



Affective Story Teller: A TTS System for Emotional Expressivity

Mostafa Al Masum Shaikh, Antonio Rui Ferreira Rebordão, Keikichi Hirose

Department of Information and Communication Engineering, University of Tokyo, Japan

{almasum, antonio, hirose}@gavo.t.u-tokyo.ac.jp

Abstract

Some Text-to-Speech (TTS) systems revealed weaknesses in their emotional expressivity but this situation can be improved by a better parameterization of the acoustic and prosodic parameters. This paper describes a system, Affective Story Teller (AST), as an example of emotionally expressive speech synthesizer. Our technique uses several linguistic resources that recognizes emotions in the input text according to its emotional affinity and assigns appropriate prosodic parameters as well as pitch accents by XML-based tagging to generate a synthesized speech sample. Then the synthesized sample is re-synthesized through TD-PSOLA based pitch manipulation in accordance to emotional connotation. The system employed MARY TTS system to readout a folk tale. The preliminary perceptual test results are encouraging and human judges, by listening to the re-synthesized speech samples of AST, could perceive "happy", "sad", and "fear" emotions much better than compared to when they listened non-affective synthesized speech.

Index Terms: speech synthesis, emotional expressivity, affective story teller, MaryXML, intelligent text processing

1. Motivation

Expressive eloquence contributes to the naturalness of synthesized speech as indicated by many studies like [1,2,3,4]. It is generally accepted that an unified tone, a proper pitch accent and a suitable intensity of speech can help conveying speech subtleties and their intent in a contextually and content-rich manner. The Figure 1 shows the relative changes of four quantitative speech variables namely, Speech Rate (SR) (i.e., syllable/sec), Pitch Average (PA), Pitch Range (PR), and Intensity (I) with respect to neutral speech. This objective evaluation matches with the findings of the studies [1,4] that interrelates the aforementioned variables with emotional expressivity in human speech. Therefore, a good TTS also should match this behavior. Therefore, if a Text-To-Speech (TTS) system can generate human-like speech, then it can convince or appeal to a particular audience more successfully. Thus, in our opinion a TTS system should produce synthesized speech that resembles speech produced by human articulation but contemporary TTS systems tend to produce synthetic speech in a way that sounds unnatural. This is partly due to some deficiencies in the syntactic analysis of the raw input text and to a lack of semantic information, affective clues, and world knowledge. Several perceptual and objective experiments that have been carried out in [5], show that the present TTS systems are weak in the characterization and expression of emotions. In [5] the authors provided affective and non-affective text to several state-of-the-art TTS systems and analyzed the synthesized speech samples. This study revealed that the pitch accent assignments in the synthesized speech were inappropriate and that their pitches were very similar to the synthesized speech samples produced out of non-affective sentences. The affective texts had obvious affective connotation (e.g., sad/happy) but this

emotional content were not present. Therefore, it is inferred that TTS systems usually fail in encoding emotional connotation in synthesized speech.

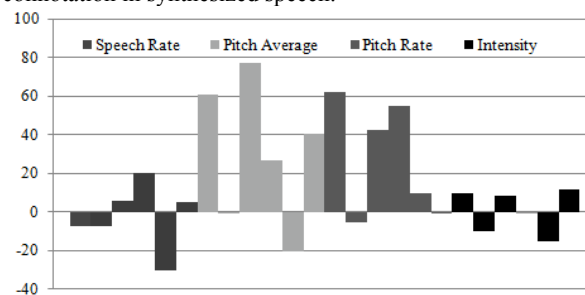


Figure 1: Changes (in percentage) of SR, PA, PR, I for 6 emotions (from left to right and per color: happy, sad, anger, fear, disgust and surprise) with respect to neutral speech for speech articulated by humans.

