MARY TTS HMM-based voices for the Blizzard Challenge 2012

Marcela Charfuelan

DFKI GmbH, Language Technology Lab
Berlin Germany
marcela.charfuelan@dfki.de

Abstract

This paper describes the first participation of MARY TTS HMM-based voices in a Blizzard challenge. An architecture for synthesis of expressive speech based on the MARY TTS system and sentiment analysis of text is proposed. The creation of several HMM-based voices in different styles using audiobook data is described. Preliminary results on perception of different voice styles and the appropriateness of a given style for a given sentence are presented. The latest developments in the open source MARY TTS 5.0 are briefly described.
Index Terms: speech synthesis, parametric speech synthesis, expressive speech, sentiment analysis, signal processing.

1. Introduction

This paper describes the fifth participation of the MARY TTS system in a Blizzard Challenge. The previous four entries were unit selection speech synthesis systems [1], this year is the first time that the MARY TTS entry is a HMM-based parametric speech synthesis system. The task in this year's challenge was to build a synthetic voice from audiobook recordings where a single speaker reads four books by Mark Twain. The narration in the audiobooks is lively and expressive and the speaker impersonates or performs several characters apart from the narrator himself.

From a theoretical point of view, narratives have been studied as a context for the integration of language and emotion. According to [2] evaluative information in narratives can be conveyed in several ways: lexically, syntactically and paralinguistically by emotional facial expression, gesture and affective prosody. Opinions, sentiment and emotions expressed in text are also studied in the relatively new area of sentiment analysis [3]. Motivated by these two ways of expressing and representing emotions we carried out a preliminary study about possible correlation of acoustic features extracted from audiobook sentences and sentiment analysis scores extracted from the corresponding text sentences [4]. In this study it was found that scores derived from movie reviews or categorisation of emotional stories seem to be more close to the acoustics in the narrative, in particular more correlated with average energy and mean fundamental frequency (F0); also it was shown that the voice style of a sentence could be, to some extent, automatically derived from textual data and a trained model.

Based on the results in [4], in this paper we propose an architecture for synthesis of expressive speech based on the MARY TTS system and sentiment analysis of text. Although the architecture is under development and the voices created do not reach yet a good quality, it was decided to participate in the challenge because it is an invaluable opportunity to get feedback, in particular in this year's challenge, regarding evaluation of expressivity in synthetic voices.

The paper is organised as follows. Section 2 describes what is new in MARY TTS 5.0, with emphasis on the support of HMM-based voice creation. Section 3 describes how different HMM-based voices in different styles were created using audiobook data; this section also describes how these voices and sentiment analysis are used in an architecture of expressive speech synthesis based on MARY TTS. Preliminary results about voice styles perception and appropriateness of a given style for a given sentence are presented in Section 4; here the results of our entry in the challenge are also discussed. Conclusions, lessons learnt and future work are presented in Section 5.

2. MARY TTS 5.0

The MARY TTS platform[1] is an open-source, modular architecture for building text-to-speech systems, including unit selection and statistical parametric waveform synthesis technologies [5]. The code in the latest release, MARY TTS 5.0, has been thoroughly restructured, main new features in this release are:

- simpler installation
- simplified use of MARY TTS in your own projects
- new MaryInterface API
- emotion Markup Language support

Information about these new features and the new modularised code can be found in the marytts github repository[2], where it is now maintained. For building HMM-based voices in the MARY TTS framework we use the latest version of the scripts provided by HTS [6], in particular MARY TTS 5.0 includes the HTS-2.2 for HTK-3.4.1 training scripts, which have been modify to:

- use monophone and full context feature labels extracted with the MARY text analyser,
- generate a questions file for tree building, depending on the MARY context features selected for training the HMMs,
- generate and use band-pass voicing strengths during training for mixed excitation generation.

Detailed description of this procedure can be found in [7]. For run-time synthesis using HMM-based voices, MARY TTS includes a ported version to Java of the latest HTS-Engine (hts_engine_API-1.05). This Java HMM-based synthesiser is fully integrated into MARY TTS and has additional possibilities like:

- support for explicit prosody specification using the "prosody" element of the Speech Synthesis Markup Language (SSML) [8]. Examples of adjusting speech rate

[1] http://mary.dfki.de
[2] https://github.com/marytts/marytts

or pitch level and shaping intonation contour using the markup are described in [9].

