it is useful to add several other attributes of the pitch contour prior to its generation, which is discussed in coming section.

5.1 Pitch Range:

Pitch range refers to the high and low limits within which all the accent and boundary tones must be realized: a floor and ceiling, so to speak, which are typically specified in Hz. This may be considered in terms of stable, speaker-specific limits as well as in terms of an utterance or passage.

5.2: Gradient Prominence:

Gradient prominence refers to the relative strength of a given accent position with respect to its neighbors and the current pitch-range setting. The simplest approach, where every accented syllable is realized as a High tone, at uniform strength, within an invariant range, can sound unnatural.

5.3: Declination

Related to both pitch range and gradient prominence is the long-term downward trend of accent heights across a typical reading-style, semantically neutral, declarative sentence. This is called *declination*.

5.4: Phonetic F0: Micro prosody

Micro prosody refers to those aspects of the pitch contour that are unambiguously phonetic and that often involve some interaction with the speech carrier phones.

6. BLOCK DIAGRAM OF FORWARD PARSING METHOD

6.1: Methodology:

Parsing is a method of scanning the text, in order to determine various points such as content of text, context of text, frequency of particular word in the text etc. While finding out the emotions present in the text, it is necessary to determine context of text. The context of the text determines the current emotions present in the text and also used to find variation in the emotion. Most of the

commercial available TTS are based on regular parsing in which the emotion present in the text is generated at the same time when the text is converted and represented in the voice form to the user. This approach followed in current text to speech synthesizers, generates delay, and reduces naturalness of the speech. [12]
The basic requirement of the system is emotion present within the text should be known before hand so that it can be used to alter the pitch of the words present in the text. This will remove the delay component as well as the voice generated will be similar to natural voice. For example if the text is consist of three paragraphs, then, when the first paragraph is presented to user in voice format, scanning of next two paragraphs is performed, and emotion present in the paragraphs is derived. This emotion is then used as pitch alteration component and will act as intensifier. The intensifier may be high, low or neutral. The value of intensifier then can be used to alter the pitch of the text present. To handle first paragraph, the pre-processing phase is performed on first paragraph, this pre-processing will scan the first paragraph and generates the emotion present within the first paragraph.

6.2: Architecture for implementing Forward Parsing

The block diagram for implementing forward parsing is as shown in the figure.

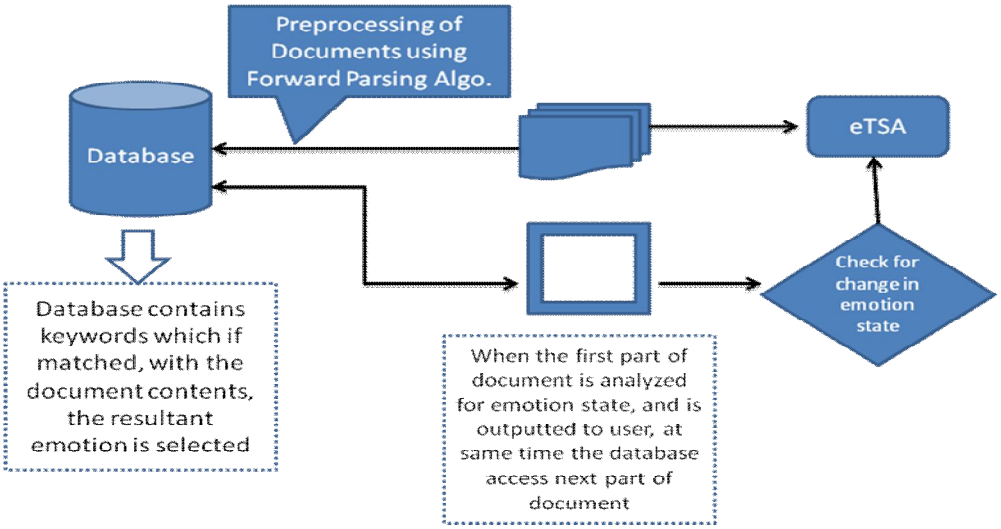

Figure 3: Architecture for Forward Parsing

As shown in the figure 3, a Database is maintained, which contains the keywords and category of emotion to which it belongs. Following types of emotions are handled using the architecture.
Anger, Joy, Surprise, Disgust, Contempt, Pride, Depression, Funny, Sorry, Boredom, Suffering, Shame
The text is scanned and keywords present in the text are compared with the contents of database. The comparison will finalize the value of emotion. Once the type of emotion is fixed the information is supplied to the composer, which then composes the wave file based on value of emotion. The value of emotion is changed based on intensity of emotion in the text. For example if the text is
I am happy: Then the intensity of emotion happy is normal and will be represented by <+>
I am very happy: Then the intensity is increase and will be represented by <++>
This methodology will help in varying the pitch of the keyword “happy”.

