Emotion Extractor: AI based methodology to implement prosody features in Speech Synthesis

M.B.Chandak[1], Dr.R.V.Dharaskar[2]

[1] Nagpur University, Shri Ramdeobaba Kamla Nehru Engg. College
Research Scholar G.H.Raisoni College of Engg. Nagpur
Nagpur, Maharashtra, India

[2] MPGI Campus, Nanded
Maharashtra, India

Abstract

This paper presents the methodology to extract emotion from the text at real time and add the expression to the documents contents during speech synthesis. To understand the existence of emotions self assessment test was carried out on set of documents and preliminary rules were formulated for three basic emotions: Pleasure, Arousal and Dominance. These rules are used in an automated procedure that assigns emotional state values to document contents. These values are then used by speech synthesizer to add emotions to speech. The system is language independent and content free.

Keywords: *Speech Synthesis, Emotions, Artificial Intelligence, Prosody, Expressive speech synthesis.*

1. Introduction

Due to development in the field of Information Technology, electronic documents are acting as major source of transmitting information. The documents are formatted based on information present in the documents so as improve the readability. Recent research on psychology and cognitive science indicates that emotions plays important role in human decisions. Modeling emotions is an interesting subject in artificial intelligence. Various studies revels that there exists relationship between the metadata and readability of document as well as the meaning the reader understands from the documents [3]. Emotion and emotion state of reader depends largely upon the document structure, layout, text formatting and the context of the document. Minski, first pointed out the importance of emotion in artificial intelligence [1] [2]. Ortony et al. presented the famous OCC theory of emotion and model was used in many studies involving emotion extraction from running text. The model suggested that the context is changed from one part of the text to another; there is change in the emotions of the reader. Despite of illusory simplicity, emotional analysis of text presents a great challenge to computer scientists due to variety of expressed meaning in texts.

Studies on Human Computer Interaction field focus on the user's emotional response during the interaction [4]. A common experimental procedure Self Assessment Test was introduced by P.J.Lang in 1985. The test procedures were used for assessment of advertisement, evaluation of readability of documents and web pages [5].

Prosody generation is a complex process that involves the analysis of several linguistic phenomena. Dynamic approaches are usually prone to errors. For instance, part-of speech (POS) identification fails in 5% of the cases for Greek using statistical taggers, while syntax and metric trees are hard to construct. A solution that overcomes that is offered by (a) limiting the domain to which the TtS applies to and thus limiting the linguistic phenomena, and (b) using a Concept-to-Speech (CtS) system. The advantage of the latter is that the generated texts are annotated with high level linguistic factors in contrast to plain texts.

Expressive speech synthesis [E.S.S] is a method of integrating emotions to speech using variation in pitch of speech and speech characteristics. There have been studies on modeling expressive speech using the Pleasure-Arousal-Dominance approach. The advantage of using this method is that the values of P.A.D. dimensions are continuous. Through an emotional state it is possible to map P.A.D. values in a specific emotion (or variations of the emotion). For example, the emotion "Anger" can have variations like "Angry", "very angry", "less angry".

In this work we propose a methodology for the real time extraction of readers' Emotional State (E.St.) from documents' metadata and the P.A.D. annotation. Analysis of metadata which conveys hidden information present in the text in the form of emotion is complex procedure.

Previous works have accomplished this "conversion" using methods like semantic extraction from documents and convey the hidden logic into acoustic modality. However there is no systematic work on creation of automated reader's emotion, which uses text formatting and structure of documents.

Using the E.St. annotated document, we are able to map the states into variations and differences of speech characteristics.

2. System Overview

In this study we propose an XML-based system. The document to be converted into voice form is preprocessed to generate the tagged document. The tags are added to recognize the emotions in the document and are denoted as emotion tags. The emotions tags are derived from comparing the document main contents with the database which contains the information about the words and their classification into specific emotion class. This will generate the effect, that emotions are automatically extracted from document just like the reader's emotional state transitions, in real-time mode and of expressions during speech synthesis. The proposed architecture performs following steps:

a) For the given document, semantic analysis is performed to produces annotated document. This document contains additional information about various types of emotions present and details about document structure.