- preliminary support for requesting expressive synthetic speech using EmotionML [10] in terms of discrete emotions: angry, happy or sad; or in terms of continuous values for emotion dimensions: arousal, pressure and dominance. EmotionML examples are available on the MARY TTS demo page[3]. EmotionML support is also available in some unit selection voices.

3. Building of HMM-based voices in different styles

We have designed an architecture for synthesising expressive speech as shown in Figure 1. Expressive speech for arbitrary text is realised by first of all extracting sentiment analysis scores from the input text, these scores together with the number of words and number of quotations in the text are passed to a voice style prediction model which determines which voice style to use for synthesis. We use MARY TTS as a base and created HMM-based voices with different styles using audiobook data. The voice style prediction model is trained with features extracted from text of the same audiobook. More details about the creation of the voice style prediction model and the voices in different styles is given below.

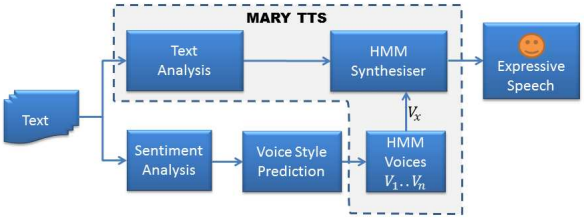

Figure 1: Expressive speech synthesis architecture based on MARY TTS 5.0.

From the four audiobooks available in the challenge we have used "The adventures of Tom Sawyer". The audiobooks were already split into prosodic phrase level chunks. The sentence segmentation and orthographic text alignment of the audiobook has been performed using an automatic sentence alignment method - LightlySupervised - as described in [11]. From the selected audiobook, we have discarded the sentences with confidence value $< 100\%$ as well as sentences with more than 30 words. The number of sentences used was 3676.

As described in more detail in [4], we have extracted sentiment analysis scores from the text sentences and acoustic features from the corresponding audio sentences, the following is a short summary of scores and features extracted:

Sentiment scores:

- Scores derived from IMDB reviews using machine learning techniques [12]:
  - ImdbEmphasis: a sentiment score for emphasis vs. attenuating
  - ImdbPolarity: a sentiment score for positive vs. negative
- OpinionLexicon, sentiment scores by lexicon lookup using Bing Liu's lexicon, which is a list of positive and negative opinion words or sentiment words for English (around 6800 words) that has been compiled over many years [13].
- SentiWordnet, wordNet entries with added sentiment scores (negative and positive value):
  - SentiWordNetNeg
  - SentiWordNetPos
- Scores derived from the Experience Project, this project is a social networking website that allows users to share stories about their own personal experiences, users write typically very emotional stories about themselves, and readers can then chose from among five reaction categories to the story [14]. Data from this project has been used to derive the following reaction scores:
  - Hugs: Sympathy reader reaction score
  - Rock: Positive-exclamative reader reaction score.
  - Teehee: Amused/light-hearted reader reaction score.
  - Understand: Solidarity reader reaction score.
  - Wow: Negative-exclamative reader reaction score.
- Predicted negative (Neg) and positive (Pos) probability derived by training a model with the previous scores:
  - Neg, Pos
  - Polar: calculated as Pos-Neg, this is a kind of predicted polarisation score.

Acoustic features:

- F0 and F0 statistics, mean, maximum, minimum and range. F0 values were extracted with the snack tool [15].
- Duration in seconds per sentence.
- Average energy, calculated as the short term energy averaged by the duration of the sentence in seconds.
- Number of voiced frames, number of unvoiced frames and voicing rate calculated as the number of voiced frames per time unit.
- F0 contours, as in [16] we have extracted slope (a1), curvature (b2) and inflexion (c3); these measures are estimated by fitting a first-, second- and third-order polynomial to the voiced F0 values extracted from each sentence:

$$y = a_1 * x + a_0 \quad (1)$$

$$y = b_2 * x^2 + b_1 * x + b_0 \quad (2)$$

$$y = c_3 * x^3 + c_2 * x^2 + c_1 * x + c_0 \quad (3)$$

- Voicing strengths estimated with peak normalised cross correlation of the input signal [17]. Seven bandpass voicing strengths are calculated, that is, the input signal is filtered into seven frequency bands; mean statistics of these measures are extracted.

