Assigning Suitable Phrasal Tones and Pitch Accents by Sensing Affective Information from Text to Synthesize Human-like Speech

Mostafa Al Masum Shaikh, Md. Khademul Islam Molla and Keikichi Hirose

Department of Information and Communication Engineering, University of Tokyo, Japan
{almasum, molla, hirose}@gavo.t.u-tokyo.ac.jp

10.21437/Interspeech.2008-97

Abstract

We have carried out several perceptual and objective experiments that show that the present Text-To-Speech (TTS) systems are weak in the relevance of prosody and segmental spectrum in the characterization and expression of emotions. Since it is known that the emotional state of a speaker usually alters the way s/he speaks, the TTS systems need to be improved to generate human-like pitch accents to express the subtle features of emotions. This paper describes a pitch accent assignment technique which places appropriate pitch accents on elements of the utterance that require particular emphasis or stress. Our pitch accenting technique utilizes commonsense knowledge-base and a linguistic tool to recognize emotion conveyed though the text itself. From these it determines whether the content of the utterance has a connotation to a particular emotion (e.g., happy, sad, surprise etc.), good or bad concepts, praiseworthy or blameworthy actions, common or vital information. It can then assign an appropriate pitch accent to one word in each prosodic phrase. The TTS component then determines the appropriate syllable to be accented in the word. Our approach can well support a TTS system's synthesis, allowing the system to generate affective version of the spoken text.

Index Terms: speech synthesis, pitch accents, TTS, emotion in speech

1. Introduction

One of the known issues to be involved in speech production and perception is how natural speech synthesis sounds. Expressive or emotion eloquence is one of the relevant issues that contribute to the natural sounding of synthesized speech. Emotional connotation is what brings a spoken text or speech to life. Whether a speaker is delighted, depressed, enraged or repulsive can be easily identified by the way the speech is being articulated. A unified tone, proper pitch accent, and suitable intensity of speech can help to convey attitude or intent of the speech both contextually and in a content-rich manner. Thus, if a Text-To-Speech (TTS) system is capable to generate human-like speech then it can serve to convince, persuade, or appeal to a particular audience more successfully.

TTS primarily aims to improve the intelligibility of generated speech so that the output resembles human-like articulation. Contemporary TTS systems tend to read text in a way that sounds unnatural. This is partly due to deficiencies in syntactic analysis of raw input text, but also due to the lack of semantic information, affective clues, and world knowledge. We have carried out several perceptual and objective experiments that show that the present TTS systems are weak in the relevance of prosody and segmental spectrum in the characterization and expression of emotions. We collected several sentences from online news sources and employed three off-the-shelf TTS systems (i.e., RealSpeak [1], Festival [16], Microsoft TTS) to synthesize the collected news texts. In Figure 1, the pitch-accent diagrams of speech samples synthesized by Nuance RealSpeak [1] for the following two sentences are given. Our sample sentences are:

Sentence 1 (S1): *The car exploded near a popular ice cream parlor, sending flames and shrapnel through the busy square and killing 17 people.*

Sentence 2 (S2): *Girls in traditional costume take part in a parade in the Azabu Juban district in Minato Ward, Tokyo, on Sunday as part of a "flower festival" to celebrate the Buddha's birthday on April 8. More than 150 children participated in the annual event.*

Since these sentences have obvious affective connotation (i.e., sad/fear and happy/joy respectively) ideally the TTS should be capable to synthesize speech to express such emotions by the produced speech samples. But we hardly find those subtle features that signal emotional affinity on analyzing the auto-synthesized speech samples. Of the three TTS systems, the output of RealSpeak sounds most natural. Therefore, in Figure 1 the output of RealSpeak is shown.