For example: in the text if the sentence is 'I am happy about your result', then annotated text will add tag to verb 'happy'. This will be considered as emotion in the text and context.

The semantic analysis will also split the text in parts, so that further analysis is simplified.

For example: in the text if the sentence is 'I was so angry, that I shouted loudly on students' will be split into two parts: 'I was very angry' and 'I shouted loudly on student' and will also generate emotion 'angry' for the context.

b) The information about text formatting and other components of documents such as image, figures etc is generated.

For example: if the content is explained with the help of figure, then header of the annotated text will store this information.

c) Maps the emotion values into speech prosodic variables in order to convey the hidden emotional state information of documents into speech.

For example: If the emotion in the context present in the text is 'happy', then prosodic variable for the context is set to 'happy' mood.

3. Part of Speech Tagging for prosody generation

POS Category	POS Features
Adjective	Degree
Noun	Common / proper
Pronoun	Personal/ relative
Participle	Sub category of verb
Article	Definite/ indefinite
Numeral	Ordinal / cardinal
Verb	Voice, mood, person, number
Conjunction	Co-ordinating / sub ordinating
Prepositions	Of, with
Adverb	
Residuals	Abbreviations / Foreign words

4. Emotion Extraction from documents

4.1 Basic System:

In Figure 1 the system's architecture is presented.
The system is divided into two modules:
i. Emotional State Extraction (E.S.E.) module.
ii. Expressive Speech Synthesis (E.S.S.) module

In the ESE module the document is parsed and some basic emotions like pleasure, anger, sorrow, dominance, etc are obtained, which are stored in the form of XML document. This XML document is access by the ESS module for implementing emotion with the speech.

The E.S.S. Module (Figure 1) is used for following purposes:

1. To preprocess the documents and analyze the emotional state present with the document text.
2. This will generate the repository of emotions.
3. The repository is used by the composer and along with acoustic modeling rules; the speech is generated with variable pitches as per the emotions involved in the text.

The E.S.E. Module (Figure 2) is used for generating meta-data about the document. The meta-data consist of following:

a. Text formatting meta-data, to provide information about the font type and size. (i.e. bold, italics, font size),
b. Text Structure meta-data, to provide information about headers and sub-headers of the document (i.e. chapter, title, paragraph),
c. Text Layout meta-data, to provide information about number of columns in the text, borders etc
d. Non-textual meta-data (i.e. Figures, Drawing, Pictures, Logos).

These meta-data are host in the tagged Document which is processed in a later stage, by the E.S.E. Module for the Emotional State Annotation.

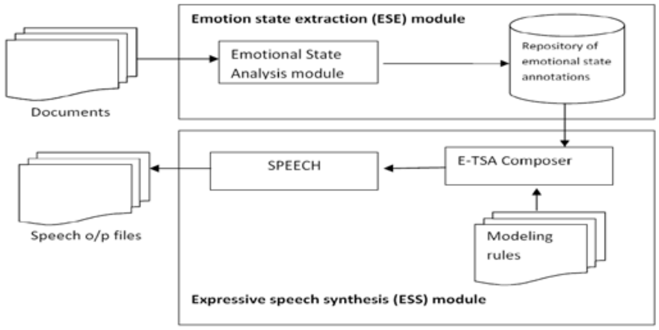

Emotion state extraction (ESE) module
Documents
Emotional State Analysis module
Repository of emotional state annotations
SPEECH
E-TSA Composer
Speech o/p files
Modeling rules
Expressive speech synthesis (ESS) module