[3] http://mary.dfki.de:59125/

3.1. Data partitioning and voice style prediction

As in [4] we have used sentiment scores to predict a measure of "expressivity" that depends on the acoustic features. Our measure of expressivity is the first principal component value (PC1) after computing principal component analysis (PCA) of all the acoustic features extracted from the data.

A PC1 value per sentence was calculated and used to split the data into several sets which correspond to several styles. Figure 2 (a) shows the distribution of data according to PC1 and the five sets in which the data was split. Informal listening test of sentences in the different sets was performed, perceptual differences were found among the different sets that seems to correspond to variations of the "arousal" dimension.

Quartile statistics of PC1 were used for partitioning the data into the following sets:

$$veryhigh \quad : \quad k2 \times Q_3 <= PC1 \tag{4}$$
$$high \quad : \quad k1 \times Q_3 < PC1 < k2 \times Q_3 \tag{5}$$
$$center \quad : \quad k1 \times Q_1 <= PC1 <= k1 \times Q_3 \tag{6}$$
$$low \quad : \quad k2 \times Q_1 < PC1 < k1 \times Q_1 \tag{7}$$
$$verylow \quad : \quad PC1 <= k2 \times Q_1 \tag{8}$$

where Q_1 and Q_3 are the first and the third quartiles of PC1 and k_1 and k_2 are constants empirically designed to generate similar densities for levels in the center and the extremes, where the data is more sparse, see Figure 2 (b).

Multiple linear regression (MLR) of sentiment scores, number of words and number of quotations were used to train a prediction model of the acoustic PC1 feature; sequential floating forward selection (SFFS) was used to find the best sentiment score predictors. The learnt parameters after the SFFS multiple linear regression are:

$$\begin{aligned} PC1 = & -0.74 - 3.55 \times Wow + 0.60 \times num\_quotes \\ & + 0.071 \times num\_words + 55.75 \times ImdbEmphasis \\ & + 5.49 \times Understand - 3.99 \times SentiWordNetNeg \\ & - 2.67 \times Hugs - 10.02 \times ImdbPolarity \\ & + 1.21 \times OpinionLexicon + 1.6 \times SentiWordNetPos \end{aligned} \tag{9}$$

Using this equation a PC1 value is predicted for the test sentences and mapped into the five possible levels (and voices) using equations 4-8. The number of sentences, average fundamental frequency (F0) and some average sentiment scores for each set are presented in Table 1. General tendencies for averaged measures among the sets can be observed in this table; for example average F0 values are particularly different among the extreme sets, which were found to be perceptually more expressive in [4]; the average number of words in these extreme sets is also relatively lower than in the center sets, which confirms the fact that shorter sentences tend to be more expressive [18]. The average values for the two sentiment scores "Wow" and "Imdb-emphasis", which were found to be some of the best predictors in equation 9, also show clear tendencies among the sets.

In the preparation of the test sentences for our entry in the challenge, we have extracted sentiment analysis scores for all the sentences and use the voice style prediction model described above to select the voice for synthesise them. In particular for paragraphs, each sentence was processed and synthesised individually and afterwards concatenated. As an example in Table 2 the pre-processing of a test paragraph, including the predicted voice style for the split sentences, is presented.

Level	No. sent.	F0	No. words	No. quotes	Wow	Imdb-Emphasis
veryhigh	765	146.5	9.2	0.84	0.003	0.00266
high	615	123.8	11.1	0.75	0.034	0.00133
center	755	112.8	12.3	0.52	0.040	0.00047
low	813	103.0	11.0	0.40	0.569	-0.00020
verylow	728	92.0	6.9	0.28	0.921	-0.00011

Table 1: Distribution of sentences among the five sets used to create HMM-based voices. Average values per set of Fundamental frequency (F0), Number of words and Number of quotes in the text, and the sentiment scores Wow and ImdbEmphasis.

