Improving TTS Synthesis for Emotional Expressivity by a Prosodic Parameterization of Affect based on Linguistic Analysis

Mostafa Al Masum Shaikh, Antonio Rui Ferreira Rebordao and Keikichi Hirose

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

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

Abstract

Affective Speech Synthesis is quite important for various applications like storytelling, speech based user interfaces, computer games, etc. However, some studies revealed that Text-To-Speech (TTS) systems have tendency for not conveying a suitable emotional expressivity in their outputs. Due to the recent convergence of several analytical studies pertaining to affect and human speech, this problem can now be tackled by a new angle that has at its core an appropriate prosodic parameterization based on an intelligent detection of the affective clues of the input text. This, allied with recent findings on affective speech analysis, allows a suitable assignment of pitch accents, other prosodic parameters and signal properties that adhere to F0 and match the optimal parameterization for the emotion detected in the input text. Such approach allows the input text to be enriched with meta-information that assists efficiently the TTS system. Furthermore, the output of the TTS system is also post-processed in order to enhance its affective content. Several preliminary tests confirm the validity of our approach and encourage us to continue its exploration.

Index Terms: speech synthesis, intelligent text processing, affect sensing, prosody

1. Motivation

Prosodic variables like tone, pitch accent and intensity are vital for fully understanding speech [1, 2, 3, 4, 5] and, it is generally accepted that, if synthetic speech includes emotion-embedded clues, it will sound more natural and human-like [3,6,7,8] and thus, increases its usefulness as an interface for human-machine systems. Therefore, ideally the output of a Speech-To-Text (TTS) system should resemble human speech and prosody may be the key to achieve it. However, TTS systems fail at incorporating emotional expressivity into their outputs [9]. In [9] the authors provided affective and non-affective text to several state-of-the-art TTS systems and analyzed the synthesized speech samples. They concluded that synthetic speech samples had inappropriate pitch accents that were very similar to the pitch accents of synthesized speech samples produced out of non-affective sentences. The texts had obvious affective connotation (e.g., sad/happy) but this features were not present in the synthesized speech samples. Our experiments suggest the same conclusion. For example, figure 1 presents the relative changes of four quantitative speech variables of human speech: Speech Rate (SR); Pitch Average (PA); Pitch Range (PR) and Intensity (I) with respect to neutral human speech. This behavior matches the findings of [1,6] but, as it can be seen in figure 3, the prosodic variables of the synthesized speech samples produced by several TTS systems do not match the expectations suggested by figure 1, partly due to deficiencies in the syntactic analysis of the raw input text and lack of semantic information, affective clues and context knowledge.

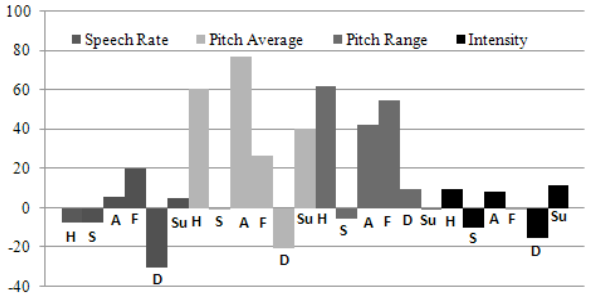

Figure 1: *Changes (in percentage) of SR, PA, PR, I for 6 emotions (from left to right and per color: happy (H), sad (S), anger (A), fear (F), disgust (D) and surprise (Su)) for human speech (with respect to neutral speech).*

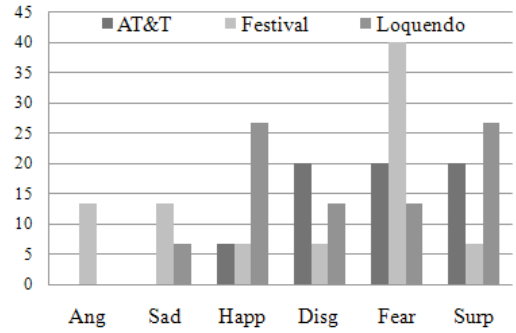

Figure 2: *Results (in percentage) of the perceptual test regarding the efficiency of the emotion recognition (six emotions) of the speech samples synthesized by AT&T, Festival and Loquendo.*

The results of some perceptual tests that we carried also confirm this lack of emotional expressivity. For example, figure 2 shows that, for three well-known TTS system, our subjects could not perceive emotions in the speech samples that they listened. Therefore, it is inferred that TTS systems are not efficient in generating synthesized speech with emotional expressivity.