Some TTS systems accept XML-like mark-up input text pre-marked with intonational information but we have noticed that very few systems make intelligent text pre-processing that can assist the synthesis process. We use commonsense knowledge and emotion recognition techniques to process the text, annotate appropriate pitch accent to words and/or phrases and adjust the prosodic parameters before the synthesis. For example, the studies [1,2,4,6] show that in order to signal "sadness", the Speech Rate (SR) (i.e., syllable/sec) and Pitch Average (PA) should be slightly slower; the Pitch Range (PR) should be slightly narrower; the Intensity of the signal should be lower; and the Pitch Change (PC) should have downward inflections with respect to neutral speech. To signal "happiness" the SR should be faster or slower; the PA should be much higher; the PR should be much wider; the Intensity should be higher; and the PC should have smooth upward inflections. Therefore our approach tries to configure dynamically those parameters by sensing the affective meaning of the input text. Although automatic prosody control in a TTS is not new, most of the previous work gave emphasis at an acoustic level and not in a text-processing level. So our primary contribution lies in this text-processing zone. Extensive linguistic processing is done at this level and appropriate speech parameters are assigned that can assist the synthesizer in generating emotion-embedded speech.

2. Related Work

Although tremendous effort has gone into the synthesis of speech from text as well as identifying emotions from human speeches, as far as we know there is no system that takes the content (e.g., typed text), evaluates its affective information and parameterizes appropriate prosodic settings that feed a TTS engine. By carefully reviewing the existing literature it is found that research regarding expressivity in synthetic speech is closely related to the following concepts: emotional text-to-

speech synthesis; control languages to guide TTS synthesis process; flexibility in TTS architecture; and emotion recognition from textual data. The following sections briefly discuss these concepts.

2.1. Emotional Speech Synthesis

Previous researches (e.g., [1,2,3,4,6]) have found that there are several features in natural speech that are adhering to specific affective connotation. These features are: different statistical values (e.g., max, mean, standard deviation etc.) of fundamental frequency F0; different statistical values of the first three formants (F1, F2, and F3); and their bandwidths (BW1, BW2, and BW3), energy, speaking rate, etc. These features are generally derived by observing how human's voice changes regarding different emotions. The studies mentioned above have established that when a speaker is in a state of fear, anger or joy, then his speech is typically faster, louder, and enunciated, with strong high-frequency energy. When the speaker is bored or sad, then his speech is typically slower and low-pitched, with very little high-frequency energy. Such pragmatic knowledge obtained from speech signal processing has inspired various kinds of synthesis methods like, formant synthesis, diphone concatenation, unit selection and prosody rules based synthesis. In [2,3] these techniques are described along with their advantages and disadvantages. Moreover, techniques like explicit prosody control, expressivity based unit selection, HMM based parametric synthesis, non-verbal vocalization, etc. are applied and obtained partial success in recognizing anger and sadness from synthesized speech samples.

2.2. XML-like Markup Languages for TTS Systems

XML-based markup languages aim at giving a non-expert user the opportunity of adding information to a text in order to improve the way it should be spoken. The languages are independent of any particular TTS system and the synthesis processors are assumed to parse the markup enriching their input and translate the information contained in it into a system-internal data representation format which in most cases is not XML-based. VoiceXML [7] is a kind of markup language used for specifying interactive voice dialogues between a human and a computer. Languages like SSML [8], JSML [9], SABLE [10] and MaryXML [11] are used to control synthesis. For example, MaryXML can be used to control the prosody, the accent and the boundary of the articulation. For example for the phrase "Please look at me!" we could have: `<prosody rate="+30%" pitch="+50%" range="-5%" volume="loud"><t accent="L+H*">please</t><t accent="L-L%">look at me!</t><boundary duration="100"/></prosody>`

2.3. Sensing Affective Information from Text

This research addresses the aspect of subjective opinion, which particularly includes the identification of different emotive dimensions and classifying texts by their emotion affinity. It can be argued that attitude's analysis and affect in texts depends on audience, context and world knowledge. The approaches for assessing affective information from text are based in using one or a combination of the following techniques: keyword spotting; lexical affinity; statistical methods; a dictionary of affective concepts and lexicon; common-sense knowledge-base; fuzzy logic; knowledge-base from facial expression; machine learning; domain specific classification and contextual valence assignment. Some researches [12,13] dealt with the above techniques. For

example, Shaikh et al. [18] implemented contextual valence assignment technique and achieved tremendous result in recognizing different emotions (e.g. happiness, sadness, anger, etc.) from text and Liu et al. [19] using common-sense knowledge could detect the basic six emotions from a given input text.