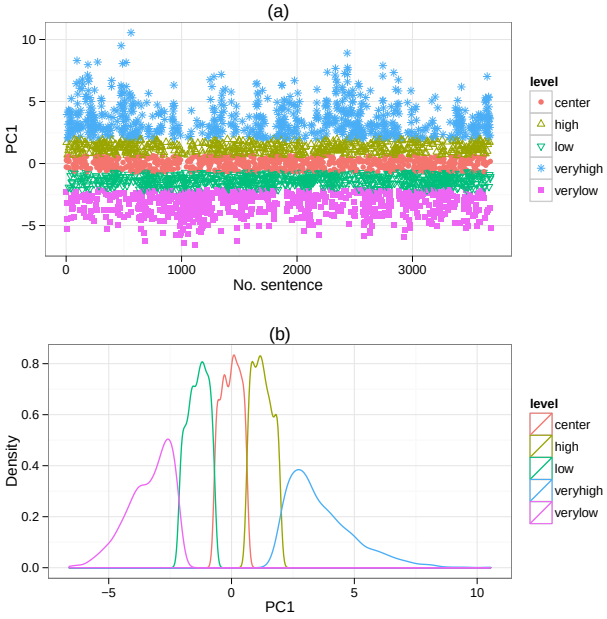

Figure 2: Data selection for building HMM-based voices in five styles: (a) distribution of sentences per set according to PC1. (b) Density of sentences in each set.

4. Listening tests results

Being able to synthesise expressivity and emotions is one of the challenges in this year's Blizzard, another challenge is to be able to evaluate expressivity features in the speech synthesised. The listening test in this year's challenge included evaluation of expressivity aspects like: pleasantness, pauses, stress, intonation, emotion and listening effort; also the traditional mean opinion score (MOS) for similarity to original speaker and naturalness, and word error rate (WER) were evaluated.

The results for our entry in the challenge were below average in all the aspects evaluated, particularly lower rates were obtained in MOS for similarity to original speaker and naturalness. We were aware of the low quality of the voices and submitted the test sentences anyway with the objective of getting some feedback regarding expressivity, in particular on the synthesis of paragraphs. It is clear from the results that the speech quality of our entry, in comparison to other entries, was a drawback. However, a more detailed analysis of the sentences used for the Blizzard evaluation indicates that the features for which

Sentence	Predicted style	Text
bookpara2_2012_0049_01	low	Lucy sat down at the table.
bookpara2_2012_0049_02	center	Miss Bartlett, who was thoroughly frightened, took up a book and pretended to read.
bookpara2_2012_0049_03	center	She would not be drawn into an elaborate speech.
bookpara2_2012_0049_04	center	She just said: "I can't have it, Mr. Emerson.
bookpara2_2012_0049_05	high	I cannot even talk to you.
bookpara2_2012_0049_06	high	Go out of this house, and never come into it again as long as I live here - " flushing as she spoke and pointing to the door.
bookpara2_2012_0049_07	center	"I hate a row.
bookpara2_2012_0049_08	high	Go please."
bookpara2_2012_0049_09	center	What -
bookpara2_2012_0049_10	center	No discussion.
bookpara2_2012_0049_11	center	But I can't -
bookpara2_2012_0049_12	verylow	She shook her head.
bookpara2_2012_0049_13	high	"Go, please.
bookpara2_2012_0049_14	center	I do not want to call in Mr. Vyse."
bookpara2_2012_0049_15	veryhigh	You don't mean, he said, absolutely ignoring Miss Bartlett - "you don't mean that you are going to marry that man?"
bookpara2_2012_0049_16	low	The line was unexpected.
bookpara2_2012_0049_17	low	She shrugged her shoulders, as if his vulgarity wearied her.
bookpara2_2012_0049_18	center	"You are merely ridiculous," she said quietly.

Table 2: Predicted voice style for split sentences of a paragraph in the Blizzard listening test.

our system was designed were hardly evaluated. Table 3 shows the predicted style (according to our trained model) for the sentences and paragraphs used in the listening test. This table indicates that actually the extreme voice styles, veryhigh and verylow, were almost not used. This might explain why users rated very low intonation and emotion, since most of the sentences and paragraphs (54%) were synthesised with the center or neutral voice style.