Some TTS systems accept XML-like mark-up input text enriched with meta-data but, as far as we know, few systems make intelligent text pre-processing that may assist the synthesis process. Our research has its core at this point and proposes a novel way of solving the problem referred in the previous paragraph. For that we use emotion recognition techniques that process the text, annotate appropriate pitch accent to words and/or phrases and adjust automatically suitable prosodic parameters before the synthesis' process.

The structure of this article is as follows: Section 2 discusses related studies and background concepts; Section 3 presents our proposed method; Section 4 explains the datasets and the experimental results; Section 5 concludes the paper and provides some insights into our future work.

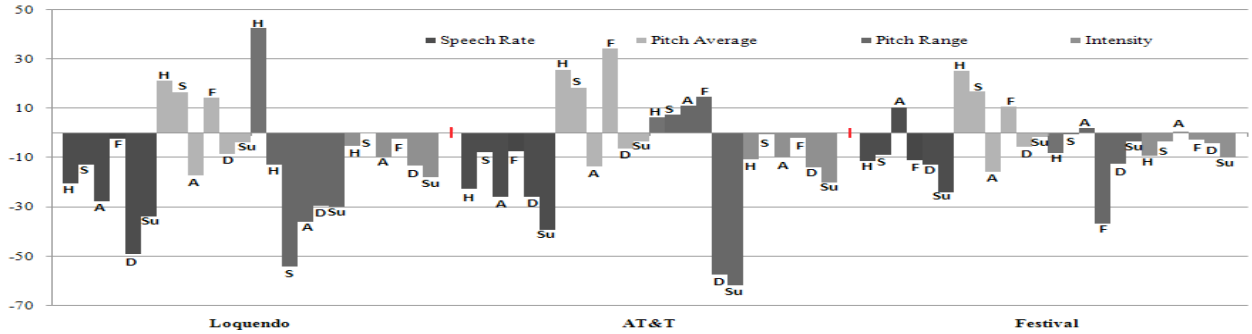

Figure 3: *Changes (in percentage) of SR, PA, PR and I with respect to neutral speech for the TTS systems Loquendo, AT&T and Festival. Each bar corresponds to an emotion (from the left to the right and per color: happy (H), sad (S), anger (A), fear (F), disgust (D) and surprise (Su).*

2. Background

Before presenting our proposed method, we would like to briefly introduce some necessary concepts.

2.1. Emotional Speech Synthesis

Previous researches (e.g., [2, 3, 4, 5, 6]) have found that there are several features in human speech that are related with its affective content. These features refer to: different statistical values (e.g., max, mean, standard deviation, etc.) of the fundamental frequency F0; different statistical values of the first three formants (F1, F2, and F2); and their bandwidths (BW1, BW2, and BW3), energy, speaking rate, etc. Generally these features are derived by observing how human's voice changes accordingly to 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 [3, 4] these techniques are described along with their advantages and disadvantages. Moreover, techniques like explicit prosody control [1, 5, 7], expressivity based unit selection [8], HMM based parametric synthesis [10], non-verbal vocalization [11], etc., are quite popular and obtained partial success for recognizing anger and sadness in synthesized speech samples.

2.2. Sensing Affective Information from Text

This research addresses the aspect of subjective opinion, particularly the identification of different emotive dimensions and the classification of texts by their emotion affinity. It can be argued that the affective content of a text and its analysis depend on the audience, context and world knowledge. The assessment of affective information from text is based or in 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 dealt with the above techniques. For example, Shaikh et al. [12], implemented a technique based on contextual valence assignment and achieved tremendous results in recognizing different emotions (e.g. happiness, sadness, anger, etc.) from text and, Liu et al. [13], using common-sense knowledge could detect the six basic emotions in a text.