Figure 1: The proposed XML architecture

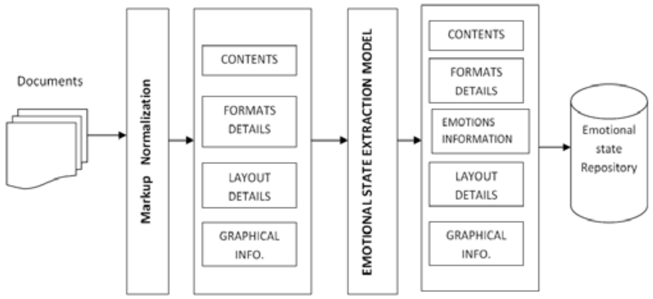


Figure 2: Emotion State Extraction Module

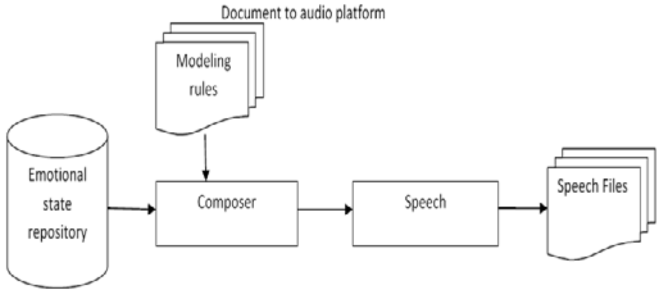


Figure 3: Expressive Speech Synthesis module

4.2 Emotion State Annotation

To understand the emotional state of user, we investigated most commonly used elements to find emotions along with the above four factors. These elements are:

1. Text formatting meta data
2. Text structure meta data
3. Text content meta data

The completion of the experimental procedure result the mapping rules for the P.A.D. annotation of documents. The rules are used in the E.S.E. module which produces an XML file (Emotional-ML) that is stored in a database called Emotional State Annotation Repository.

5. Expressive Speech Synthesis and Metadata

To implement expressive speech synthesis, the metadata generated after processing the speech is used to construct the new format of the document called as Emotional version of the document. This format is stored in emotional state annotation repository. The forward parsing approach is used to generate the real time emotions. In this method when one part of the document is used by speech synthesizer, the forward part or the subsequent parts are scanned and emotions present will be stored in repository. This is continuous process and result in generating real time emotions.

Similarly to generate actual effects, to each generated emotion, a degree of activation is added. This will help in converting one emotion to its higher or lower state.

The advantage of this model, over the generally described methodology, is the immediate correlation of emotions like Pleasure – Arousal – Dominance (named also as Evaluation – Activation – Power respectively) with the prosody variations. This correlation gives us the opportunity to map the documents' metadata into speech elements using the standard dimensions as medium. In Figure 3 the E.S.S. Module is described. The mapping is performed using Document-to-Audio platform. This platform makes use of the prosodic rules generated from the documents, which are to be converted into voice format. The voice dictionary is used for mapping the text content to audio form. The dictionary only contains the basic phonemes of source language, which are further with the help of grammar of the language and concatenation process generates the actual words. These words are applied with the prosodic rules to generate the desired voice output.

6. Implementation Details of System

6.1 Method to calculate emotion

The system makes use of following to calculate emotion and its intensity value:

a. information from dictionary
b. matching patterns
c. empirical matching patterns from own studies

Information from dictionary

Various dictionaries in the form of 'Wordnet', 'General Inquirer' 'Levin verbs' are used to generate group of emotion words.

Matching Patterns

The communication grammar of English is used to generate group of matching pattern. Some of the examples are:

1. Interjection: Ex: Oh, What a beautiful day!
2. Exclamations: Ex: What a master piece!
3. Repetition: Ex: The machine is very very expensive
4. Intensifying adverbs and modifiers: Ex: We are completely speechless.
5. Exclamatory question: Ex: Hasn't they came till now!
6. Emphasis: Ex: How did you manage all these?
7. Intensifying a n egative sentence: We were not informed about this at all.