	Predicted level				
Type	veryhigh	high	center	low	verylow
bookpara1	6	15	42	13	0
bookpara2	3	15	32	17	0
booksent	0	4	23	10	0
news	0	0	16	4	0
sus	0	1	6	13	0
Total	9	35	119	57	0
Total(%)	4%	16%	54%	26%	0%

Table 3: Predicted voice style for test sentences in the Blizzard listening test.

4.1. Appropriateness of a style for a sentence and perception of a style

So far what it is clear from the Blizzard results is that: (i) the speech quality affected heavily the evaluation, (ii) users did not perceive expressivity variation among the high, center and low styles and (iii) the selection of the test sentences has an impact on the evaluation, in particular for our system. In fact, as it was shown in [19], where evaluation of synthetic speech in audiobook reading tasks is investigated, the selection of text has a significant influence to the subjective assessment of synthetic speech.

In our study we are interested to know whether users perceive that an style fits better or is more appropriate for a given sentence and if their preference is somehow in agreement with the style automatically predicted through sentiment analysis. Another aspect to evaluate is whether users can perceive the different styles of the voices created, in particular for the extreme styles, which were practically not used in the Blizzard test.

In order to test these aspects, we have performed a preliminary informal listening test, where two experiments were designed:

1. in the first one, users were presented with a sentence synthesised in three styles: veryhigh, center and verylow, and asked to select one that in their opinion fit better or is more appropriate for the given sentence.
2. in the second experiment, the same sentences were presented plus the original audio file from the audiobook and users were asked to choose among the three synthetic voices the one that is more close to the reference in terms of voice style.

Ten sentences of each style, according to our voice style prediction, were selected; as a reference the 30 sentences are presented in Table 6. In both experiment users were given the opportunity to select "none", when they could not decide and the text was presented on the screen. Six users, non-native speakers of English participated in the two experiments, four of the listeners are speech experts. The users listened ten sentences of each style in random order. There was no training phase, so the users were not familiar with the three voice styles in the first test, this was also intended to avoid influencing any preference.

In order to overcome a bit the problem of speech quality, the ten sentences in each style were selected from the sets with which the voices were trained. Some average features of the three sets are presented in Table 4. Clear tendencies of the features can be observed in this table, since the sentences were selected from the two extreme sets and center; they present quite a difference in average regarding F0, also the sentiment scores "Wow" and "Imdb-Emphasis" present clear differences.

Although these tests were performed with few listeners it gave us some insights regarding the aspects investigated in this paper. The test results of the informal test are presented in Table 5, main observations are:

- first, users seem to agree on a style preference for the selected sentences, and that style is also in agreement with the style automatically predicted; this effect is more clear for the sentences in extreme styles where the users agree 56% of the times.

Level	F0	No. words	No. quotes	Wow	Imdb-Emphasis
veryhigh	173.8	8.8	0.7	0.039	0.00457
center	102.1	13.6	0	0.070	-0.00089
verylow	79.7	9.3	0	0.114	-0.00142

Table 4: Average measures for 30 sentence in the informal listening test, the selected sentences are presented in Table 6.

- second, the different styles were perceived by the users, regardless of the low quality of the speech, again the extreme styles seem to be easier to identify with 71.6% for veryhigh style and 65% for verylow style.

The low percentages for none also indicate that users most of the time have a defined preference, which is contrary to the experiments reported in [20], where significant differences among subject's individual voice style preferences for particular sentences are reported. This again might have to do with the selection of sentences and their content, as described in [19].

	(a) Preferred style by users %			
Predicted style	veryhigh	center	verylow	none
veryhigh	56.6	20.0	18.3	5.0
center	15.0	40.0	40.0	5.0
verylow	6.6	33.3	56.7	3.3
	(b) Perceived style by users %			
Original style	veryhigh	center	verylow	none
veryhigh	71.6	10.0	6.6	11.6
center	6.6	51.7	33.3	8.3
verylow	3.3	16.7	65.0	15.0

Table 5: (a) Appropriateness of a style for a sentence, diagonal agreement: 51.1%, (b) Perception of a style, diagonal agreement: 62.8%.