2.3. Mary TTS: A Flexible TTS System

The MARY TTS system [14] is a client-server application written in Java and created at DFKI GmbH. MaryXML serves as the configuration input language of this system and is a flexible toolkit for speech synthesis research. We have chosen MARY TTS system because it allows the dynamic creation of MaryXML with appropriate prosodic and accent properties related with the intended emotion and allows access to all intermediate processing results for purposes of debugging and analysis.

3. Our Approach

Our system deals with the six basic emotions: happy, sad, fear, anger, surprise and disgust. It performs affective evaluation of the input text and, accordingly to the emotional content of the input sentence, produces MaryXML that matches the desired prosodic parameters and the findings reported in [1, 2, 4, 5, 6,15]. This Dynamic MaryXML is used as input for MARY TTS system to assist the speech synthesis process.

3.1. System Architecture

The system architecture is based in the following modules: Language Processing, Affective Text Sensing and Generation of Dynamic MaryXML.

3.1.1. Language Processing

For each input sentence the language processing module outputs triplet(s) consisting of a subject or agent, a verb and an object. A XML formatted syntactic and functional dependency information for each word of the input text is obtained using the Machinese Syntax parser [21] and this output constitutes the basis for further processing that generates the triplet(s). For example, the sentence *"The car exploded near a popular ice cream parlor, sending flames and shrapnel through the busy square and killing 17 people."* produces three triplets as shown in Table 1.

3.1.2. Affective Text Sensing

We used the output of the system SenseNet developed by Shaikh et al. [12] that can process the triplet-formatted input of a sentence. SenseNet can perform affective sentence 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 sentence that is used as input. For example, SenseNet outputs -10.76 for the sentence referred above as an example. The output value indicates a numerical measure of negative or positive sentiments carried by the sentence. SenseNet implements a cognitive theory of emotion using rules as suggested by the model. Therefore it can classify input texts according to eight types of emotions, namely, happy, sad, hope, fear, admiration, shame, love and hate, plus a neutral category. Following an experimental study [12], the accuracy of SenseNet to assess sentence-level negative/positive sentiment is 91% and classification accuracy of eight emotion types is 82%.

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

Triplets processed by Semantic Parser	
Triplet 1	[[['*Actor:*', 'car', *'Actor-Type:'*, 'object', *'Actor-Attrib:'*, ['DET: the']], ['*Action-Name:*', 'explode', *'Action Status:'*, 'Past', *'Action-Attrib:',* ['place: near a popular ice cream parlor']], ['*Object-Name:*', '', *'Object-Type:'*, '', *'Object-Attrib:*', ['']]]
Triplet 2	[[['*Actor:*', '', *'Actor-Type:'*, '', *'Actor-Attrib:',* []], ['*Action-Name:',* 'send', *'Action-Status:'*, 'Present Progressive', *'Action-Attrib:'*, ['place: through the busy square']], ['*Object-Name:'*, 'flame and shrapnel', *'Object-Type:'*, 'N NOM', *'Object-Attrib:'*, ['']]]
Triplet 3	[[['*Actor:*', '', *'Actor-Type:'*, '', *'Actor-Attrib:'*, []], ['*Action-Name:'*, 'kill', *'Action-Status:'*, 'Present Progressive', *'Action-Attrib:'*, []], ['*Object-Name:'*, 'people', *'Object-Type:'*, 'N NOM', *'Object-Attrib:'*, ['Quantity: 17']]]]

3.1.3. Dynamic MaryXML Generation

After the input text has been processed as mentioned above, we obtain the affective assessment of the text: the overall emotion carried by the text; the positive or negative meaning of the events represented by the triplet(s); and the attributes (e.g., location, time, etc.) of the events that are considered important. First, several speech parameters are set for the overall negative or positive affective connotation of the text and then parameters like pitch, pitch-dynamics, number-of-pauses, etc., are adjusted accordingly to the detected emotions. For example, if a sentence 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. The phrasal tones (L-L%, L-H%, H-H%, and H-L%) and the pitch accents (peak, low, scooped, and rising peak) are considered at word and phrase level and are assigned using ToBI notation.

Thus, our system, from plain text, creates prosody-rich MaryXML that can be processed by the MARY TTS system in an affective context.