Empirical matching patterns from own studies

The emotions on the basis of different examples were classified into five classes: neutral, positive, negative, low positive, low negative. The intensifier is calculated from the available text and is applied to emotion word. For example

Example	Pattern
I am so happy	<Intensifier> <Emotional word+> → <Result++>
I am not happy	<Negation> <Emotional word+> → <Result->
I am not very happy	<Negation> <Intensifier> <Emotional word+> → <Result->

In the above table:
<Emotional word+> is a low positive emotion word
<Result++> is high positive result
<Result-> is low negative result

6.2 Method to link part of text

If a sentence is consist of two subparts then emotions present in the subparts are calculated. The emotion of one part is imposed on other part. This will allow merging the emotion of two or more subpart and generating common emotion factor for the entire sentence.

Pattern for linking sub-sentences	Example
<Sup++><Sup+> → <Result++>	It is very good play and acting is excellent
<Sup++> <Sub-> → <Result+>	It is very good play, but acting in the first part is not good.

In above table, as shown the pattern for linking sub-sentence is a pattern that matches the example. <Sub++> represents high positive emotional meaning of super dominant sub sentence. <Sub+> represents low positive meaning of super dominant sub sentence. <Sub-> is low negative meaning of super dominant sub sentence. <Result++> is high positive result of sense of sentence, <Result+> low positive result of sense of sentence.

7.0 Complete System

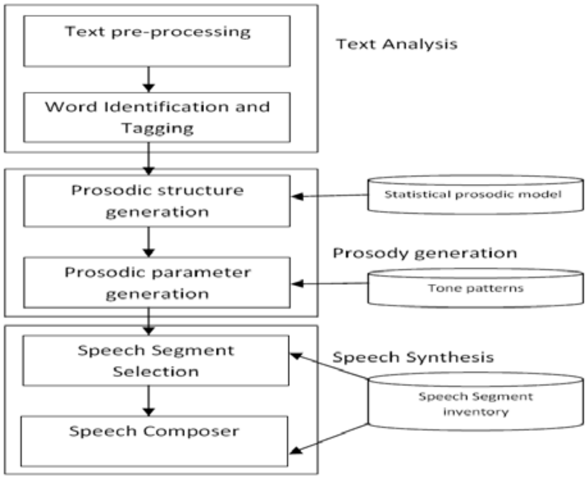
As shown in the figure 4, the system is consists of three main modules: text analysis module, prosody generation module and the speech synthesis module.

The text analysis module transforms an input sentence into word sequences with POS tagging. This module is capable of handing the abbreviations and all special characters of the language.

The prosody generation module will generate the correct prosodic information of the document and store it in the repository. This module is also capable of selecting the male and female voices for generating output.

The speech module is used to select the speech waves from the dictionary and add the necessary emotion factor to the speech waves before actually being converted to voice format for output.

The generated voice will have the prosodic feature which will make the speech output more realistic.



8. Effectiveness Experiments

A total of 50 students and staff participated in experiment. The self assessment test was carried out on a standard document in various phases. The test was designed using guidelines presented by P.J.Lang. [11]. The document used was consist of various parts like title, chapter, subtitles, body text etc. The participants were provided with a t abular format to mark the presence of emotion and intensity of emotion in the text.

In the first phase, the participants were asked to assess the document on various metadata related with the document.

In second phase the same document was used and emotion assessment was carried with the system developed. The system generated emotions their intensity values were stored in database. The database values and values generated by participants were then compared to revel the conclusions.

8.1 Experimental Results & Comparison

The participants were asked to access the contents of text for various types of emotions such as anger, boredom, disgust, anxiety, happiness, sadness and neutral. These emotions were also tested with the system design. Following table shows the details of the emotions in percentage generated from participants and comparison of the results with the results generated by the system.