5. Conclusions

We have described an architecture for synthesis of expressive speech based on the MARY TTS system and sentiment analysis of text. The creation of several HMM-based voices in different styles using audiobook data is explained. We have described how we use sentiment analysis scores extracted from text sentences and acoustic features extracted from the corresponding audio sentences to build a prediction model of voice style. Also we have described how this model can be used together with the set of HMM-based voices to synthesise expressive speech.

General conclusions from the Blizzard listening test results regarding our entry are: (i) the low speech quality of our entry affected heavily the evaluation, (ii) users did not perceive expressivity variation among the high, center and low styles and (iii) the selection of the test sentences has an impact on the evaluation, in particular for our system.

These results partially helped us to evaluate aspects of our system, although aspects like the use of sentiment analysis to predict a voice style, or the actual perception of extreme voice styles were not covered. Therefore we carried out an informal listening test in which we evaluate these aspects. Two important observations can be preliminary concluded from this informal test: first, users seem to agree on a style preference (51.1%) for particular sentences, and that style is also in agreement with the style automatically predicted; second, the voice styles of the extreme voices and center were perceived by the users (62.8%), regardless of the low quality of the speech. These results encourage us to continue researching in the main ideas presented in this work and improving the overall quality of the synthetic speech. In future experiments we might consider to reduce the number of voice styles, at least until we manage to improve the quality of the synthetic speech. In this sense the experience of other participants in the Blizzard challenge would be very interesting in particular the entries that also participate with HMM-based voices.

6. Acknowledgements

This work is supported by the EU project SSPNet (FP7/2007-2013). The author would like to thank Christopher Potts for providing the sentiment analysis of the data; Marc Schröder for the restructuring and implementation of the new features in MARY TTS 5.0 system and Holmer Hemsen for helpful discussion and support with the listening test.

7. References

[1] M. Schröder, S. Pammi, and O. Türk, "Multilingual MARY TTS participation in the Blizzard Challenge 2009," in *Blizzard Challenge Workshop*, Edinburgh, UK, 2009.

[2] J. Reilly and L. Seibert, "Language and emotion," in *Handbook of Affective Sciences*, R. J. Davidson, K. R. Scherer, and H. H. Goldsmith, Eds. Academic Press, 2003, ch. 27, pp. 535–559.

[3] B. Liu, "Sentiment analysis and subjectivity," in *Handbook of Natural Language Processing, Second Edition*, N. Indurkhya and F. J. Damerau, Eds. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2010.

[4] M. Charfuelan and M. Schröder, "Correlation analysis of sentiment analysis scores and acoustic features in audiobook narratives," in *ES3-LREC Workshop*, Istanbul Turkey, 2012.

[5] M. Schröder, M. Charfuelan, S. Pammi, and I. Steiner, "Open source voice creation toolkit for the MARY TTS Platform," in *Interspeech*, Florence, Italy, 2011.

[6] K. Tokuda, K. Oura, S. Shiota, k. Hashimoto, T. Nose, H. Zen, J. Yamagishi *et al.*, *HMM-based Speech Synthesis System (HTS)*, 2012. [Online]. Available: http://hts.sp.nitech.ac.jp/

[7] M. Charfuelan, *HMMVoice creation for MARY TTS 5.0*, 2012. [Online]. Available: https://github.com/marytts/marytts/wiki/HMMVoiceCreation

[8] *Speech Synthesis Markup Language (SSML) Version 1.1*, W3C Recommendation, 2010. [Online]. Available: http://www.w3.org/TR/speech-synthesis11/#edef_prosody

[9] S. Pammi, "Prosody control in HMM-based speech synthesis," DFKI Speech Technology Lab., Tech. Rep., 2011.

[10] *Emotion Markup Language (EmotionML) 1.0*, W3C Recommendation, 2012. [Online]. Available: http://www.w3.org/TR/emotionml/

[11] N. Braunschweiler, M. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Interspeech*, Makuhari, Chiba, Japan, 2011.