3.2. An example of dynamic MaryXML

The following code is an example of the dynamic MaryXML for the sentence referred in 3.1.1.

```
<?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>
```

4. Experiments and Results

We conducted some experiments to assess the validity of our approach and hypothesis. Experiment 1 analyzes the performance of TTS systems for conveying positive and negative emotions, with and without our approach. Experiment 2 has the same goal but, in this case, the systems were evaluated for 6 different emotions types.

4.1. Data Set

In Experiment 1, the dataset consists of 40 synthesized speech samples that were produced from 20 brief texts (RSS feeds) collected from online sources like BBC News, etc. For each text we created 2 versions of synthesized speech samples using the MARY TTS system. One is the output obtained from the plain text input and the other is produced by inputting dynamic MaryXML pre-marked with intonational information created by our approach. Both cases use the voice Mbrola-us2 version 3.5.0 and the length of each synthesized speech audio sample is 7 seconds on average.

In Experiment 2, the dataset consists of 18 speech samples (12 synthetic speech samples and 6 speech samples spoken by humans). For each of the six affective texts (each one pertaining to a different emotion), we created two versions of synthesized speech samples using the MARY TTS system. One is the output obtained by using the plain text and the other is obtained from the dynamic MaryXML outputted by our approach. Both cases use the voice Mbrola-us2, version 3.5.0, and the length of each synthesized speech audio sample is 17 seconds on average.

4.2. Experimental Procedure

In experiment 1, the two systems, the plain text input system (S1) and the dynamic MaryXML input system (S2). The goal is to assess how TTS systems convey positive and negative emotions, with and without our approach. The survey was conducted online at http://research.rebordao.net/emonews/ and a total of 30 people participated. The subjects had to listen to synthesized speech audio samples produced from S1 and S2 and assessed if they could perceive any emotion, or not. If an emotion would be perceived, it would be asked them to classify it as either positive or negative.

In experiment 2, we also have two systems, the plain text input system (S1) and the dynamic MaryXML input system (S2). The goal is to assess how TTS systems perform for different emotions, with and without our approach. We conducted a perceptual test through an online survey at http://research.rebordao.ne/emostory/ and had a total of 15 participants. The subjects had to listen to the synthesized speech audio samples produced by S1 and S2. They were asked to assess if they could perceive any emotion, or not. If an emotion would be perceived, it would be asked them to select an emotion among of the six basic emotions.

4.3. Results and Discussion

In Experiment 1, we considered the scores obtained from the web-survey for which, either one or both systems, received an emotion perception score. From Chi-Square test it is evident that the evaluation scores of the systems are statistically significant ($P<0.001$) regarding its emotional expressivity. The figure 4 shows that the systems performed almost similarly (i.e., accuracy 52.2%, 47.3% for S1 and S2 respectively) for conveying negative emotions. But figure 5 shows that S2 performed significantly better than S1 (i.e., accuracy 6.0%, 67.4% for S1 and S2 respectively) to convey positive emotions. In figure 5 it is evident that S1 has tendency to synthesize speech with a negative tone that the subjects perceive as sad or boring. These results are encouraging from two perspectives, one of them is that S1 is very weak to convey positive emotions (like happiness) and our approach can solve this problem. The system S1 has tendency to express negative emotions (like sadness) and our approach can be applied to incorporate different levels of negativism within the phrases of a synthesized sentence.