6.3: Prosodic Markup Language

To incorporate the emotion component in the text and allow the synthesizer to determine the intensity of the particular word in the text following tags are designed and the text is modified
For prosodic processing, text may be marked with tags that have scope, in the general fashion of XML. Some examples of the form and function of a few common TTS tags for prosodic

processing are discussed below. Other tags can be added by intermediate subcomponents to indicate variables such as accents and tones [10].

a. **Pause** or **Break:** These commands can accept absolute duration of silence in millisecond or relative duration of silence like large, medium or small. For example, a "," (comma) in text may allow to pause for some duration and then continue the next part of text.
b. **Rate:** This parameter controls the speed of output. The usual measurement is *words per minute*, which can be a bit vague, since words are of very different durations. However, this metric is familiar to many TTS users and works reasonably well in practice.
c. **Baseline Pitch:** This parameter specifies the desired average pitch: a level around which, or up from which, pitch is to fluctuate.
d. **Pitch Range:** It specifies within what bounds around the baseline pitch level line the pitch of output voice is to fluctuate.
e. **Pitch:** This parameter commands can override the system's default prosody, giving an application or document author greater control. Generally, TTS engines require some freedom to express their typical pitch patterns within the broad limits specified by a Pitch markup.
f. **Emphasis:** This parameter emphasizes or deemphasizes one or more words, signaling their relative importance in an utterance. Its scope could be indicated by XML style tag. Control over emphasis brings up a number of interesting considerations. For one thing, it may be desirable to have degrees of emphasis [11]. The notion of gradient prominence—the apparent fact that there are no categorical constraints on levels of relative emphasis or accentuation—has been a perpetual thorn in the side for prosodic researchers. This means that in principle any positive real number could be used as an argument to this tag. In practice, most TTS engines would artificially constrain the range of emphasis to a smaller set of integers, or perhaps use semantic labels, such as *strong*, *moderate*, *weak*, *none* for degree of emphasis [15].

7. RESULTS AND DISCUSSION

In this paper, we have presented a high-quality English text-to-speech system. The system can transfer English text into natural speech based on part-of-speech analysis, prosodic modeling and non-uniform units. These technologies significantly improve the naturalness and quality of the TTS system. The system is also modularized for easily incorporating to many applications with speech output.

The TTS designed is tested with 10 different set of documents, the output generated is compared with standard TTS commercially available. Following results are noted after performing the test.

a. The TTS designed is more precisely determining the emotions in the text scanned and converted into voice format.
b. The TTS designed is capable of shifting the emotions from one state to another with smooth transition, which can be noted while listening to the output generated.
c. The matrix of emotions is generated for both TTS designed and standard commercially available TTS and it is found that the emotion recognized by TTS designed are on the higher side.
d. Experimental results demonstrated that the intended emotions were perceived from the synthesized speech, especially "anger", "surprise", "disgust", 'sorrow", "boredom", "depression", and "joy". Future work includes incorporating voice quality in addition to prosody, compensating the duration of phonemes, and applying the proposed framework to other context factors.[11,12]

8. REFERENCEES

[1] Bender, O., S. Hasan, D. Vilar, R. Zens, and H. Ney. 2005. Comparison of generation strategies for interactive machine translation. In *Proceedings of the 10th Annual Conference of the European Association for Machine Translation (EAMT05)*, pages 33–40, Budapest

[2] Casacuberta, F. and E. Vidal. 2007. Learning finite-state models for machine translation. *Machine Learning*, 66(1):69–91.

[3] Tom´as, J. and F. Casacuberta. 2006. Statistical phrase-based models for interactive computer-assisted translation. In *Proceedings of the 44th Annual Meeting of the Association for Computational Linguistics and 21th International Conference on Computational Linguistics (COLING/ACL 06)*, pages 835–841, Sydney.

[4] I. Titov and R. McDonald. 2008. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. ACL-2008

[5] Allen, J., M.S. Hunnicutt, and D.H. Klatt, *From Text to Speech: the MITalk System*, 2007, Cambridge, UK, University Press.

[6] J. Wiebe, and T. Wilson. 2002. Learning to Disambiguate Potentially Subjective Expressions. CoNLL-2002.

[7] F. Casacuberta et al. Some approaches to statistical and finite-state speech-to-speech translation. *Computer Speech and Language*,18:25–47, 2004.

[8] D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2000

[9] Fangzhong Su and Katja Markert. 2008. From word to sense: a case study of subjectivity recognition. In *Proceedings of the 22nd International Conference on Computational Linguistics*, Manchester

[10] Andrea Esuli and Fabrizio Sebastiani. 2007. PageRanking wordnet synsets: An application to opinion mining.In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 424–431, Prague, Czech Republic, June

[11] Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Conference on Empirical Methods in Natural Language Processing* , pages 129–136, Sapporo,Japan.

[12] B. Pang and L. Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *(ACL-04)*, pages 271–278, Barcelona, ES. Association for Computational Linguistics

[13] Laxmi-India, Gr.Noiida, March 2010. Development of Expert Search Engine for Web Environment. In International Journal for Computer Science and Security, pages 130-135, Vol 4. Issue 1, CSC Journals, Malaysia.

[14] J. Yuan, J. Brenier, and D. Jurafsky, “Pitch accent prediction: Effects of genre and speaker,” in *Proc. Interspeech 2005*, Lisbon, Portugal, 2005.

[15] V. Strom, R. Clark, and S. King, “Expressive prosody for unit-selection speech synthesis,” in *Proc. Interspeech*, Pittsburgh, 2006.