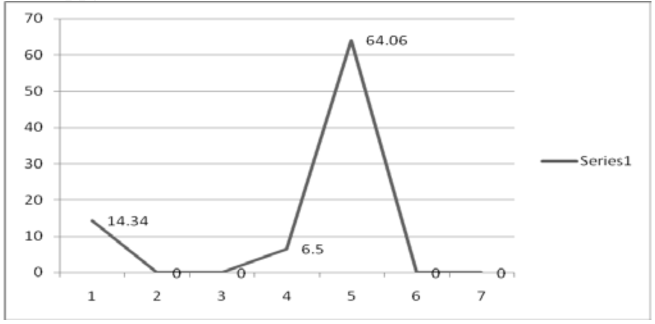
P/M	A	B	C	D	E	F	G
A	82.12	0.0	0.0	4.12	14.34	0.0	0.0
B	0.0	90.0	4.23	0.0	0.0	0.0	5.6
C	0.0	3.23	85.34	6.5	0.0	4.2	0.0
D	8.04	2.04	0.0	74.78	6.50	0.0	3.68
E	27.56	0.0	1.23	6.54	64.06	0.0	0.0
F	0.0	0.0	0.0	0.0	0.0	95.66	4.58
G	1.2	6.6	0.0	1.2	0.0	1.2	90.34

A=Anger; B=Boredom; C=Disgust; D=Anxiety

E=Happiness; F=Sadness; G=Neutral, P=Participants, M=Machine

The value shown in the above table is % match between the values generated by the participants and values generated by the system.

The graphical representation is as shown below for E=Happy emotion:



9. CONCLUSION

We have presented an approach for the automatic emotional state annotation of documents using the various emotion dimensions. The proposed methodology maps documents' meta-data into speech prosodic variations. The output of the Emotional State Extraction Module can be utilized in different Expressive Speech Synthesis systems. Limiting the emotional state variations and consequently the prosodic variations, unit selection technique would be more appropriate, increasing the naturalness of the speech output.

REFERENCES

1. M. Minsky, *The Society of Mind*, New York Simon & Schuster, 1985.
2. M. Minsky, *The Emotion Machine*, New York: Simon and Schuster, 2006.
3. A. Ortony, G.L. Clore, A. Collins, *The cognitive structure of emotions*, New York: Cambridge University Press, 1988.
4. C.-H. Wu, C.-C. Hsia, T.-H. Liu and J.-F. Wang, “Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis”, *IEEE Transactions on A udio, Speech and Language Processing*, vol. 14, No 4, 2006, pp. 1109-1116.
5. S. Kettebekov, M. Yeasin, and R. Sharma, “Prosody based co-analysis for continuous recognition of co-verbal gestures,”*Proc. ICMI’02*, 2002.
6. M Slaney, G McRoberts, Baby Ears: A Recognition System for Affective Vocalizations, Proceedings of ICASSP, May 12-15, 1998, Seattle, WA.
7. K. Holmqvist, J. Holsanova, M. Barthelson and D. Lundqvist, “Reading or scanning? A study of newspaper and net paper reading”, In Hyönä, J. R., and Deubel, H. (Eds.)*, The mind's eye: cognitive and appl ied aspects of eye movement research,* Elsevier Science Ltd, 2003, pp. 657-670.
8. R. J. McAulay and T.F. Quatieri, “Shape Invariant Time-scale and Pitch Modification of Speech”, IEEE Trans. On Signal Processing, 1992, pp. 497-5 10
9. “Special Section on Expressive Speech Synthesis”, as presented in the Editorial of *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no 4, July 2006, pp.1097-1098
10. Holmberg N., *Eye movement patterns and newspaper design factors. An experimental approach*, Master Thesis, Lund University Cognitive Science, Sweden, 2004.
11. Lang P.J., M.Bradley, *International affective picture system (IAPS): Instruction Manual and Affective Ratings,* Technical Report A-6, The Centre for Research in Psychophysiology, University of Florida, U.S.A., 2005
12. P. Boersma, “Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound,” *Proc. of the Inst. of Phonetic Sciences*, vol. 17, pp. 97–110, 1993
13. Barrien F., *Color and Human response*, Van Norstrand Reinhold, New York , 1978.
14. L.A.Streeter, N.H.Macdonald, R.M.Krauss, W.Apple, K.M.Galotti, “Acoustic and perceptual indicators of emotional stress”, .l.Acoust.Soc.Am., ~01.73, No.4, April 1983, pp.1354-1360.