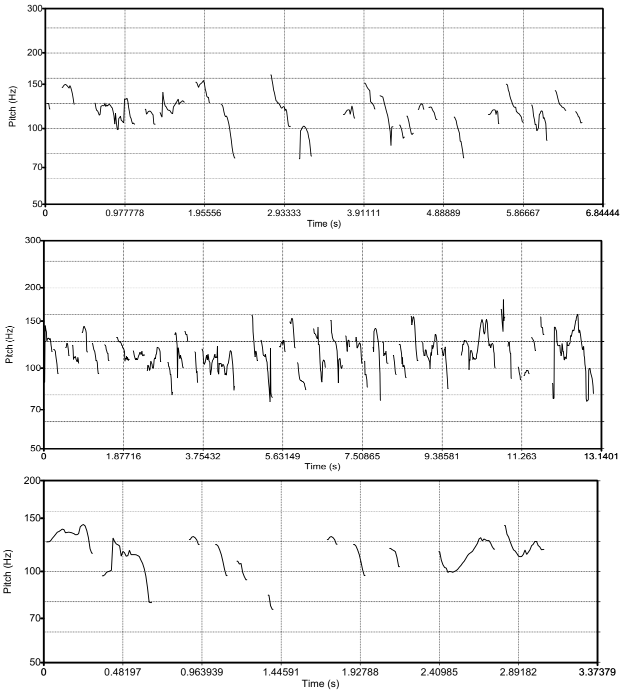

Figure 1. Pitch contour diagram of RealSpeak synthesized samples for S1 (top), S2 (mid), and a neutral sentence (bottom).

In Figure 1, we notice that there is inappropriate pitch accent assignment in the speech and moreover the pitches of both speech samples are very similar relative to speech sample of a neutral emotion carrying sentence. But the emotional affinities are totally different in this case. Table 1 shows different speech parameters of the synthesized speech samples of the aforementioned two sentences in order to indicate the failure to encode emotional connotation in the synthesized speech. Though some TTS systems accept input text pre-marked with intonational information but there is hardly any good annotation system we noticed that make necessary text pre-processing. Our research finds the niche at this point. We perform text processing to annotate appropriate pitch accent to words or phrases with the help of commonsense knowledge and emotion sensing technique for texts before the synthesis.

Table 1. *Speech parameters of the synthesized speech samples of the two sentences analyzed by PRAAT*

TTS Name	Input	SR	PA	PR	I	PC
TTS 1	S1M	4.82	118.46	75.76 163.37	61.26	216.65
	S1F	4.40	171.07	90.85 230.05	70.79	362.91
	S2M	5.40	115.80	70.89 180.97	67.72	280.97
	S2F	5.20	173.08	92.14 338.67	73.24	423.20
TTS2	S1M	4.15	102.77	80.45 127.00	73.35	67.43
	S1F	3.98	171.80	137.03 208.80	74.58	208.49
	S2M	3.85	100.02	74.38 205.83	73.23	75.78
	S2F	3.94	171.16	103.97 224.34	74.82	213.49
TTS3	S1M	3.58	98.94	75.42 120.70	87.62	90.52
	S1F	3.50	209.11	140.29 313.59	88.01	210.13
	S2M	3.78	99.83	75.55 126.97	87.37	109.97
	S2F	3.58	199.55	140.87 279.40	88.13	194.95

SXX: Sentence 1 or 2 Synthesized by Male or Female Voice SR: Speech Rate (syllable/sec), PA: Pitch Average, Pr: Pitch Range, I: Intensity (dB), and PC: Pitch Changes (Hz/sec).

According to [2][3], we know that in order to signal "sadness" in speech the SR and PA should be slightly slower; PR should be slightly narrower; Intensity should be lower; and PC should be downward inflections with respect to a neutral speech. On the contrary, to signal "happiness" the SR should be faster or slower; PA should be much higher; PR should be much wider; Intensity should be higher; and PC should be smooth upward inflections. However, analyzing the values mentioned in Table 1, we cannot observe the aforementioned phenomena while evaluating the values with respect to a neutral speech sample of those TTS systems.

2. Related Work

Though tremendous effort has gone into the synthesis of speech from text as well as identifying of emotions from human spoken speeches, as far as we know there is no system that takes the content (e.g., typed text from a speaking-impaired person like Stephen Hawking) and the affective values of the content are automatically generated with the right settings for a TTS engine. Instead, we have found a system called *Affect Editor* [3] where the speaker has to adjust the affect parameters by hand. Affect Editor takes an acoustic and linguistic description of an utterance and generates synthesizer instructions for a DECtalk3 synthesizer to produce speech with a desired affect.

Previous research (e.g., [2],[4]) have found that there are several features in speech that are adhering to specific affective connotation. These features are, namely, different statistical values (e.g., max, mean, standard deviation etc.) of fundamental frequency F0, different statistical values of first three formants (F1, F2, and F2) and their bandwidths (BW1, BW2, and BW3), energy, speaking rate, etc. These features are generally derived by observing how human's voice changes with different emotions. Several studies (e.g., [5]) have established the fact 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.