2.4. MARY TTS: A Flexible TTS System

The MARY TTS system [20] is a client-server application written in Java and created at DFKI. MaryXML serves as the configuration input language of the system and thus it has become a very flexible toolkit for speech synthesis research.

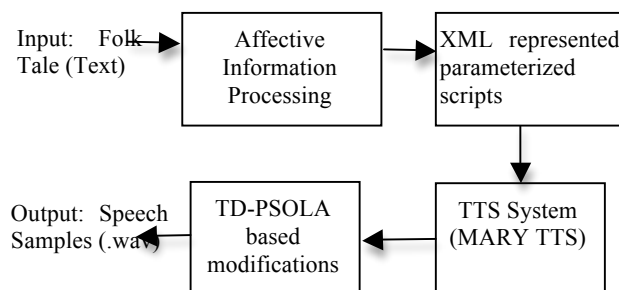
We have chosen MARY system because we can dynamically create MaryXML with appropriate prosodic and accent properties related to emotion and still be able to access all intermediate processing results for purposes of debugging and analysis.

3. Our Approach

Our system deals with five basic emotions: happy, sad, fear, anger, and surprise. It performs affective evaluation of the input text and, accordingly to the emotional content of the input sentence, it produces MaryXML that matches the desired prosodic parameters and the findings reported in [1,2,4]. This Dynamic MaryXML is used as input to MARY system. Then for the obtained synthesized speech samples F0 mean, range, and shape modifications were performed using the Time Domain Pitch Synchronous Overlap and Addition (TD-PSOLA) algorithm as implemented in the Praat software [14].

3.1. System Architecture

Figure 2 indicates the pipelined architecture of the Affective Story Teller (AST).



3.1.1. Affective Information Processing

For each input sentence the language processing module outputs triplet(s) consisting of a subject or agent, a verb and an object. Each member of the triplet may or may not have associated attribute(s) (e.g. adjective, adverb etc.). First XML-formatted syntactic and functional dependency information of each word of the input text is obtained using the Machine Syntax parser [15] and this output constitutes the basis for further processing that generates the triplet(s). Since a triplet is initiated with an occurrence of a verb in the sentence, the semantic parsing may obtain more than one such triplet if there are multiple verbs in the sentence. Basically a triplet encodes information about "who is associated with what and how" with a notion of semantic verb frame analysis. For example, a sample sentence of the story "I like sleeping in my hut with a nice fire to keep me warm." produces three triplets as shown in Table 1.

Table 1: Triplet output of parsing for the example sentence.

Triplets processed by Semantic Parser	
Triplet 1	[[['Actor:','I','Actor-Type:','self','Actor-Attrib:',''], ['Action-Name:','like','Action-Status:','Present','Action-Attrib:',['dependency: to']], ['Object-Name:',' ','Object-Type:',' ','Object-Attrib:',['']]]
Triplet 2	[[['Actor:','I','Actor-Type:','self','Actor-Attrib:',''], ['Action-Name:','sleep with','Action-Status:','Present','Action-Attrib:',['place: in my hut','dependency: and']], ['Object-Name:','fire','Object-Type:','N NOM','Object-Attrib:',['nice: A ABS']]]
Triplet 3	[[['Actor:','fire','Actor-Type:','N NOM','Actor-Attrib:',''], ['Action-Name:','keep','Action-Status:','Present','Action-Attrib:',['warm: A ABS']], ['Object-Name:','me','Object-Type:','PRON PERS ACC SG1','Object-Attrib:',['']]]