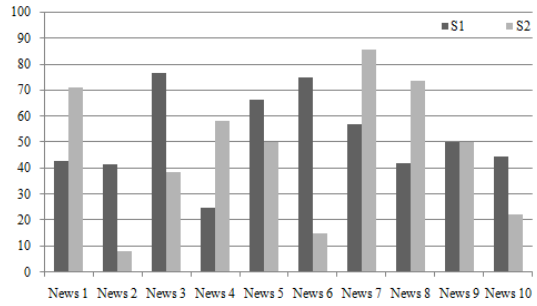

Figure 4: *Recognition rates of S1 and S2 for 10 negative news.*

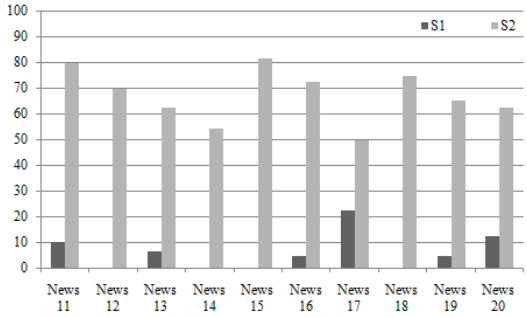

Figure 5: *Recognition rates of S1 and S2 for 10 positive news.*

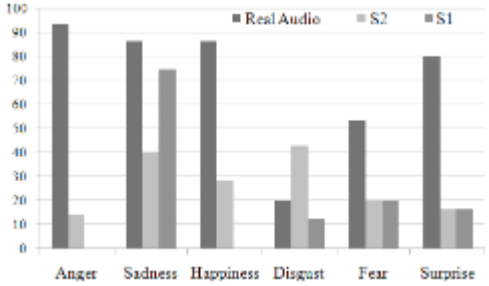

Figure 6: *Emotion recognition efficiency rates of the perceptual test for the Real Audio, S2 and S1.*

In Experiment 2 we considered the scores for which, either one or both systems, received an emotion perception score. Figure 6 shows that the subjects perceived easily the emotion from the real audio (excepting for disgust). Furthermore, the system S2 performed significantly better than S1 for conveying anger (improvement of 14.3%), disgust (improvement of 30.4%) and happiness (improvement of 28.6%) but for conveying sadness, S1 performs better (40% for S2 and 75% for S1). This could occur due to the above referred tendency of S1 to produce synthesized speech with intonational information related to negative emotions.

5. Conclusion

In our study we have found that several well-known TTS systems cannot synthesize affective speech. However, this situation can be improved by pre-processing the input in two manners, first by recognizing the emotions conveyed through the plain text and then controlling the synthesis process by assigning appropriate prosodic parameters that suit the detected emotions. Thus, the output of our system is an enhanced version of MARY-XML of the plain input text that is given to the TTS system (i.e., MARY TTS) to process. In future we plan to perform analysis on the synthesized samples and then re-synthesis of the samples doing TD-PSOLA based modification as suggested by Murtaza et al [15].

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] Oudeyer, P., The production and recognition of emotions in speech: features and algorithms. International Journal of Human-Computer Studies, 59, pp. 157–183, 2003.

[3] Schröder, M., Expressive Speech Synthesis: Past, Present, and Possible Futures, Affective Information Processing (Tao, J., Tan, T., eds.), pp. 111-126, 2009.

[4] 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.

[5] Morrison, D., Wang, R., & De Silva, L. C., Ensemble methods for spoken emotion recognition in call-centres. Speech Communication, 49, pp. 98–112, 2007.

[6] 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.

[7] Burkhardt, F., & Sendlmeier, W.F., Verification of acoustical correlates of emotional speech using formant synthesis. In Proc. ISCA Workshop on Speech and Emotion, pp. 151–156, 2000.

[8] Fernandez, R., & Ramabhadran, B., Automatic exploration of corpus-specific properties for expressive text-to-speech: A case study in emphasis. In Proc. 6th ISCA Workshop, Bonn, Germany, pp. 34–39, 2007.

[9] 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, pp. 326–329, Brisbane, 2008.

[10] Zen, H., & Toda, T., An overview of Nitech HMM-based speech synthesis system for Blizzard Challenge 2005. In Proc. of InterSpeech, Lisbon, Portugal, pp. 93–96, 2005.

[11] Campbell, N., Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues. In Proc. of the Intl. Congress of Phonetic Sciences, Saarbrücken, Germany, pp. 343–348, 2007.

[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 *Proc. Of the 8th international Conference on intelligent User interfaces,* IUI '03, ACM, Miami, USA, pp. 125-132, 2003

[14] M. Schröder and J. Trouvain, The German text-to-speech synthesis system MARY: A tool for research, development and teaching, Intl. J. Speech Technology, vol. 6, pp. 365–377, 2003

[15] Murtaza Bulut, Sungbok Lee, and Shrikanth Narayanan. Recognition for synthesis: Automatic parameter selection for resynthesis of emotional speech from neutral speech. In Proceedings of ICASSP, Las Vegas, Nevada, April 2008.