There are several concept-to-speech systems (e.g., [6]) that generate spoken descriptions of something. Usually such systems have a Natural Language Generation (NLG) component that generates textual descriptions and then those descriptions are marked with intonational information. The NLG component has a support for assigning pitch accent which places appropriate pitch accents on elements of the utterance requiring particular emphasis or stress. Such systems employ specific knowledgebase or corpus support to identify appropriate word to assign pitch accent. But we did not find such a system that also pre-process affective information before synthesizing the speech. There are researches that do automatic prosody modeling to improve TTS by constructing and systematic analysis of a prosodic database. Such systems process prosody usually in acoustic, phonetic, and phonological levels.

Hence it is evident that perception of human emotion is plausible from the prosodic properties of speech and researches have enlisted most of the features of a speech to signal emotion, intonation etc. We also have synthesizers that can be manually tweaked to output human-like speech from an input text. Though automatic prosody controlled TTS are attempted, most of them had given less emphasis on the text-processing rather emphasized on acoustic level knowledge. So our primary contribution lies in the text-processing zone where extensive linguistic processing is done to assign appropriate speech parameters for a synthesizer to synthesize emotion-embedded speech.

3. Our Approach

As mention in [7], the vocal intonation of how something is said elucidates two components: cues emphasizing which content in the message is most important, and cues arising from the speaker's affective state. Our approach targets both of the components from linguistic standpoints. We assign phrasal tones, e.g., L-, H-, etc., and pitch accents, e.g., H*, L+H*, etc., according to ToBI [8] annotation.

3.1. System Architecture

We propose a pipelined architecture with the following phases: Parse, Assess, Annotate, and Synthesize. The system components are briefly described as following.

3.1.1. Semantic Parser

For each input sentence the Semantic Parser outputs triplet(s) consisting of a subject, a verb, and an object. Each member of the triplet may or may not have associated attribute(s) (e.g. adjective, adverb etc.). We first obtain XML-formatted syntactic and functional dependency information of each word of the input text using the Machinese Syntax parser [9] and this output constitutes the basis for further processing to generate the triplet(s). Basically a triplet encodes information about *"who is associated with what and how"*. For example, the aforementioned first example sentence produces three triplets as shown in Table 2.

Table 2. *Triplet Output of Semantic Parser*

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']]]]

3.1.2. Valence Assignment

Our system 'SenseNet' described in [10][11] begins with a lexicon of words with prior valence values using WordNet [12] and ConceptNet [13], and assigns the contextual valence of each semantic verb-frame by applying a set of rules. For example, the Triplet 1 encodes an event as "car explodes" having "place" attribute. In our system each word, i.e., "car", "explode", "near", "popular", "ice cream", "parlor" has an initial score (i.e., valence) either pre-assigned or automatically assigned (for new words). Thereby we have negative valence for the word "explode" and positive valence for the other words associated with event. The system applies several computational linguistic rules to assign a resultant value (i.e., contextual valence) for the triplet. Thus Triplet 1 finally gets a negative score due to having a negative action associated with other positive entities. We have utilized cognitive and commonsense knowledge resources [13] to assign semantic orientation (SO) [14] of the event in order to know whether the event is praiseworthy or blameworthy. Thus we get a negative score for the event "car explodes" that qualifies the event as blameworthy. An event is also tagged as "common" or "uncommon" according to the average familiarity valence obtained from ConceptNet as described in [15]. The object of an event is tagged as "attractive" or "not attractive" based on two scores, namely, 'object valence' and 'familiarity valence' as described in [15].

3.1.3. Affect Sensing From Text

SenseNet assesses 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 examples, SenseNet outputs -11.16 and +9.57 for the inputs, "*The attack killed three innocent civilians*." And *"The President called the space shuttle Discovery on Tuesday to wish the astronauts well, and congratulate them"*, respectively. These values indicate a numerical measure of negative and positive sentiments carried by the sentences. SenseNet can classify the news-texts according to eight emotion-types namely, Happy, Sad, Hopeful, Fearful, Admirable, Shameful, Loveable, and Hatred plus a Neutral category. The accuracy of SenseNet to assess sentence-level negative/positive sentiment is 91% and classification accuracy of eight emotions is 82% in an experimental study.

3.1.4. Pitch Accent Annotator

After the input text has been processed as mentioned above, we obtain several linguistic cues like: the overall emotion carried by the text, positive or negative feeling of the events mentioned in the text, the attributes (e.g., location, time, etc.) of the events, quality of the events (i.e., praiseworthy/ blameworthy, common/uncommon), and quality of the targeted object of the event. First, several speech parameters are set for the overall synthesis adhering to the overall affective connotation of the text with respect to neutral emotion expressing speech. For example, if the text would have "happy" 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. Then following rules are applied to annotate suitable phrasal tone and pitch accent to be processed by synthesizer during synthesis.