We used the output of the system SenseNet developed by Shaikh et al. [18] that can process the triplet-formatted input of a sentence. SenseNet can perform sentence level affective sensing by assessing the contextual valence of the words using rules and prior-valence values of the words. It outputs a numerical value ranging from -15 to +15 flagged as the ‘sentence-valence’ for each input sentence. As example, SenseNet outputs -10.76 for the given input sentence. The output value indicates a numerical measure of negative or positive sentiments carried by the sentence. SenseNet implements a cognitive theory of emotion known as the OCC emotion model [16] by developing rules for the model defined emotions. Therefore it can classify input texts according to eight emotion-types namely, happy, sad, hope, fear, admiration, shame, love, and hate plus a neutral category. In this system we map the output of SenseNet to the basic six emotions in the following manner: happy, hope and love are considered as happiness, sad as sadness, fear as fear, admiration as surprise, shame as anger and hate as disgust. Following an experimental study, the accuracy of SenseNet to assess sentence-level negative/positive sentiment is 91% and classification accuracy of eight emotions is 82%.

3.1.2. XML represented parameterized scripts

After the input text has been processed as mentioned above, we obtain affective assessment of the text like: the overall emotion carried by the text; positive or negative meaning of the events represented by the triplet(s); the attributes (e.g., location, time, etc.) of the events that appear as important information. First, several speech parameters are set for the overall synthesis adhering to the overall negative or positive affective connotation of the text and then parameters like pitch, pitch-dynamics, number-of-pauses, etc. are adjusted with respect to neutral emotion expressing speech according to the detected named emotions. For example, if the text would have to express “happiness” then the overall speech rate is set faster, pitch average is set higher, pitch range is set much wider, intensity is made higher, and pitch changes are set as smooth upward. MaxyXML offers a rich set of such prosody attributes to realize emotion specific desired setup.

At present MARY TTS system has the following natural language components: Tokenize; Preprocessing; and Tagger & Chunker. These components can process an input given in MaryXML format and our system, at present, has nothing to do with these components. It just creates MaryXML formatted input from a given plain text and inputs it into

MARY TTS. In future we plan to add a pre-processing module to MARY system that implements our approach by performing emotion recognition from the plain text and automatically generating MaryXML in accordance to the recognized emotion. The following is an example of the dynamic MaryXML of a given example sentence that expresses “fear”.

```
<?xml version="1.0" encoding="UTF-8"?>
<maryxml
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns="http://mary.dfki.de/2002/MaryXML" version="0.4"
xml:lang="en">
<prosody pitch="-5%" pitch-dynamics="-25%"
range="5.32st" range-dynamics="+26%" preferred-accent-
shape="falling" accent-slope="+75%" accent-
prominence="+58%" preferred-boundary-type="low" rate="-
0%" number-of-pauses="+23%" pause-duration="-7%"
vowel-duration="-5%" nasal-duration="-5%" liquid-
duration="-5%" plosive-duration="+41%" fricative-
duration="+41%" volume="61">
The car exploded near a popular ice cream parlor, sending
flames and shrapnel through the busy square and killing 17
people.
</prosody>
</maryxml>
```

3.1.3. TD-PSOLA based modifications

The modification parameters are taken according to the explanation given in [17,18] and the following modifications are applied on the obtained synthesized speech samples each adhering to a specific emotion. First, the mean was modified by shifting the F0 contour up or down. The following modifications were applied: a) Increasing F0 mean by 15%, 25% and 50% on the synthesized speech samples predefined as ‘Happy’, ‘Anger’ and ‘Surprise’ respectively; b) Decreasing F0 mean by 25% and 50% for ‘Sad’ and ‘Fear’ emotion while keeping the neutral speech samples unchanged. c) These modifications made the F0 mean equal to 150, 250, 300, 250, and 300 Hz for ‘Happy’, ‘Anger’, ‘Surprise’, ‘Sad’, and ‘Fear’ emotion respectively. Multiplying the F0 contour with a constant and then shifting the contour up or down so that the mean will be the same as the original mean value the range was modified. For ‘Surprise’, ‘Anger’ and ‘Happy’ the scaling range is 2.0; for ‘Sad’ and ‘Fear’ the scaling range is 0.75, and 1.5 respectively. Finally, by stylizing the F0 contour the shape of the F0 contour of the utterances was altered. The following modifications were applied: Stylizing the F0 contour by a 2, 5, 10, 15, 20, and 40 semitone frequency resolution for ‘Surprise’, ‘Happy’, ‘Anger’, ‘Sad’, ‘Fear’ and ‘Neutral’ emotion. Stylization of the F0 contour was performed using the Praat software. The logic behind the stylization algorithm is explained in [18].