[12] P. Bo, L. Lillian, and V. Shivakumar, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10*, ser. EMNLP '02, Stroudsburg, PA, USA, 2002, pp. 79–86.

[13] B. Liu, "Opinion mining, sentiment analysis, and opinion spam detection," http://www.cs.uic.edu/∼liub/FBS/sentiment-analysis.html, 2011.

Sentence	Predicted style	Text
chp08_00040	veryhigh	I want to go home."
chp09_00062	veryhigh	I never heard the beat of that in all my days!
chp01_00163	veryhigh	"Well I WILL, if you fool with me."
chp17_00426	veryhigh	Tom Sawyer's Gang!
chp17_00394	veryhigh	I never see such a woman!
chp01_00160	veryhigh	I could lick you with one hand tied behind me, if I wanted to."
chp06_00159	veryhigh	"Tom, what on earth ails that cat?"
chp06_00076	veryhigh	And you said, 'Dont torment me so -- I'll tell!'
chp01_00043	veryhigh	Forty times I've said if you didn't let that jam alone I'd skin you.
chp06_00210	veryhigh	some people think they're mighty smart, -- always showing off!"
chp09_00226	center	He wished there was some way to get that boy into trouble without much risk to himself.
chp08_00006	center	They had a famous fried-egg feast that night, and another on Friday morning.
chp08_00226	center	This was to knock off being pirates, for a while, and be Indians for a change.
chp05_00031	center	It was on a hill, about a mile and a half from the village.
chp07_00200	center	But the talk soon began to drag, and then died.
chp03_00333	center	Tom knew that when his name was pronounced in full, it meant trouble.
chp05_00004	center	Tom lay awake and waited, in restless impatience.
chp17_00184	center	Huck began to dig and scratch now.
chp16_00231	center	The village was illuminated; nobody went to bed again; it was the greatest night the little town had ever seen.
chp02_00112	center	Mary took his book to hear him recite, and he tried to find his way through the fog:
chp05_00027	verylow	Huckleberry Finn was there, with his dead cat.
chp17_00033	verylow	Huck's face saddened.
chp05_00391	verylow	This was worse than a thousand whippings, and Tom's heart was sorer now than his body.
chp05_00084	verylow	A muffled sound of voices floated up from the far end of the graveyard.
chp05_00289	verylow	The boys clasped each other suddenly, in an agony of fright.
chp16_00129	verylow	Becky's face paled, but she thought she could.
chp09_00393	verylow	The stillness continued; the master searched face after face for signs of guilt.
chp05_00059	verylow	This was a damper, and conversation died again.
chp05_00217	verylow	whispered Tom, in short catches between breaths.
chp12_00160	verylow	Joe's knife struck upon something.

Table 6: Predicted voice style and corresponding text for sentences used in the informal listening test.

[14] C. Potts, "Sentiment Symposium Tutorial: Lexicons. Section 3.4 Experience Project reaction distributions," http://sentiment.christopherpotts.net/lexicons, 2011.

[15] K. Sjölander, "The snack sound toolkit," http://www.speech.kth.se/snack, 2012.

[16] C. Busso, S. Lee, and S. Narayanan, "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 17, no. 4, pp. 582–596, 2009.

[17] W. C. Chu, *Mixed excitation linear prediction*, ser. Speech coding algorithms Foundations and Evolution of Standardized Coders. Wiley, 2003, ch. 17, pp. 454–485.

[18] S. Mohammad, "From once upon a time to happily ever after: Tracking emotions in novels and fairy tales," in *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Portland, OR, USA: Association for Computational Linguistics, June 2011, pp. 105–114. [Online]. Available: http://www.aclweb.org/anthology/W11-1514

[19] F. Hinterleitner, G. Neitzel, S. Möller, and C. Norrenbrock, "An evaluation protocol for the subjective assessment of text-to-speech in audiobook reading tasks," in *Blizzard Challenge Workshop*, Turin, Italy, 2011.

[20] E. Székely, J. P. Cabral, M. Abou-Zleikha, P. Cahill, and J. Carson-Berndsen, "Evaluating expressive speech synthesis from audiobook corpora for conversational phrases," in *LREC*, Istanbul, Turkey, 2012.