Phrasal tones are assigned at every intermediate or intonation phrase. Four types of phrasal tones, L-L%, L-H%, H-H%, and H-L% are considered. The rules to annotate phrasal tones are:

- Tones are assigned by considering verb-phrase, noun-phrase and object-phrase at Triplet level. A Triplet basically has two parts, the event and event's attribute. Both event and event attributes may have affective values. Based on their affective values tones are annotated.
- If an event shows negative affect and is associated with a positive actor (e.g., car exploded) then H-L% else L-L% (e.g., the attack killed). But, if an event shows positive affect associated with positive actor and action (i.e., girls take-part in a parade) then H-H% else if both actor and action are negative (e.g., the criminal was shot) then L-H% is assigned.
- If an event shows negative affect associated with a positive action and negative object (e.g., sending suicide-bomber) then H-L% else if an event has negative action with a positive object (e.g., killing people) then L-L% is assigned.
- If an event's attribute has a positive adjective with a positive entity (e.g., busy square) then H-H%, if negative adjective with either positive or negative entity (e.g., alone in the apartment, terrible murder) then L-L%, if positive adjective with negative entity (e.g., popular crime zone) then H-L% is assigned.

Pitch accent tones are marked at every accented syllable. The system annotates for peak accent (H*), low accent (L*), scooped accent (L*+H) and rising peak accent (L+H*) on word level considering whether the word represents a verb or object or attribute cue of an event. In this case some of the rules are:

- If a verb word has negative value associated with an event having certain values of blameworthy and uncommon variables then L*+H is assigned.

- If a verb word has positive value associated with an event having certain values of praiseworthy and uncommon variables then L+H* is assigned.
- If a verb has positive or negative meaning but the event doesn't have certain values for either of those two variables then H* and L* are assigned respectively.
- If the word is an attribute cue (e.g., near, on) that complements the description of location, time etc. of an event then H* assigned to emphasize it as vital information being spoken.
- Similar rules like verb words are applied to nouns considering the value of 'attractive' variable. Details are not given due to space limitation.

3.1.5. Speech Synthesizer

We have used Festival [16]. Festival is kept running in server mode setting it to ToBI input mode. Festival generates an audio file which is considered as the output of the system.

3.2. Examples and Evaluation

Annotation of the aforementioned two sentences are given and Figure 2 shows the pitch-accent diagram of Festival [16] synthesized speech of the first sentence that expresses "sad" emotion (i.e., slower pitch average, lower intensity, downward pitch-change comparing to the output produced by Festival for the same un-annotated sentence).

Annotation for the first sentence: (H-L% the <H*>car <L*>exploded) <H*>near (H-H% a <H*>popular ice cream parlor), (H-L% <L*+H>sending flames and shrapnel) <H*>through (H-H% the <H*>busy square) and (L-H% <L*>killing <H*>17 people)

Annotation for the second sentence: (H-H% Girls in traditional costume <H*>take part in a <L+H*>parade) <H*> in (the Azabu Juban district in Minato Ward, Tokyo,) <H*>on (H-H% Sunday as part of a <L+H*>"flower festival") (H-H% to <H*>celebrate the Buddha's <H*>birthday) <H*>on (April 8). (H-H% More than <H*>150 children <L+H*>participated) in (H-H% the <H*>annual event)

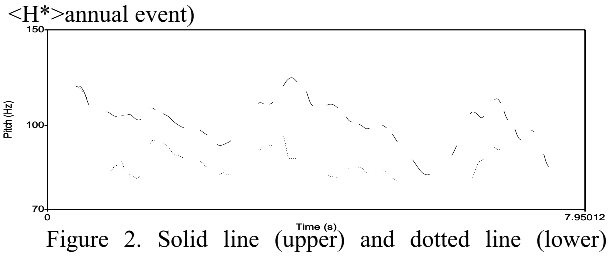

Figure 2. Solid line (upper) and dotted line (lower) indicate the pitch accent of speech samples synthesized from un-annotated and annotated input (i.e., S1) respectively.