4. Experiments and Results

For each affective text, using MARY TTS, we created two versions of synthesized speech samples. One is the output obtained from just plain text input (i.e., S1) and the other is produced by our approach (i.e., S2). Both cases used the male and female voices of Mbrola-us2 version 3.5.0 and a total of 12 people participated (all of them were non-English speaking natives). The tested folk story have 12 happy, 12 sad, 12 anger, 12 fear, 6 surprise, and 9 neutral emotion bearing sentences. Some speech samples and the text of the folk story can be found at www.almasum.com/ast/. Real human beings also articulated the story. The subjects had to listen to the human-spoken version (R) and synthesized audio samples of

Table 2: Confusion Matrix for emotion perception of human-spoken speech, synthetic speech samples of S1 and S2

Predicted	Actual																	
	H			S			A			F			Su			N		
	R	S1	S2	R	S1	S2	R	S1	S2	R	S1	S2	R	S1	S2	R	S1	S2
H	107	65	87	5	5	16	7	7	12	3	3	11	14	15	17	8	12	14
S	6	28	15	109	116	98	29	43	38	11	38	23	2	2	4	3	25	5
A	3	8	5	10	10	9	82	42	69	6	18	14	4	11	3	7	13	10
F	4	6	8	13	6	11	13	26	10	115	66	87	4	8	2	9	11	8
Su	10	23	13	3	3	4	7	10	8	5	9	6	45	32	39	6	9	7
N	14	14	16	4	4	6	6	16	7	4	10	3	3	4	7	75	38	64

Table 3: Measure of Precision (P), Recall (R), F-Measure (F) and Accuracy of the three instances

	H	S	A	F	Su	N
R	P=74.31 R=74.31 F=74.31	P=68.13 R=75.69 F=71.71	P=73.21 R=56.94 F=64.06	P=72.78 R=79.86 F=76.16	P=59.21 R=62.5 F=60.81	P=70.75 R=69.44 F=70.09
Avg. P=69.73; Avg. R=69.79; Avg. F=69.52 and Accuracy=70.50						
S1	P=60.75 R=45.14 F=51.79	P=98.31 R=80.56 F=88.55	P=41.18 R=29.17 F=34.15	P=53.66 R=45.83 F=49.43	P=37.21 R=44.44 F=40.51	P=44.19 R=35.19 F=39.18
Avg. P=55.88; Avg. R=46.72; Avg. F=50.60 and Accuracy=47.49						
S2	P=55.41 R=60.42 F=57.81	P=96.08 R=68.06 F=79.67	P=62.72 R=47.92 F=54.33	P=69.05 R=60.42 F=64.44	P=50.65 R=54.17 F=52.35	P=62.14 R=59.26 F=60.66
Avg. P=66.01; Avg. R=58.37; Avg. F=61.54 and Accuracy=58.73						

the story produced by S1, and S2. They were asked to label each utterance using any of the following labels: happy (H), sad (S), anger (A), fear (F), surprise (Su), and neutral (N). In this manner every judge evaluated 63 samples producing 756 evaluations (i.e., 63*12) for each instance. The confusion matrix of individual emotion perception from the three sources of speech samples is given in Table 2. Table 3 summarizes the precision, recall, and f-measure values of the emotion perception for those three sources of speech samples. The result shows that S2 has performed better than S1.

The results are encouraging in two manners, firstly S1 is very weak to convey positive emotions (e.g., “happy”, “surprise”), so our approach (i.e., S2) can solve this problem and secondly, S1 has a tendency to add negative emotion (e.g., “sadness”). Thus we are optimistic that our approach can be applied to incorporate different levels of negativism/positivism within different phrases of a sentence and that this can help to a better emotion perceivedness of the synthesized speech.

5. Conclusion

In our study we have found that a well-known TTS system (e.g. MARY TTS) does not produce affective synthesized speech. However, this situation can be improved by pre-processing the input in two manners, first by recognizing the emotion conveyed through the plain text and then controlling the synthesis process by assigning appropriate prosodic parameters that suit the detected emotion. In the second step, re-synthesis is performed on the synthesized samples applying emotion specific TD-POSLA modifications. A perceptual test was performed using the synthesized speech samples and the results support that the speech samples produced by our approach are more affectively expressive than the speech samples synthesized from the plain text. As future work we plan to build a tool combining all the resources discussed in our approach and add it as an add-on to MARY TTS system.

6. References

[1] Cahn, J.E., The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, 8, 1–19, 1990.

- [2] Schröder, M., *Expressive Speech Synthesis: Past, Present, and Possible Futures*, Affective Information Processing (Tao, J., Tan, T., eds.), pp. 111-126, 2009.
- [3] Schröder, M., Approaches to emotional expressivity in synthetic speech. In: K. Izdebski (Ed.) *The emotion in the human voice*, vol 3, Plural, San Diego, 2008.
- [4] Murray, I. R., & Arnott, J. L., Towards the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion. *Journal of the Acoustic Society of America*, 93(2), pp. 1097–1108, 1993.
- [5] Shaikh, M. A. M., Molla, M. K. I. and Hirose, K., Assigning suitable phrasal tones and pitch accents by sensing affective information from text to synthesize human-like speech. In *Proceedings of InterSpeech*, pages 326–329, Brisbane, 2008.
- [6] M. Schröder and J. Trouvain, “The German text-to-speech synthesis system MARY: A tool for research, development and teaching.” *Intl. J. Speech Technol.*, vol. 6, pp. 365–377, 2003.
- [7] <http://www.w3.org/TR/voicexml20/>
- [8] <http://www.w3.org/TR/speech-synthesis/>
- [9] <http://www.w3.org/TR/jsml/>
- [10] <http://www.bell-labs.com/project/tts/sable.html>
- [11] <http://mary.dfki.de/documentation/maryxml>
- [12] Shaikh, M. A. M., Prendinger, H., and Ishizuka, M., Sentiment assessment of text by analyzing linguistic features and contextual valence assignment. *Applied Artificial Intelligence*, Vol.22, Issue 6, pp.558-601, Taylor & Francis, 2008.
- [13] Liu, H., Lieberman, H., and Selker, T. 2003. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th international Conference on intelligent User interfaces* (Miami, Florida, USA, January 12-15, 2003). IUI '03. ACM, New York, NY, 125-132.
- [14] Boersma, P., and Weenink, D. 2007. “Praat: doing phonetics by computer”, version 5.1.19 computer program, <http://www.fon.hum.uva.nl/praat/>, last retrieved Feb 10, 2010.
- [15] Machine Syntax <http://www.connexor.eu/technology/machine/>
- [16] A. Ortony, G. L. Clore, and A. Collins, *The Cognitive Structure of Emotions*. Cambridge University Press, July 1988.
- [17] Bulut, M., Lee, S., and Narayanan, S., Recognition for synthesis: Automatic parameter selection for resynthesis of emotional speech from neutral speech. In *Proceedings of ICASSP*, Las Vegas, Nevada, April 2008
- [18] Bulut, M., Narayanan, S., On the robustness of overall F0-only modifications to the perception of emotions in speech. *The Journal of the Acoustical Society of America*, vol. 123, issue 6, p. 4547-4558