Twenty sentences from news genre are collected from online sources and auto-annotations are done by the technique mentioned. Of the input texts, 8 are positive (i.e., 6 and 2 sentences express 'happy' and 'surprise' emotion respectively), 10 are negative (i.e., 5, 3 and 2 sentences express 'sad', 'fear' and 'disgust' emotion) and 2 are neutral sentences. The speech samples of the annotated texts are generated by Festival speech synthesizer. Obtained speech samples are considered for perceptual evaluations by five listeners. We got 85% average accuracy to distinguish negative, positive and neutral speech samples as well as 72% average accuracy of recognizing of individual emotions by the judges. Judges have the highest and lowest recalls for "happy" (92%) and "neutral" (46%) speech samples.

4. Conclusion

In this paper we have shown how a synthesizer can be benefited by providing suitable ToBI annotation utilizing several linguistic resources. The output of synthesized speech of the annotated text seems more natural while comparing to sample speeches synthesized for the same input by the TTS engine. Moreover the synthesized speech of annotated text contains the features that are regarded as the essential speech parameters to signal specific emotions. We plan to build a tool combining all the resources discussed above so that a speech impaired person can type a text and then synthesized speech is generated conveying appropriate emotion. Such a system maybe used to read out story books for the kids.

5. References

[1] Nuance RealSpeak, Online: http://www.nuance.com/realspeak/

[2] Murray, I.R. and Arnott, J.L., "Toward the simulation of emotion in synthetic speech: A review of the literature on human vocal emotion", Journal Acoustical Society of America, 93(2):1097-1108, Feb. 1993.

[3] Cahn, J.E., "The generation of affect in synthesized speech", Journal of the American Voice I/O Society, 8:1-19, July 1990.

[4] Barra, R., Montero, J.M., Macias-Guarasa, J., D'Haro, L.F., San-Segundo, R. and Cordoba, R., "Prosodic and segmental rubrics in emotion identification", In Proc. ICASSP, pp. 1085-1088, May. 2006.

[5] Pollermann, B.Z. and Archinard, M., "Acoustic patterns of Emotions", In E. Keller, G. Bailly, A. Monaghan, J. Terken, and M. Huckvale, editors, Improvements in Speech Synthesis, pp. 237-245, April 2002.

[6] Williams, S., "Generating Pitch Accents in a Concept-To-Speech System Using a Knowledge Base", In Proc. 5th ICSLP, Vol. 4, pp. 1159-1162, Dec 1998.

[7] Picard, R. W, Affective Computing. Cambridge, MA: The MIT Press, 1997.

[8] Silverman, K., Beckman, M., Petrelli, J., Ostendorf, M., Wrightman, C., Price, P., Pierrehumbert, J. and Hirschberg, J., "ToBI: a standard for labeling English prosody", ICSLP92, vol. 2, pp. 867-870, 1992.

[9] Machinese Syntax, http://www.connexor.com/connexor/

[10] Shaikh, M. A. M., Prendinger, H., and Ishizuka, M., "SenseNet: A Linguistic Tool to Visualize Numerical-Valance Based Sentiment of Textual Data", In Proc. ICON-07, pp. 147-152, India, 2007.

[11] Shaikh, M. A. M., Prendinger, H., and Ishizuka, M., "Sentiment assessment of text by analyzing linguistic features and contextual valence assignment", to appear in Journal of Applied Artificial Intelligence, Taylor & Francis.

[12] Fellbaum, C. (ed.): WordNet: An Electronic Lexical Databases, MIT Press, Cambridge, Massachusetts, 1999

[13] Liu, H. and Singh, P.: ConceptNet: A Practical Commonsense Reasoning Toolkit, BT Technology Journal 22(4), 211-226. Kluwer Academic Publishers, 2004

[14] Grefenstette, G., Qu, Y. Evans, D., and Shanahan, J., "Validating the Coverage of Lexical Resources for Affect Analysis and Automatically Classifying New Words along Semantic Axes", In J. Shanahan, Y. Qu, and J. Wiebe, editors, Computing Attitude and Affect in Text: Theory and Applications, pp. 93-107. The Information Retrieval Series Vol.20, Netherlands: Springer Verlag, 2004.

[15] Shaikh, M. A. M., Prendinger, H., and Ishizuka, M., "Roles of Emotions: A Linguistic Interpretation of an Emotion Model for Affect Sensing from Texts", A. Paiva, P. Prada and R. W. Picard, editors, Affective Computing and Intelligent Interaction, Springer LNCS 4738, pp.737-738.

[16] Black, A.W. and Taylor, P.A., "The Festival Speech Synthesis System: System documentation", Technical Report HCRC/TR-83, Human Communciation Research Centre, University of Edinburgh, Scotland, UK, 1997. Online: http://www.cstr.ed.ac.uk/projects/festival.html